e-ISSN: 2148-7456

a peer-reviewed online journal

hosted by DergiPark

International Journal of Assessment Tools in Education

Volume: 9 Issue: 1 March 2022

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

IJATE

e-ISSN 2148-7456

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

Volume 9 Issue 1 2022

Editor : Dr. Eren Can AYBEK
Address : Pamukkale University, Education Faculty,
Kinikli Campus, 20070 Denizli, Turkiye
Phone : +90 258 296 1174
Fax : +90 258 296 1200
E-mail : ijate.editor@gmail.com

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

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

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

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

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

International Journal of Assessment Tools in Education

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

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

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

Indexes and Platforms:

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

CONTENTS

Research Articles

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 1, 1–19

https://doi.org/10.21449/ijate.946171

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

The Collective Teacher Efficacy Behaviours Scale: A Validity and Reliability Study

Seyfettin Kapat[1], Sevilay Sahin[2], Mevlut Kara[3,*]

[1]Republic of Turkey, Ministry of Education, Gaziantep, Turkiye
[2]Gaziantep University, Faculty of Education, Department of Educational Sciences, Gaziantep, Turkiye
[3]Gaziantep University, Nizip Faculty of Education, Department of Educational Sciences, Gaziantep, Turkiye

ARTICLE HISTORY

Received: June 1, 2021
Revised: Nov. 29, 2021
Accepted: Jan. 13, 2022

Keywords:
Collective efficacy behaviours,
Teacher,
Scale,
Scale development,
Social cognitive theory.

Abstract: The concept of collective efficacy that can be defined as "a belief in their common ability to organize and realize plans to achieve goals" (Bandura, 1997, p. 477) has gained utmost importance in educational contexts. Therefore, there arises an emergent need to develop scales to evaluate teachers' collective efficacy behaviours. To this end, the present study aimed to develop an instrument to assess collective teacher efficacy behaviours. For this purpose in mind, an item pool was created in line with the related literature and face-to-face interviews with teachers. Two participating groups were included in the study. There was a total of 833 participants, 475 of which were in the first group and 358 in the second group. The preliminary version of The Collective Teacher Efficacy Behaviours Scale (CTEBS), consisted of 26 items. Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) were employed to test the construct validity of the scale with the available datasets. As a result of the EFA, a two-factored structure, namely social and professional relationship and professional development, was identified with 20 items. The two factors explained 58.798% of the total variance. Confirmatory factor analysis (CFA) was used to test the validity of the structure based on the EFA results. It was found that the CFA fit indices were χ^2/df=3.174, RMSEA=.076, SRMR=.435, NFI=.902, CFI=.930, IFI=.931, and GFI=.872. The results implied that The Collective Teacher Efficacy Behaviours Scale, consisting of two dimensions and 20 items, was a valid and reliable instrument.

1. INTRODUCTION

Human beings, as social entities, may overcome difficult tasks, adapt to society, and accelerate personal and professional development more easily with a collective lifestyle. The individuals' faster integration into society and their concordant actions are directly related to their acceptance by society. In this regard, individuals tend to meet their requirements to communicate effectively and act in cooperation with their environment. Similarly, individuals need to coexist with such concerns as preventing potential problems, improving working conditions, and increasing the existing level of achievement (Demir, 2019). Individuals who act together and support each other may be more effective and efficient in solving the problems

*CONTACT: Mevlut KARA ✉ mevlutkara85@gmail.com Gaziantep University, Nizip Faculty of Education, Department of Educational Sciences, Gaziantep, Turkiye

e-ISSN: 2148-7456 **/© IJATE** 2022

they are likely to encounter. Collective power is supposed to arise among the individuals working together if they meet those requirements. Direct and indirect learning may emerge as a result of the relationships between individuals in organizations based on a collective understanding. The development and consequences of social relations between individuals is an area of interest for social cognitive theory. According to the mutual causality principle of social cognitive theory, individual and organizational factors mutually affect each other in organizations (Kurt, 2012). From this point of view, Bandura (1982) discussed the concept of collective efficacy, which implies a greater phenomenon than the sum of self-efficacies in an organization, and pointed out that people never live in isolation from the social environment and some hard work can only be achieved through working together.

The literature review introduces a great many definitions for the concept of collective efficacy. Bandura (1997) embraced the concept of collective efficacy from an organizational perspective and defined it as "a belief in their common ability to organize and realize plans to achieve goals" (p. 477). Tschannen-Moran and Barr (2004) dealt with this notion in terms of schools as educational organizations and described it as "a school characteristic that creates a difference on students' schooling, unlike teachers' self-efficacy" (p. 190). The main goal of schools is to improve the quality of education and training and ultimately increase student achievement. In the harsh conditions of our age, it is far from reality for teachers to increase student achievement solely with their separate efforts (Yılmaz & Turanlı, 2017). It is of great importance that teachers support each other and act in harmony for in-school activities in order to be more effective and efficient during the educational processes. The creation of an organization with collective efficacy depends on employees' support for each other, teamwork, cooperation (Yılmaz & Uslu, 2018), and solidarity (Demir, 2019).

From the perspective of student outcomes, Abedini et al. (2018) identified the concept of collective efficacy as "the educators' perceptions for their ability to positively affect student outcomes" (p. 2). Goddard et al. (2000) characterized it as "the interactive product of the group members at school for student achievement" (p. 483). As collective efficacy includes interactive, coordinated, and synergistic social dynamics unlike self-efficacy (Yorulmaz & Erdem, 2017), the concept of collective efficacy becomes even more and more important for school climate and the school outcomes concerning student achievement. Collective teacher efficacy has started to be frequently investigated, especially in educational studies, because of its positive effects on school outcomes (Koçak & Özdemir, 2019). Goddard et al. (2000) asserted that teachers are more effective on students at schools with high collective teacher efficacy. Collective efficacy, which is an organizational characteristic of schools (Schechter & Tschannen-Moran, 2006) is a phenomenon that positively or negatively affects cooperative teacher behaviours (Lee et al., 2011), instructional school decisions (Goddard, 2002), higher expectations and openness to new ideas (Donohoo, 2018), and their performance qualities (Abedini et al., 2018). In terms of students' academic development, it can be alleged that schools with high collective efficacy may positively affect students' development (Belfi et al., 2015), result in student learning and achievement (Eells, 2011; Goddard et al., 2000), significantly predict the level of success between schools (Goddard, 2002), reduce the negative effects of students' sociodemographic variables (Ramos et al., 2014), and help teachers motivate their students better (Erdoğan & Dönmez, 2015).

Collective teacher efficacy is significant not only for student achievement (Goddard, 2002; Tschnann-Moran & Barr, 2004) but also in terms of teachers' job satisfaction, commitment to their students, positive attitudes towards students, and professional development (Donohoo, 2018). Strengthening collective efficacy at schools would be possible with the development of co-working behaviours, the adoption of school vision by teachers, getting everyone's ideas and opinions in problem-solving, creating encouraging environments for student learning, and teachers' keeping themselves up to date (Turhan & Yaraş, 2014).

An examination of the related literature shows that a great many instruments have been developed to measure collective efficacy perceptions (Schwarzer & Jarusalem, 1999; Goddard et al., 2000; Goddard, 2002; Tschannen-Moran & Barr, 2004; Carroll et al., 2005; Pepe et al., 2008; Kurt, 2009; Erdoğan & Dönmez, 2015; Abedini et al., 2018). The social cognitive theory asserts that there are differences between people's levels of perception and their behaviours, that is, not every piece of knowledge and skill could be observed explicitly (Kurt, 2009). In this regard, the present study focused on measuring collective teacher efficacy behaviours, unlike their levels of perception, with a specific purpose to contribute to the relevant literature by developing a valid and reliable instrument to assess collective teacher efficacy behaviours. Besides, this scale is considered as the development of first original in the Turkish context concentrating on teachers' collective efficacy behaviours.

2. METHOD

In this part of the study, the scale development procedures for Collective Teacher Efficacy Behaviours Scale (CTEBS) are explained in detail. During this process, the following stages proposed by DeVellis (2017) were followed:

- Determining the behaviours to be assessed,
- Creating an item pool,
- Determining the measurement method,
- Taking the opinions of field experts,
- Implementing the scale,
- Analysing the items,
- Finalizing the scale based on the analyses.

2.1. Study Group

The research data were collected from two different groups and at different times in the academic year of 2020-2021. Exploratory Factor Analysis (EFA) was employed on the dataset of the first study group and Confirmatory Factor Analysis (CFA) was performed on the dataset of the second study group. The teachers in the study groups were active at schools. The demographics of the first and second study groups are presented in Table 1.

According to data in Table 1, it can be asserted that both sample groups were similar as the percentages of the variables for the first group in which the EFA was employed and the second group in which the CFA was performed were quite close to each other. Besides, exploratory factor analysis and correlation analysis were performed on the dataset of the first study group, while confirmatory factor analysis, 27% lower-upper group analysis, Cronbach's alpha, and composite reliability analysis were performed on the dataset of the second study group.

Table 1. *The demographics of the first and second study groups.*

First Study Group				Second Study Group			
Variable	Group	N	%	Variable	Group	N	%
Gender	Female	261	54.9	Gender	Female	203	56.7
	Male	214	45.1		Male	155	43.3
Age	22-30	206	43.4	Age	22-30	167	46.6
	31-40	193	40.6		31-40	127	35.5
	41-50	64	13.5		41-50	51	14.2
	51 and over	12	2.5		51 and over	13	3.6
Level of education	Associate	4	0.8	Level of education	Associate	6	1.7
	Bachelor	415	87.4		Bachelor's	311	86.9
	Master	53	11.2		Master's	37	10.3
	PhD	3	0.6		PhD	4	1.1
Professional seniority	1-5	215	45.3	Professional seniority	1-5	160	44.7
	6-10	96	20.2		6-10	60	16.8
	11-20	120	25.3		11-20	102	28.5
	20 and over	44	9.3		20 and over	36	10.1
School type	Primary school	176	37.1	School type	Primary school	141	39.4
	Secondary school	175	36.8		Secondary school	129	36.0
	High school	124	26.1		High school	88	24.6

2.2. Scale Development Process

An item pool was initially created to assess collective teacher efficacy behaviours. During the formation of the item pool, both the literature was reviewed and the teachers were interviewed. A 41-item pool was created as a result of the review of related literature (Abedini et al., 2018; Bandura, 1997; Blatti et al., 2019; Borgogni et al., 2010; Çelik et al., 2018; Donohoo, 2017; Donohoo et al., 2018; Eells, 2011; Goddard et al., 2000; Goddard et al., 2004; Guskey & Passaro, 1994; Gürçay et al., 2009; Kurt, 2009; Özcan, 2017; Parker et al., 2006; Ross et al., 2003; Ross & Bruce, 2007; Schwarzer & Jarussalem, 1999; Turhan & Yaraş, 2014; Uğurlu et al., 2018; Ware & Kistantas, 2007; Yılmaz & Uslu, 2018; Yorulmaz & Erdem, 2017) and face to face interviews with 19 teachers individually.

A draft version of the instrument was developed based on the item pool. Three experts in the field of educational administration were consulted to examine the content validity of the draft version. The experts were asked to choose among the options of *"appropriate, should be improved,* and *inappropriate"* and were encouraged to express their opinions under the option of *"explanations"*. Büyüköztürk et al. (2018) uttered that the necessary arrangements should be made in case the items are unsatisfactory, and the inappropriate ones should be removed based on expert opinions. Accordingly, the scale was reduced to 26 items after the exclusion of 15 items that were deemed to be inappropriate for measuring similar behaviours by the experts. For the face validity, the 26-item draft form was edited by two assistant professors who are experts in the field of the Turkish language. A pilot scheme was conducted with 22 teachers to determine the level of understandability of the draft scale, which was finalized in line with their feedback.

A five-point Likert-type grading was used to determine whether teachers agree with the items in the scale. The options in the scale were "*1- Do not agree at all*, 2- *Disagree*, 3- *Neutral*, 4- *Agree*, and 5- *Completely agree*".

2.3. Data Collection

The research data were collected during the Covid 19 pandemic in 2021. Due to the closure of the schools, the computer-assisted survey was formed to reach the teachers. The instrument consisted of three sections. The first part included the purpose of the study and an informed consent section where the participants declared their voluntary participation in the study. The second part consisted of five questions (gender, age, level of education, professional seniority, and school type) to determine their demographics. And in the third part a 26-item scale was given.

In the beginning, the sample size for the analyses was determined. While Nunally (1978) asserted that reaching a sample of 10 times the number of items would be sufficient, Tabachnick and Fidell (2001) pointed out that the sample size of 300 was acceptable and that of 1000 was perfect. In this regard, 475 participants were included in the first study group for the 26-item version and 358 participants were covered in the second study group for the 20-item version based on the statistical analyses for the first group data. The research data were collected in December 2020. Before the analyses, the research data were examined and a total of 17 surveys were excluded as the presence of outliers in the data set would affect the correlation size (Best & Kahn, 2017).

2.4. Data Analysis

At the first step, the EFA was employed to determine the construct validity of the draft version of the scale. Principal Axis Factoring (PAF) was used in the EFA since the researcher may prefer the principal axis factoring method to understand the latent variables among the observed ones (Karaman et al., 2017). This method also yields a composite result by combining the common and unique variables (Karaman, 2015). On the other hand, Tabachnick and Fidel (2012) specified that promax rotation may be preferred since the results to be obtained by a researcher to perform oblique rotation would be more applicable than the direct oblimin rotation for the future. In this regard, promax rotation was chosen among oblique rotation methods while performing exploratory factor analysis. Tabachnick and Fidel (2012) suggested that the factor loadings should be greater than .32 for the item to be a member of any factor. The present study conforms to the aforementioned criteria.

The CFA was employed to test and verify the structure obtained as a result of the EFA. There are a great many indices in the CFA to reveal the compliance of the structure. This study examined the chi-square goodness test, Goodness of Fit Index (GFI), Adjusted Goodness of Fit Index (AGFI), Comparative Fit Index (CFI), Parsimony Normed Fit Index (PNFI), Parsimonious Goodness of Fit Index (PGFI), Normed Fit Index (NFI), Incremental Fit Index (IFI), the Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR).

In order to test the criterion validity, the relationship between the current scale and the Collective Teacher Efficacy Scale (CTES) adapted by Erdoğan and Dönmez (2015) was examined. The reliability of the data was tested through Cronbach’s Alpha and composite reliability methods. SPSS 22.0, AMOS 21, and Microsoft Excel were used for all calculations in the study. The composite reliability coefficient was estimated with formula-1 using path coefficients and error variances obtained from the CFA.

$$CR = \frac{(\sum_{i=1}^{n} \lambda_i)^2}{(\sum_{i=1}^{n} \lambda_i)^2 + (\sum_{i=1}^{n} \delta_i)} \qquad \text{(Formula-1)}$$

3. FINDINGS

Findings regarding the validity and reliability of The Collective Teacher Efficacy Behaviours Scale are presented in this section. The EFA and the CFA were employed to test the construct validity of the data obtained by using the CTEBS.

3.1. Exploratory Factor Analysis

The suitability of the dataset for the analyses was initially examined to employ the exploratory factor analysis. To that end, KMO and Bartlett's sphericity tests were conducted. KMO test is supposed to be greater than 0.60 and the Bartlett sphericity test is to be significant for the adequacy of sample size (İslamoğlu & Alnıaçık, 2016). Kalaycı (2006) affirmed that the KMO coefficient over 0.90 indicates a perfect dataset for analysis. In this study, the KMO coefficient was estimated to be 0.966, and Bartlett's test of sphericity (x^2=9619,895, *df*=325, *p* <.000) was found to be significant. Based on these, the dataset was determined to be convenient for the analysis.

Büyüköztürk (2002) pointed out that the eigenvalues should be scrutinized, the factors with a score greater than 1 should be assumed valid, and the line graph (scree plot) for the factor eigenvalues should be reviewed to determine the number of factors. On the other hand, Uyar (2012) noted that the most consistent criterion in determining the number of factors is parallel analysis while Pallant (2007) asserted that parallel analysis results should be included in the process of reporting the findings of studies in the fields of education and psychology. This method was used in the present study as it is claimed that the number of factors should be determined with the parallel analysis method (Brown, 2006). Table 2 indicates the factor eigenvalues of the draft version of the scale and the factor eigenvalues after the parallel analysis.

Table 2. *The findings of EFA and parallel analysis eigenvalues.*

Factor	EFA Eigenvalues	PA Eigenvalues	Conclusion
1	14.050	1.455	Accepted
2	1.715	1.384	Accepted
3	1.306	1.333	Rejected

Factors with eigenvalues of and above 1 are considered noteworthy in factor analysis (Pedhazur & Pedhazur Schmelkin, 1991). In determining the factors through parallel analysis, it is necessary to compare the eigenvalues of the real dataset with the randomly selected data and exclude factors up to the point where the eigenvalues of the real data are greater than those of the random data (Akbaş et al., 2019). When Table 2 is examined according to the aforementioned criteria, it can be seen that the first and second factors were accepted since the EFA eigenvalues were higher than those of the parallel analysis were. It can be explained that the scree plot, which is used as an auxiliary graph to decide the number of factors, will be cut in the area where the points are flattened and the following eigenvalues will be small and approximate (Çokluk et al., 2012). Graph 1 displays the results of the analyses. It was obvious that the slope in the line graph decreased significantly after the second factor. When the factor eigenvalues and scree plot were considered together, it was concluded that the scale could have a two-factor structure.

Graph 1. *Scree plot.*

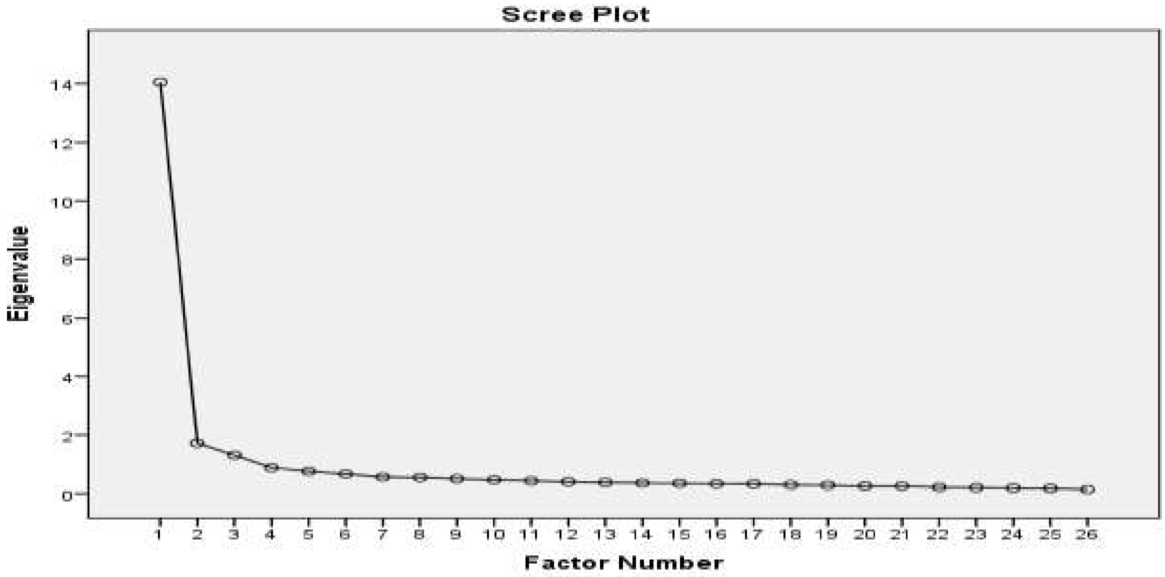


Table 3. *Item factor loadings.*

	Factor	
Item No	1	2
i12	.868	
i18	.803	
i10	.782	
i21	.773	
i9	.741	
i19	.739	
i23	.737	
i24	.687	
i11	.672	
i15	.666	
i8	.664	
i16	.646	
i22	.636	
i20	.633	
i13	.631	
i26	.580	
i14	.547	
i7		.954
i5		.907
i2		.838
i1		.774
i4		.736
i3		.634
i25		.481
i6	.374	.467
i17	.339	.406

According to Table 3, the item factor loadings varied between .339 and .954. The analysis results implied that there were statistical problems associated with item overlapping between the factors. Regarding the item overlap correlation, Büyüköztürk (2012) suggested that the difference between two high loadings should be .10 at least. Moreover, the lower limit of the item factor loadings was determined to be .32 in factor analysis (Tabachnick & Fidel, 2001). In this vein, the analyses were reiterated by respectively excluding the overlapping items (i6 and i17) and those deemed to be incompatible with the factors based on expert opinion (m8, i14, i23, and i25). The emergent structure is presented in Table 4. According to Table 4, a structure consisting of two factors and 20 items was obtained as a result of the reiterated EFA. The factor loadings of the items varied between .499 and .919.

Table 4. *EFA loadings after the exclusion of items threatening construct validity.*

Item		Factor Loadings	
No	Common Variance	Factor 1	Factor 2
i12	.722	.908	
i18	.639	.825	
i21	.717	.824	
i10	.578	.823	
i9	.698	.780	
i19	.437	.736	
i11	.568	.704	
i13	.625	.673	
i15	.528	.651	
i16	.579	.617	
i22	.533	.601	
i24	.433	.554	
i26	.472	.543	
i20	.340	.499	
i7	.752		.919
i5	.736		.865
i2	.595		.811
i1	.614		.751
i4	.617		.669
i3	.575		.628
Explained Total Variance: %58.798		52.544%	6.254%

3.2. Factor Labelling

The factors of the EFA were re-examined in terms of the expression of the items and the factors were labelled as in Table 5.

Table 5. *Factor labelling.*

Factor	Number of Items	Items	Sample items All of us as teachers …
Social and Professional Relationship (SPR)	14	9, 10, 11, 12, 13, 15, 16, 18, 19, 20, 21, 22, 24, 26	... are with our fellow teachers in their special occasions and hard times. … unite when a colleague of us is exposed to an unfairness.
Professional Development (PD)	6	1, 2, 3, 4, 5, 7	… keep ourselves up to date to ensure our professional development. … make every attempt for our students' academic achievement.

3.3. Confirmatory Factor Analysis

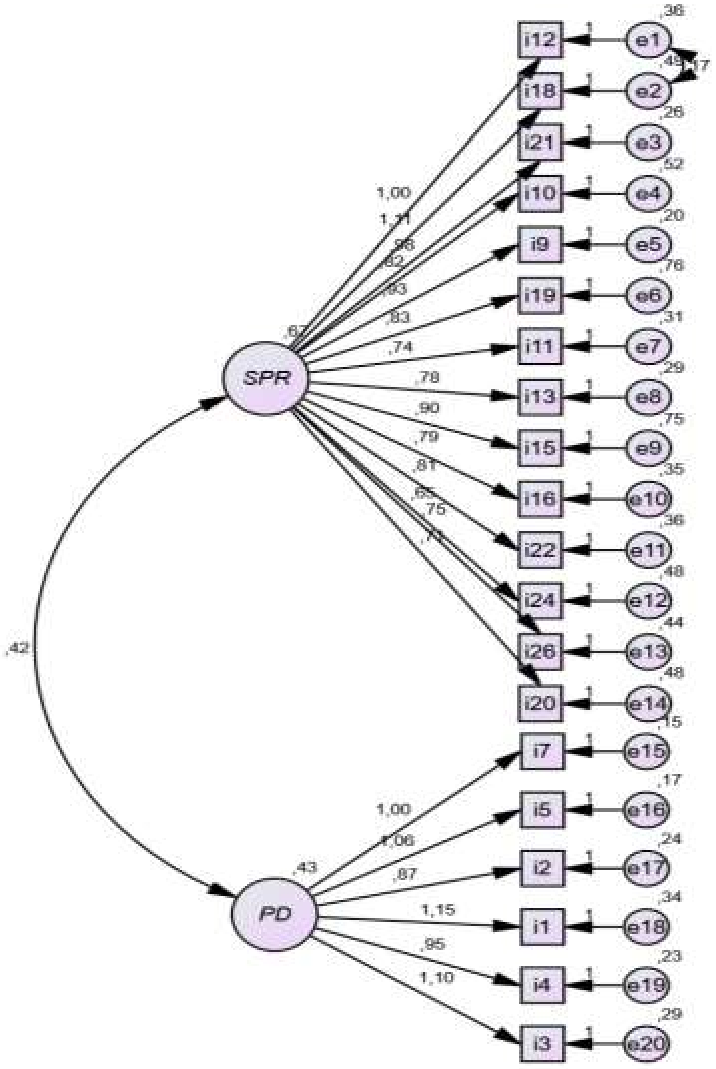
In the confirmatory factor analysis, the researcher tests the hypothesis suggested based on theoretical grounds (Balcı, 2016). The CFA was employed to determine the compliance of the emergent two-dimensional structure with 20 items as a result of the EFA. Crowley and Fan (1997) recommended that various fit indices should be used as parameters in the CFA. In this vein, fit indices and the results based on the CFA are submitted in Table 6.

Table 6. *Acceptable indices and the results of the CFA.*

Fit Index	Index	Perfect Fit	Acceptable Fit	Result
χ^2/df	3.174	$0 < \chi^2/df \leq 3$	$3 < \chi^2/df \leq 5$	Acceptable Fit
RMSEA	.076	$.00 \leq \text{RMSEA} \leq .05$	$.05 < \text{RMSEA} \leq .08$	Acceptable Fit
SRMR	.0435	0<SRMR≤.05	.05<SRMR≤.10	Perfect Fit
CFI	.930	.95 ≤ CFI ≤ 1.00	.90 ≤ CFI ≤ .95	Acceptable Fit
GFI	.872	.90 ≤ GFI ≤ 1.00	.85 ≤ GFI ≤ 90	Acceptable Fit
NFI	.902	.95 ≤ NFI ≤ 1.00	.90 ≤ NFI ≤ .95	Acceptable Fit
IFI	.931	.95 ≤ IFI ≤ 1.00	.90 ≤ IFI ≤ .95	Acceptable Fit
PNFI	.797	.95 ≤ PNFI ≤ 1.00	.50 ≤ PNFI ≤ .95	Acceptable Fit
PGFI	.698	.95 ≤ PGFI ≤ 1.00	.50 ≤ PGFI ≤ .95	Acceptable Fit

When the reference ranges of the indices in Table 6 and the results for the dataset were examined together, it was clear that the two-factor model had an acceptable fit (Bentler & Bonett, 1980; Byrne & Campell, 1999; Schumacker & Lomax, 2010; Tabachnick & Fidel, 2012; Doğan, 2013; İlhan & Çetin, 2014; Karagöz, 2017). The path diagram for the model and factor loadings based on the CFA are presented in Figure 1.

Figure 1. *Path diagram for the model and factor loadings based on the CFA.*



As a result of the CFA, the structure consisting of a total of 20 items, 14 of which are in the Social and Professional Relationship dimension and 6 of them in the Professional Development dimension, was confirmed. As can be seen in Figure 1, factor loadings ranging from .61 to .86 for the sub-dimensions support the model fit.

3.4. Criterion Validity

The correlation coefficients between The Collective Teacher Efficacy Behaviours Scale (CTEBS) and the Collective Teacher Efficacy Scale (CTES) adapted by Erdoğan and Dönmez (2015) were analysed with the data obtained from 53 teachers within the scope of criterion validity. The results are presented in Table 7.

Table 7. *Criterion validity findings.*

	CTES	Student Discipline	Instructional Strategies
CTEBS	.069	.047	.083
Social and Professional Relationship	.074	.051	.088
Professional Development	.008	.003	.012

An examination of Table 7 indicates that the correlation coefficient between the CTEBS and CTES is .069 and ranges from .003 to .088 for the dimensions. This finding implies that CTES, focusing on the collective perceptions of teachers, and CTEBS, concentrating on collective behaviours, intend to assess different aspects.

3.5. Reliability

The reliability of the emergent scale was determined through Cronbach's alpha and composite reliability coefficients. The scores are given in Table 8.

Table 8. *Reliability coefficients for CTEBS.*

Factors	Cronbach's Alpha	Composite Reliability
Social and Professional Relationship	.912	.939
Professional Development	.919	.853
Overall	.938	.962

As displayed in Table 8, Cronbach's Alpha coefficient of .938 for the *overall scale*, .912 for the *Social and Professional Relationship* dimension, and .919 for the *Professional Development* dimension were estimated. On the other hand, composite reliability coefficients were computed based on the factor loadings and error variances in the CFA. Composite reliability coefficients were determined as .962 for the *overall scale*, .939 for the *Social and Professional Relationship* dimension, and .853 for the *Professional Development* dimension. As the reliability coefficients over .70 indicate that an instrument is reliable (Liu, 2003), it can be alleged that the reliability coefficients for CTEBS are satisfactorily high in our study.

3.6. Item Analysis

Lower-upper group item analysis was conducted to determine item discrimination (Tezbaşaran, 1997). In this vein, independent samples t-test was performed for the lower (n=97) and upper (n=97) groups, based on rankings according to the highest and lowest scores for each item, and the item-total correlations are submitted in Table 9.

Table 9. *Item-total correlations and lower-upper group item analysis results.*

Item No	Item Total Correlation	*t*	*p*
i1	.701	15.226	.000
i2	.653	11.735	.000
i3	.720	15.525	.000
i4	.730	13.691	.000
i5	.734	15.222	.000
i7	.714	14.316	.000
i9	.824	18.825	.000
i10	.736	13.937	.000
i11	.752	13.540	.000
i12	.816	19.841	.000
i13	.792	13.280	.000
i15	.748	13.042	.000
i16	.774	13.830	.000
i18	.788	2.573	.011
i19	.652	12.633	.000
i20	.618	11.987	.000
i21	.826	19.636	.000
i22	.745	13.510	.000
i24	.686	11.560	.000
i26	.708	13.627	.000

An examination of the findings in Table 9 yields that t values for 20 items in the scale are between 2.573 and 19.841. Accordingly, the significance of t values implies that the items are discriminatory. It can also observed that the item-total correlations vary between .652 and .826, which implies that each item is coherent with the scale. The examination of item analysis for the lower-upper groups results indicate that all the items in the scale have a high level of reliability and item discriminations are significant (Büyüköztürk, 2012).

4. DISCUSSION and CONCLUSION

This study aimed to develop a valid and reliable instrument for collective teacher efficacy behaviours. For this purpose, the scale development stages suggested by DeVellis (2017) were followed. In this vein, a pool of 41 items was initially created by reviewing the relevant literature and interviewing the teachers. 30 items in the pool were created based on the literature review (Abedini et al., 2018; Bandura, 1997; Bandura, 2000; Blattivd, 2019; Borgogni et al., 2010; Çelik et al., 2018; Donohoo, 2017; Eells, 2011; Goddard et al., 2000; Goddard et al., 2004; Gürçay et al., 2009; Kurt, 2009; Kurt, 2012; Lee & Smith, 1996; Parker et al., 2006; Ross et al., 2003; Schwarzer & Jarussalem, 1999; Tschannen-Moran & Barr, 2004; Turhan & Yavaş, 2014; Ware & Kistantas, 2007; Yılmaz & Uslu, 2018; Yorulmaz & Erdem, 2017), and 11 items were based on the interviews with the teachers. Expert opinion was taken for the content and face validity of the scale. Based on the expert opinions, 15 items were eliminated and a 26-item draft scale was created. A five-point Likert-type grading including "*do not agree at all, disagree, neutral, agree,* and *completely agree*" was used for the items in the scale.

Exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) were employed to test the construct validity of the CTEBS. As a result of the first EFA, the overlapping two items (items 6 and 17) and theoretically incompatible four items (items 8, 14, 23, and 25) with the dimensions were respectively excluded and the EFA was reemployed. As a result of the second EFA, a two-factor structure consisting of 20 items and explaining 58.798% of the total variance was obtained. The factors were labelled as *Social and Professional Relationship* and *Professional Development* in line with the relevant literature. Bandura (1997) highlighted that individuals working in a group cannot be socially isolated from group members and Goddard et al. (2000) asserted that collective teacher efficacy is a result of the emergent interactive dynamics within a group. Therefore, the first factor was labelled as "social and professional relationship". On the other hand, Parker et al. (2006) pointed out that collective efficacy is a crucial factor for explaining the differences in student achievement and expertise by experience is essential in improving student achievement. Considering that collective efficacy affects student success in the classroom (Ross & Gray, 2006) by conducting the necessary activities to create positive student outcomes (Goddard et al., 2004), the second factor was labelled as "professional development".

The CFA was employed to determine whether the model based on the EFA was verified or not. The fit indices were reported together with the model (path diagram in Figure 1) based on the CFA. It was observed that the fit indices were within acceptable limits, and a two-factor model consisting of 20 items was confirmed. The examination of the EFA and the CFA results implied that the scale developed had construct validity. On the other hand, the "Collective Teacher Efficacy Scale" adapted by Erdoğan and Dönmez (2015) was used as a criterion and the correlation coefficients between the overall scores of the two scales and their dimensions were estimated to test the criterion validity of the scale. According to the results of statistical analysis, it was concluded that CTEBS and CTES assessed different aspects of collective efficacy.

Cronbach's Alpha, composite reliability coefficient, and 27% lower-upper item analysis were computed to test the reliability of the data obtained by using the CTEBS. Cronbach's Alpha coefficient of .938 for the *overall scale*, .912 for the *Social and Professional Relationship* dimension, and .919 for the *Professional Development* dimension were found. Composite reliability coefficients were determined as .939 for the *Social and Professional Relationship* dimension, and .853 for the *Professional Development* dimension. The reliability coefficients over .70 indicate that the instrument is reliable (Liu, 2003; Tezbaşaran, 1997). Moreover, it was concluded that all the items were discriminatory based on the 27% lower-upper group analysis conducted to determine the distinctiveness of the items in the CTEBS.

The findings revealed that the CTEBS (Appendix 1) is a valid and reliable instrument to be used to test the collective efficacy behaviours of teachers working in primary, secondary, and high schools. For future studies, it may be suggested that the validity of the scale should be tested on preschool teachers as a different sample group. Considering the theoretical background of collective teacher efficacy, its relationship with such variables as organizational culture, academic achievement, leader-member exchange, teacher leadership, and organizational citizenship can be scrutinized.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research and publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number:** Gaziantep University/Institute of Educational Sciences, E-39083294-050.06-185

Authorship Contribution Statement

Seyfettin Kapat: Investigation, Resources, Methodology, Visualization, Software, Data Collection, Formal Analysis, and Writing the Original Draft. **Sevilay Sahin**: Framing, Methodology, Supervision, and Validation. **Mevlut Kara**: Framing, Data Collection, Investigation, Methodology, Supervision and Validation, Software, Formal Analysis, and Writing the Original Draft.

Orcid

Seyfettin Kapat https://orcid.org/0000-0003-2211-3025
Sevilay Sahin https://orcid.org/0000-0002-7140-821X
Mevlut Kara https://orcid.org/0000-0002-6381-5288

REFERENCES

Abedini, F., Bagheri, M.S., & Sadighi, F. (2018). Exploring Iranian collective teacher efficacy beliefs in different ELT settings through developing a context–specific English language teacher collective efficacy scale. *Cogent Education*, *5*(1), 1552340. https://doi.org/10.1080/2331186X.2018.1552340

Akbaş, U., Karabay, E., Yıldırım-Seheryeli, M., Ayaz, A., & Demir, Ö.O. (2019). Türkiye ölçme araçları dizininde yer alan açımlayıcı faktör analizi çalışmalarının paralel analiz sonuçları ile karşılaştırılması [Comparison of exploratory factor analysis studies in Turkish measurement tools index according to parallel analysis results]. *Kuramsal Eğitimbilim Dergisi*, *12*(3), 1095-1123. https://doi.org/10.30831/akukeg.453786

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

Bandura, A. (1982). The assessment and predictive generality of self-percepts of efficacy. *Journal of Behavior Therapy and Experimental Psychiatry*, *13*(3), 195-199. https://doi.org/10.1016/0005-7916(82)90004-0

Bandura, A. (2000). Exercise of human agency through collective efficacy. *Current Directions in Psychological Science*, *9*(3), 75-78. https://doi.org/10.1111/1467-8721.00064

Balcı, A. (2016). *Sosyal bilimlerde araştırma yöntem, teknik ve ilkeler* (12. Baskı) [*Research methods, techniques and principles in social sciences* (12. Baskı)]. Pegem Akademi.

Belfi, B., Gielen, S., De Fraine, B., Verschueren, K., & Meredith, C. (2015). School-based social capital: The missing link between schools' socioeconomic composition and collective teacher efficacy. *Teaching and Teacher Education*, 45, 33-44. https://doi.org/10.1016/j.tate.2014.09.001

Bentler, P.M., & Bonett, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588-606. https://psycnet.apa.org/doi/10.1037/0033-2909.88.3.588

Best, J.W., & Kahn, J.V. (2017). *Eğitimde araştırma yöntemleri* [*Research methods in education*], (O. Köksal, Çev.). Eğitim Yayınevi.

Blatti, T., Clinton, J., & Graham, L. (2019). Exploring collective teacher efficacy in an international school in Shanghai. *International Journal of Learning, Teaching and Educational Research*, *18*(6), 214-235. https://doi.org/10.26803/ijlter.18.6.13

Borgogni, L., Petitta, L., & Mastrorilli, A. (2010). Correlates of collective efficacy in the Italian Air Force. *Applied Psychology*, *59*(3), 515-537. https://doi.org/10.1111/j.1464-0597.2009.00410.x

Brown, T.A. (2006). *Confirmatory factor analysis for applied research*. Guilford Press.

Büyüköztürk, Ş. (2002). Faktör analizi: Temel kavramlar ve ölçek geliştirmede kullanımı [Factor analysis: Basic concepts and using to development scale]. *Educational Administration: Theory and Practice*, *8*(4), 470-483.

Büyüköztürk, Ş. (2012). *Sosyal bilimler için veri analizi el kitabı [Manual of data analysis for social sciences]* (16. Baskı). Pegem Akademi.

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö.E., Karadeniz, Ş. & Demirel, F. (2018). *Bilimsel araştırma yöntemleri [Scientific research methods]* (25. Baskı). Pegem Akademi.

Byrne, B.M., & Campbell, T.L. (1999). Cross-cultural comparisons and the presumption of equivalent measurement and theoretical structure: A look beneath the surface. *Journal of Cross-Cultural Psychology*, *30*(5), 555-574. https://doi.org/10.1177/0022022199030005001

Carroll, J.M., Rosson, M.B., & Zhou, J. (2005, April). Collective efficacy as a measure of community. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1-10). https://doi.org/10.1145/1054972.1054974

Crowley, S.L., & Fan, X. (1997). Structural equation modeling: Basic concepts and applications in personality assessment research. *Journal of Personality Assessment*, *68*(3), 508-531. https://doi.org/10.1207/s15327752jpa6803_4

Çelik, K., Gören, T., & Kahraman, Ü. (2018). The relationship between elementary school teachers' levels of collective efficacy and morale levels. *Journal of Human Sciences*, *15*(4), 2644-2656. https://doi.org/10.14687/jhs.v15i4.5651

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik SPSS ve LİSREL uygulamaları [Multivariate statistics for social sciences: SPSS and LISREL applications]* (2.Baskı). Pegem Akademi.

DeVellis, R.F. (2017). *Scale development theory and applications*. Sage Publications.

Doğan, M. (2013). *Doğrulayıcı faktör analizinde örneklem hacmi, tahmin yöntemleri ve normalliğin uyum ölçütlerine etkisi [Influence of sample size, estimation method and normality on fit indices in confirmatory factor analysis]* [Unpublished master's thesis]. Eskişehir Osmangazi University.

Demir, S. (2019). Kolektif öğretmen yeterliğinin öğretmen iş doyumundaki rolü üzerine yapısal eşitlik modellemesi [Structural equation modeling on the role of teacher's collective efficacy in teacher job satisfaction]. *OPUS Uluslararası Toplum Araştırmaları Dergisi*, *10*(17), 444-463. https://doi.org/10.26466/opus.496333

Donohoo, J. (2017). Collective teacher efficacy research: Implications for professional learning. *Journal of Professional Capital and Community*, *2*(2), 101-116. https://doi.org/10.1108/JPCC-10-2016-0027

Donohoo, J. (2018). Collective teacher efficacy research: Productive patterns of behaviour and other positive consequences. *Journal of Educational Change*, *19*(3), 323-345. https://doi.org/10.1007/s10833-018-9319-2

Donohoo, J., Hattie, J., & Eells, R. (2018). The power of collective efficacy. *Educational Leadership*, *75*(6), 40-44.

Eells, R.J. (2011). *Meta-analysis of the relationship between collective teacher efficacy and student achievement* [Unpublished doctoral dissertation]. Loyola University.

Erdoğan, U., & Dönmez, B. (2015). Kolektif öğretmen yeterliği ölçeğinin Türkçeye uyarlanması: Geçerlik ve güvenirlik çalışması [Adaptation of collective teacher efficacy scale into Turkish: Validity and reliability study]. *Kuram ve Uygulamada Eğitim Yönetimi Dergisi*, 21(3), 345-366. http://dx.doi.org/10.14527/kuey.2015.013

Goddard, R.D., Hoy, W.K., & Hoy, A. W. (2000). Collective teacher efficacy: Its meaning, measure, and impact on student achievement. *American Educational Research Journal*, *37*(2), 479-507. https://doi.org/10.3102/00028312037002479

Goddard, R. (2002). A theoretical and empirical analysis of the measurement of collective efficacy: The development of a short form. *Educational and Psychological Measurement*, *62*(1), 97-110. https://doi.org/10.1177/0013164402062001007

Goddard, R.D., Hoy, W.K., & Hoy, A.W. (2004). Collective efficacy beliefs: Theoretical developments, empirical evidence, and future directions. *Educational Researcher*, *33*(3), 3-13. https://doi.org/10.3102/0013189X033003003

Goddard, R.D., LoGerfo, L., & Hoy, W.K. (2004). High school accountability: The role of perceived collective efficacy. *Educational Policy*, *18*(3), 403-425. https://doi.org/10.1177/0895904804265066

Guskey, T.R., & Passaro, P.D. (1994). Teacher efficacy: A study of construct dimensions. *American Educational Research Journal*, *31*(3), 627-643. https://doi.org/10.3102/00028312031003627

Gürçay, D., Yılmaz, M., & Ekici, G. (2009). Öğretmen kolektif yeterlik inancını yordayan faktörler [Factors predicting teachers' collective efficacy beliefs]. *H. U. Journal of Education*, *36*(36), 119-128.

İlhan, M., & Çetin, B. (2014). Sınıf değerlendirme atmosferi ölçeğinin (SDAÖ) geliştirilmesi: Geçerlik ve güvenirlik çalışması [Development of classroom assessment environment scale (CAES): Validity and reliability study]. *Education and Science*, *39*(176), 31-50 http://dx.doi.org/10.15390/EB.2014.3334

İslamoğlu, A.H., & Alnıaçık, Ü. (2016). *Sosyal bilimlerde araştırma yöntemleri [Research methods in social sciences]* (5.Baskı). Beta Yayınları.

Kalaycı, S. (2006). *SPSS uygulamalı çok değişkenli istatistik teknikleri [Multivariate statistical techniques with SPSS applied]*. Asil Yayınevi.

Karagöz, Y. (2017). *SPSS ve AMOS uygulamalı nicel-nitel-karma bilimsel araştırma yöntemleri ve araştırma etiği [SPSS and AMOS applied quantitative-qualitative-mixed scientific research methods and research ethics]*. Nobel Yayıncılık.

Karaman, H. (2015). *Açımlayıcı faktör analizinde kullanılan faktör çıkartma yöntemlerinin karşılaştırılması [The comparison of factor extraction strategies used in exploratory factor analysis]* [Unpublished master's thesis]. Hacettepe University.

Karaman, H., Atar, B., & Çobanoğlu Aktan, D. (2017). Açımlayıcı faktör analizinde kullanılan faktör çıkartma yöntemlerinin karşılaştırılması *[The comparison of factor extraction strategies used in exploratory factor analysis]*. *Gazi University Journal of Gazi Educational Faculty (GUJGEF)*, *37*(3), 1173-1193.

Koçak, S., & Özdemir, M. (2019). Kolektif öğretmen yeterliğinin dört çerçeve liderlik modeli perspektifinden değerlendirilmesi [Evaluation of collective teacher efficacy from the perspective of four-frame leadership model]. *Education and Science*, *45*(203), 347-365. http://dx.doi.org/10.15390/EB.2019.8325

Kurt, T. (2009). *Okul müdürlerinin dönüşümcü ve işlemci liderlik stilleri ile öğretmenlerin kolektif yeterliği ve öz yeterliği arasındaki ilişkilerin incelenmesi [Examination of* *relationships between transformational and transactional leadership styles of school principals and collective efficacy and self-efficacy of teachers]* [Unpublished doctoral dissertation]. Gazi University.

Kurt, T. (2012). Öğretmenlerin öz yeterlik ve kolektif yeterlik algıları [Self-efficacy and collective-efficacy perceptions of teachers]. *Türk Eğitim Bilimleri Dergisi*, *10*(2), 195-227.

Lee, J. C.K., Zhang, Z., & Yin, H. (2011). A multilevel analysis of the impact of a professional learning community, faculty trust in colleagues and collective efficacy on teacher commitment to students. *Teaching and Teacher Education*, *27*(5), 820-830. https://doi.org/10.1016/j.tate.2011.01.006

Lee, V.E., & Smith, J.B. (1996). Collective responsibility for learning and its effects on gains in achievement for early secondary school students. *American Journal of Education*, *104*(2), 103-147. https://doi.org/10.1086/444122

Liu, Y. (2003). Developing a scale to measure the interactivity of websites. *Journal of Advertising Research*, *43*(2), 207-216. https://doi.org/10.2501/JAR-43-2-207-216

Nunally, J.C., & Bernstein, I.H. (1978). Psychometric theory – 25 years ago and now. *Educational Researcher*, *4*(10), 7-21. https://doi.org/10.3102/0013189X004010007

Özcan, S. (2017). *Özgün (authentic) liderliğin duygusal örgütsel bağlılık üzerindeki etkisinde işyerindeki esenlik, kolektif yeterlik ve kurumsal itibar değişkenlerinin rolü [The mediating roles of well-being at work, collective efficacy and organizational reputation in the relationship between authentic leadership and affective organizational behaviour]* [Unpublished doctoral dissertation]. Gebze Teknik University.

Pallant, J. (2007). *SPSS survival manual: A step by step guide to data analysis using the SPSS program* (3. Edition). McGraw Hill.

Parker, K., Hannah, E., & Topping, K.J. (2006). Collective teacher efficacy, pupil attainment and socio-economic status in primary school. *Improving Schools*, *9*(2), 111-129. https://doi.org/10.1177/1365480206064965

Pedhazur, E. J., & Pedhazur Schmelkin, L. (1991). *Measurement, design and analysis: An integrated approach*. Taylor & Francis Group.

Pepe, S., Sobral, J., Gómez-Fraguela, J.A., & Villar-Torres, P. (2008). Spanish adaptation of the adolescents' perceived collective family efficacy scale. *Psicothema*, *20*(1), 148-154.

Ramos, M.F.H., Costa, S.S., Pontes, F.A. R., Fernandez, A.P.O., & Nina, K.C.F. (2014). Collective teacher efficacy beliefs: A critical review of the literature. *International Journal of Humanities and Social Science*, *4*(7), 179-188.

Ross, J., & Bruce, C. (2007). Professional development effects on teacher efficacy: Results of randomized field trial. *The Journal of Educational Research*, *101*(1), 50-60. https://doi.org/10.3200/JOER.101.1.50-60

Ross, J.A., Hogaboam-Gray, A., & Gray, P. (2003, April). *The contribution of prior student achievement and school processes to collective teacher efficacy in elementary schools* [Conference presentation abstract]. American Educational Research Association, Chicago, IL, United States.

Ross, J.A., & Gray, P. (2006). Transformational leadership and teacher commitment to organizational values: The mediating effects of collective teacher efficacy. *School Effectiveness and School Improvement, 17*(2), 179-199. https://doi.org/10.1080/09243450600565795

Schechter, C., & Tschannen-Moran, M. (2006). Teachers' sense of collective efficacy: an international view. *International Journal of Educational Management*, *20*(6), 480-489. https://doi.org/10.1108/09513540610683720

Schumacker, R. E. & Lomax, R. G. (2010). *A beginner's guide to structural equation modeling.* Taylor & Francis Group.

Schwarzer, R., & Jerusalem, M. (1999). *Skalen zur Erfassung von Lehrer- und Schülermerkmalen: Dokumentation der psychometrischen Verfahren im Rahmen der wissenschaftlichen Begleitung des Modellversuchs Selbstwirksame Schulen*. Freie Universität Berlin.

Tabachnick, B.G., & Fidel, L.S. (2001). *Using multivariate statistics*. (4. Edition). Allyn & Bacon, Inc.

Tabachnick, B.G., & Fidell, L.S. (2012). *Using multivariate statistics* (6. Edition). Pearson.

Tezbaşaran, A.A. (1997). *Likert tipi ölçek geliştirme kılavuzu [Likert scale development guide]*. Türk Psikologlar Derneği.

Tschannen-Moran, M., & Barr, M. (2004). Fostering student learning: The relationship of collective teacher efficacy and student achievement. *Leadership and Policy in Schools*, *3*(3), 189-209. https://doi.org/10.1080/15700760490503706

Turhan, M., & Yaraş, Z. (2014). İlkokul yöneticilerinin program liderliği davranışlarını gösterme düzeylerinin öğretmenlerin kolektif yeterlik algısına ve örgütsel öğrenme düzeyine etkisi [The influence of instructional leadership behavior of school administrators on perceived collective competencies of teachers and the level of organizational learning]. *Journal of Educational Sciences,* 39, 175-193. https://doi.org/10.15285/EBD.2014397404

Uğurlu, C.T., Beycioğlu, K., & Abdurrezzak, S. (2018). Bilgi okuryazarlığı, kolektif öğretmen yeterliği ve etkili okul: Yapısal eşitlik modellemesi [Information literacy, collective teacher adequacy and effective school: Structural equation modeling]. *Elementary Education Online*, *17*(4), 1988-2005.

Uyar, S. (2012). *Açımlayıcı faktör analizinde boyut sayısını belirlemede kullanılan yöntemlerin karşılaştırılması [Comparision of procedures for determining the number of dimensions in exploratory factor analysis]* [Unpublished master's thesis]. Hacettepe University.

Ware, H., & Kitsantas, A. (2007). Teacher and collective efficacy beliefs as predictors of professional commitment. *The Journal of Educational Research*, *100*(5), 303-310. https://doi.org/10.3200/JOER.100.5.303-310

Yılmaz, M., & Turanlı, N. (2017). Öğretmenlerin kolektif yeterlik algılarının incelenmesi: Altındağ ilçesi örneği [Examination on teachers' collective efficacy perception: Altındag district sample]. *The Journal of International Lingual Social and Educational Sciences*, *3*(2), 151-158.

Yılmaz, M. & Uslu, Ö. (2018). Güdülenmiş öğrenmeyi destekleme öz-yeterlik algısının kollektif yeterlik, tükenmişlik ve teknolojiyle bütünleşmeyle ilişkisi [Relationship between supporting motivated learning self-efficacy, collective efficacy, burn-out and technology integration]. *Ege Journal of Education, 19*(1), 225-244. https://doi.org/10.12984/egeefd.375587

Yorulmaz, R., & Erdem, R. (2017). Hastane çalışanlarında kontrol odağının öz ve kolektif yeterlilik üzerine etkisi [The effect on self and collective efficacy of locus of control in the hospital employess]. *Visionary E-Journal*, *8*(19), 77-92. https://doi.org/10.21076/vizyoner.317182

APPENDIX

The Collective Teacher Efficacy Behaviours Scale (CTEBS)

Dimension	Item No	All of us as teachers … Öğretmenler olarak hepimiz…	Do not agree at all	Disagree	Neutral	Agree	Completely agree
Social and Professional Relationship	1	... are with our fellow teachers in their special occasions and hard times. …öğretmen arkadaşlarımızın özel ve zor günlerinde yanında oluruz.					
	2	… have a dynamic relationship with our colleagues. …meslektaşlarımızla aramızda dinamik bir ilişki vardır.					
	3	… act synergistically. …sinerjik bir şekilde hareket ederiz.					
	4	… work in coordination. …eşgüdüm halinde görev yaparız.					
	5	… have discussions with our colleagues to improve teaching activities. …öğretim faaliyetlerinin geliştirilmesi için meslektaşlarımızla tartışmalar yaparız.					
	6	… frequently communicate with our colleagues to support our students' development. …öğrencilerimizin gelişimlerini desteklemek amacıyla meslektaşlarımızla sık sık iletişime geçeriz.					
	7	… unite when a colleague of us is exposed to an unfairness. …bir meslektaşımıza karşı adaletsiz bir durum olduğunda birlik oluruz.					
	8	… organize various outdoor activities (trips, social events, etc.) with our colleagues. …meslektaşlarımızla okul dışı zamanlarda çeşitli etkinlikler (gezi, sosyal etkinlik vb.) düzenleriz.					
	9	… express opinions in decisions concerning the entire school. …okulun tamamını ilgilendiren kararlarda fikirlerimizi belirtiriz.					
	10	… find solutions to the in-school problems with a common sense. …okul içerisinde meydana gelen problemler karşısında çözüm yollarını ortak akılla buluruz.					
	11	… share with our colleagues when we learn new professional knowledge. …mesleki anlamda yeni bir bilgi öğrendiğimizde bu bilgiyi meslektaşlarımızla paylaşırız.					
	12	… try to help each other improve their teaching methods and techniques. …birbirimize öğretim yöntem ve tekniklerini geliştirmeleri konusunda yardımcı olmaya çalışırız.					
	13	… ask for feedback from teacher colleagues at school in improving education. …okulumuzdaki meslektaşlarımızdan eğitim-öğretimin geliştirilmesi ile ilgili geri bildirim alırız.					
	14	… trust each other in professional matters. …birbirimize mesleki konularda güveniriz.					
	15	… attempt to motivate our students in their learning process. …öğrencilerimizi öğrenme süreçlerinde motive etmek için çaba sarf ederiz.					

Professional Development	16	… make every attempt for our students’ academic achievement. …öğrencilerimizin akademik başarı elde etmeleri amacıyla her türlü girişimde bulunuruz.					
	17	… strive for our students to be successful. …öğrencilerimizin başarılı olması için çaba gösteririz.					
	18	… fairly treat our students with different levels of academic achievement. …akademik başarısı birbirinden farklı olan öğrencilerimize adaletli davranırız.					
	19	… use different teaching strategies for student development. …farklı öğretme stratejilerini öğrencilerin gelişimi için kullanırız.					
	20	… keep ourselves up to date to ensure our professional development. …mesleki gelişimimizi sağlamak için güncel gelişmeleri takip ederiz.					

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 1, 20–38

https://doi.org/10.21449/ijate.1058055

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

Examining Science Process Skills Tests: A Case of Turkey

Okan Sibic [1,*], Burcin Acar Sesen [1]

[1]İstanbul University-Cerrahpaşa, Hasan Ali Yücel Faculty of Education, Department of Mathematics and Science Education, Science Education, Turkiye

ARTICLE HISTORY

Received: Feb. 26, 2021
Revised: Nov. 28, 2021
Accepted: Jan. 13, 2022

Keywords:

Graduate theses in science education,
Science process skills,
Meta-synthesis study.

Abstract: One of the main goals of science education is to make students gain science process skills. Thus, it is significant to measure whether students gain those skills or not. For this purpose, various tests have been produced and used in various studies. This study aims to examine science process skills tests which have been used in the theses produced in the field of science education from the perspectives of the originality, question types used, and the science process skills measured in the tests, and the number of questions for each measured science process skill. Within the scope of this meta-synthesis study, 82 master's theses and 34 doctoral dissertations from Turkey were analyzed. The findings indicate that science process skills were measured with multiple-choice tests, and only in smaller number of studies, original tests were developed for the corresponding study. It was also discovered that some science process skills were measured more frequently than others. As a result of the study, some suggestions were provided.

1. INTRODUCTION

In the last decades, there have been rapid development and changes in science and technology to make life easier. In order to keep up-to-date, countries revise their educational programs, including their science programs, continuously. Similarly, educational programs were reviewed in Turkey to be able to catch up with the recent trends in the world, and various changes were made so as to fulfil new necessities. Thus, the science curriculum in Turkey was revised to keep up with the rapid developments in science and technology in 2018 (Ministry of National Education, 2018).

The aim of the revised science curriculum of Turkey is to make students obtain basic knowledge and skills about science and engineering applications, to teach students nature-human interactions and to make them able to produce solutions towards the problems that occur as a result of nature-human interactions by benefiting from science process skills, to raise consciousness about sustainable development in students, to make students obtain knowledge and ability to use science process skills towards everyday life problems that they might be confronted, to help students develop science career knowledge, to teach students how scientists produce scientific knowledge and help them develop the science process skills in order to produce scientific knowledge, to emphasize the importance of reliability and validity in scientific studies, to help students develop interest and positive attitudes towards the nature, to

*CONTACT: Okan SIBIÇ ✉ okan.sibic@iuc.edu.tr İstanbul University-Cerrahpaşa, Hasan Ali Yücel Faculty of Education, Department of Mathematics and Science Education, Science Education, Turkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

emphasize the socio-scientific issues and by using such issues, to help students develop scientific reasoning skills, habit of scientific thought, and decision-making skills (Ministry of National Education, 2018). Additionally, it is possible to say that with the success of the new program, students will be curious and sensitive about the events and problems they confront in their surroundings and about the solution of the problems, and they will behave like a scientist. In the new program, it is realized the development and use of science process skills are emphasized frequently.

Science process skills (SPSs) are defined as the process skills that scientists practise during scientific knowledge (Aslan et al., 2016; Temiz & Tan 2013). On the other hand, in the literature, it is possible to find some other definitions of SPSs. Ostlund (1992) and Charleswoth and Lind (2012) defined SPSs as skills which are used during the production of scientific knowledge, regulating the produced knowledge and also analyzing and solving the problems occurred in the process of producing scientific knowledge. In a similar vein, Anagün and Yaşar (2009) define SPSs as thinking skills used during the production of scientific knowledge and reasoning skills used about problems occurred during the process of producing scientific knowledge. Çepni et al. (1997) emphasized SPSs as some basic skills which make students active during learning by placing them into the center of learning and make learning easy and permanent and lead students to take responsibilities of their own learning. In addition to that, they emphasized that SPSs are the skills used in science laboratories and laboratory approach (Çepni et al., 1997). American Association for the Advancement of Science (AAAS) (AAAS, 1993) defines SPSs as the skills used during the production of scientific knowledge and behaviors, which are accepted in most of the science disciplines (Tan & Temiz, 2003). When all the definitions are analyzed, it is observed that although there are some differences between them, they generally emphasize the same points.

In the literature, there are different studies which explain SPSs. Although there are some differences, all the studies point out the same skills as science process skills (Aslan et al., 2016). When the studies examined, it is observed that SPSs are generally divided into one group, two groups or three groups. For instance, Rezba et al. (2007) divided SPSs into two categories: basic SPSs and integrated SPSs. While the skills of observing, communicating, classifying, measuring metrically, inferring and predicting are basic SPSs, the skills of identifying variables, constructing a table of data, constructing a graph, describing relationships between variables, defining relationships between variables, acquiring and processing your own data, analyzing investigations, constructing hypotheses, defining variables operationally, designing experiments are included in integrated SPSs (Rezba et al., 2007). On the other hand, SPSs such as observing, classifying, measuring, finding space and time relationship, using numbers, prediction, inferring, communicating, making operational definition, defining and controlling the definitions, formulating hypothesis, making experiments, interpreting data, and creating models are included into one group named as science process skills by AAAS (AAAS, 1993). Çepni et al. (1997) grouped 13 SPSs into different categories at the end of the project undertaken by the Council of Higher Education/World Bank Development of National Education in 1997. In this study, the last classification of SPSs was considered. According to this classification, while observing, classifying, measuring, and finding space and time relationships are taken into basic SPSs group, making predictions, determining variables, interpreting data and inferring are handled in causal SPSs group. In the third group, which is experimental process skills, formulating a hypothesis, using data and formulating models, designing-making experiments, controlling variables and decision-making were included by the researchers. While Çepni et al. (1997) emphasized the importance of bringing students basic SPSs, they also pointed out that bringing students basic SPSs also makes it easier for them to develop higher order thinking skills. Additionally, Çepni et al. (1997) point out that causal SPSs comprise the skills used in the process of testing hypothesis and skills that are used until making logical results after testable studies. On the other hand, experimental SPSs are defined as complicated, versatile, requiring higher order thinking skills, which contains one or more basic process skills (Çepni et al., 1997).

In order to develop scientific knowledge, as a scientist would do, SPSs are important practical skills (Ondowo & Indoshi, 2013), and have an important role while students are producing scientific knowledge and learning the nature of science directly (Erkol & Ugulu, 2014). Therefore, it is important to measure whether students gain those skills or not. To do that, different tests have been produced in the literature. Among those tests, today, the most frequently used one is the SPS tests developed by Burns et al. (1985). In this test, there are thirty-six multiple-choice questions, and five different process skills are measured through those questions. The process skills measured are determining variables, formulating hypothesis, making operational definition, and interpreting data and graphs. The adaptation study of this test to Turkish was conducted by Geban et al. (1992). It is possible to encounter different SPSs tests in the literature apart from this test (for example Enger & Yager, 1998; Smith & Welliver, 1994; Temiz, 2001; Tobin & Capie, 1981). In addition to all those tests, it is possible to say that in some studies, different SPSs tests were developed by considering the aim, sample, subject area, etc. In Turkey, for example, Daşdemir (2012), Demirörs (2018), Gültekin (2018), and Tatar (2006) developed new SPSs tests to measure students' gains towards SPSs. When the features of the SPSs tests were examined, it is possible to say that most of the tests were developed at primary levels and in science education. However, there were also some tests which were developed at secondary levels (9-11), e.g., Kazeni (2005). Burns et al. (1985) also developed their tests at primary and secondary levels (7-12).

Because SPSs are the skills that scientists use to reach scientific knowledge, they are standing out to as the skills which students should gain those skills under the scope of science education. At that point, the importance of bringing students the ability to use science process skills are emphasized in the new science education curriculum frequently as in the previous curriculum and different studies have been conducted to bring those skills to students. Additionally, different SPSs tests to measure whether the students were brought science process skills or not been observed in the literature.

This study aims to examine the SPSs tests used in the master's theses and doctoral dissertations (graduate theses) done in Turkey in the science education area from the perspectives of the originality of the tests, the question types, the SPSs measured, the number of questions of measured SPSs. The research questions of the study are as follows:

1. How is the distribution of originality of the SPS tests in graduate theses dissertations prepared in science education study area in Turkey?
2. How is the distribution of the question types (multiple-choice, open-ended, etc.) of the SPS tests in graduate theses prepared in science education study area in Turkey?
3. Which and to what frequency are the science process skills in SPS tests in graduate theses prepared in science education study area in Turkey measured?

2. METHOD

A meta-synthesis study which is also known as thematic content analysis method (Walsh & Downe, 2005) was conducted in the present study. Meta-synthesis study is a methodology in which qualitative and quantitative studies are used together. In meta-synthesis studies, qualitative and quantitative studies are used as data or unit of analysis. Meta-synthesis studies are principally "concerned with understanding and describing key points and themes contained within a research literature on a given topic" (Bair, 1999, p. 4). To follow a meta-synthesis study, the required steps were explained by Walsh and Downe (2005) as follows:

1. Determining the appropriate studies

2. Searching and evaluating the studies
3. Conceptualizing and comparing the studies
4. Synthesizing and reporting the findings.

In the theses -both master and doctoral- prepared in Turkey, SPSs tests were used widely. Therefore, in this study, in accordance with the definition of meta-synthesis studies, the SPSs used in the master and doctoral theses were examined from the perspectives of test style and type; the type of items of the tests; which process skills were measured, how many questions were used to measure each process skills; and the originality of the test, whether they were originally developed or adapted.

2.1. Criteria for Constructing the Sample of the Study

In the present study, in line with the aim of the study, the SPS tests were obtained from the master and doctoral theses published in Turkey between 2000-2019. To obtain the theses which contain SPS tests, 'Council of Higher Education Thesis Center' was used by the researchers. On the database, by typing 'bilimsel süreç becerileri' (Turkish translation of science process skills) keyword, detailed scanning was made, and 188 theses studies were reached. Firstly, all of the studies were reviewed in general and 69 out of 188 studies were decided as not compatible with the aim of the study. In Table 1, the reasons for excluding the studies from the sample were represented in terms of frequencies and percentage.

Table 1. *Reasons for excluding studies.*

Situation	Frequency	Percentage (%)
The sample of the studies not appropriate	26	37.68
Not contain SPS test	22	31.88
Not open to access	13	18.84
Not provide any information about SPS tests	8	11.59
Total	69	100

As presented in Table 1, 37.68% of the theses were not included in the study since the sample of these studies comprise preschool or elementary school students and 31.88% of the studies were not included in the sample since they did not have any SPS tests. Since 18.84% of them do not have open access and 11.59% of them do not have any information about the SPS tests which were used, those studies were not also included into the sample of the study.

After examining the theses, 116 of them were included in the examination. While 29.32% of the theses found were doctoral dissertations, it was determined that 70.68% were master's theses. In Table 2, distribution by years of the theses studies was presented in detail.

Table 2. *Distrubution by years.*

Year	Thesis ID	Frequency	Percentage
2001	T1	1	0.86
2006	T2, T3	2	1.72
2007	T4-T7; T83-T86	8	6.89
2008	T8-T15; T87, T88	10	8.62
2009	T16-T21; T89-T94	12	10.34
2010	T22-T26; T95-T98	9	7.75
2011	T27-T38; T99-T101	15	12.93
2012	T39-T50; T102, T103	14	12.06

Table 2. *Continued*

2013	T51-T62; T104-T107	16	13.79
2014	T63-T67; T108-T112	10	8.62
2015	T68-T73; T113	7	6.03
2016	T74-T78; T114-116	8	6.89
2017	T79-T81	3	2.58
2018	T82	1	0.86
Total		116	100

In Table 2, it is presented that SPS tests were started to be used in 2001, however, after this year, no study in which SPS tests were used was found until 2006. After 2006, it was observed that the interests towards SPS tests increased in the studies and between 2011-2013, SPS tests were included frequently in the theses. Besides, after 2006, SPS tests were included in at least one of the theses every year.

2.2. Data Collection

In the present study, an evaluation form which was prepared by the researchers in consistence with the aim of study was used as a data collection tool. The form, through which the features of SPS tests in graduate theses were determined, is presented in the 'Appendix' Section. The form was prepared by considering the SPSs stated as a result of the project of the Council of Higher Education/World Bank Development of National Education. Every SPS of the tests used in the theses was examined in detail and question numbers for each process skills were recorded in the corresponding form as data. In the cases in which different SPSs were observed except determining before (used for the first time), a new column was added to the right of the table, and new process skills were shown in the table to include those process skills.

2.3. Data Analysis

In the present study, to analyze the data, SPS tests were subjected to the content analysis technique. The SPS tests placed in the master's theses and doctoral dissertations were examined from the perspectives of what kind of tests they are (adaptation, original, etc.), type of test questions (open-ended, multiple-choice, etc.), the process skills measured, and the question numbers for each process skill; and all the findings were represented in different tables. All the SPS tests were included in the analysis without considering whether the same SPS tests were used in different theses or not.

2.4. Validity and Reliability of the Study

To conduct a valid and reliable study, first of all, an evaluation form was developed by one of the researchers of the present study so as to examine the SPS tests in the master's theses and doctoral dissertations. To provide the reliability of the form, three different researchers who are working in science education study area worked with the form independently to examine 10 different SPS tests found in the theses. After each independent examination, 98% consensus were built between them. Therefore, a reliable evaluation form was created.

To achieve credibility and conformability (Guba, 1981; Lincoln & Guba, 1985), the two researchers examined the SPS tests found in 116 theses independently and they built 95% consensus on the examinations. In order to achieve transferability, the whole process which were followed by the researchers to conduct this meta-synthesis study were explained in detail and each step was shown explicitly. In addition, all the details of the selection process of the theses were represent explicitly.

3. FINDINGS

SPS tests used in master's theses and doctoral dissertations were examined from the perspectives of originality. The findings represented in Table 3 point out that the 40.5% of the SPS tests used in theses were adaptations into Turkish; 37.1% of the SPS tests used in the theses were originally developed for corresponding theses; and 22.41% of the SPS tests used in the theses were revised versions of the adapted SPS tests to Turkish by considering the features of the sample, the aim of the corresponding study, etc. In Table 4, the findings were represented in detail.

Table 3. *Derivatives of Science Process Skills Tests.*

Derivatives	Frequency	Percentage (%)
Adapted	47	40.5
Originally Developed	43	37.1
Revised the Adapted Version	26	22.41
Total	116	100

In 16 (47.05%) doctoral dissertations and 31 (37.8%) master's theses, it was determined that an SPS test which was adapted to Turkish was used by the researchers of the corresponding studies. Additionally, it was observed that in 26 studies (20 master's theses and 6 doctoral dissertations), revised versions of the SPS tests according to the sample group of the corresponding studies were used. When graduate studies were examined, in 43 studies, it was seen that SPS tests were originally developed in each of those studies. It was found that among 34 doctoral dissertations, in 12 of them, SPS tests were originally developed; and, among 82 master's theses, in 31 of them, SPS tests were originally developed for each of the studies.

Table 4. *Sources of Adapted SPS Tests.*

Source (Author, Year)	Frequency	Percentage (%)
Test of Integrated Process Skills II (Burns, Okey & Wise, 1985)	30	63.8
Science Process Test (Enger & Yager, 1998)	8	17.02
The Science Process Assessment for Middle School Students (Smith & Welliver, 1994)	6	14.2
Test of Integrated Process Skills (Tobin & Capie, 1981)	2	4.25
Total	47	100

In the present study, examinations were made to find the adapted SPS tests placed in the graduate theses. The result of the examinations is represented in Table 4. When Table 4 is examined, it comes to the forefront that the adapted SPS test which was the most frequently used (63.8%) in the master's theses and doctoral dissertations is the one developed by Burns et al. (1985). It was found that in 12 master's theses and 18 doctoral dissertations, these SPS tests were used. In the theses, it was also found that the adapted version of SPS tests developed by Enger and Yager (1998) and Smith and Welliver (1994) were used. In two master's theses, the SPS test developed by Tobin and Capie (1981) was used.

Within the scope of second questions of the present study, the question types of the SPS tests were examined. In Table 5, the question types of SPS tests are presented in terms of frequency and percentage. When the table is examined, it can be explicitly observed that approximately in all theses, SPS tests which contained questions in multiple-choice format were used (104, 89.6%). On the other hand, it was seen that 8 SPS tests consisted of questions with open-ended and multiple-choice formats, two consisted of open-ended questions alone, and the other two contained questions with a mixed format.

Table 5. *Distributions of science process skills tests according to question types.*

Question Type	Frequency	Percentage (%)
Multiple-choice	104	89.6
Open Ended + Multiple Choice	8	6.90
Open Ended	2	1.72
Mixed	2	1.72
Total	116	100

When SPS tests in the master's theses and doctoral dissertations examined separately, it was found that among 34 doctoral dissertations, 27 of them had SPS tests with multiple-choice question format, 5 of them had open-ended and multiple-choice question format and the other two doctoral dissertations had SPS tests with mixed type (multiple-choice, open-ended, matching, etc.) question format. Similarly, it was observed that, in the master's theses, SPS tests with multiple-choice question format were used more frequently when compared to SPS tests in doctoral theses. While among 82 master's theses, 77 of them had SPS tests with multiple-choice question format, 3 of them had open-ended and multiple-choice question format and the other two master's theses have SPS tests with open-ended questions, no SPS tests were found in the master's theses and doctoral dissertations whose questions type were mixed.

The findings regarding which process skills and with how many questions those skills were measured were obtained through analysis by using the evaluation form (Appendix 6.1) prepared for data analysis are represented in Table 6. When Table 6 is examined, it is observed that there are 3800 questions in total in SPS tests used in the theses. In Table 7, SPSs typed in italic are the skills stated in the project of the Council of Higher Education/World Bank Development of National Education. Except for those skills, every process skill found in SPS tests has been included in Table 6. Every SPS test has been included for the examinations without considering whether those tests were used in other graduate theses of the sample. As a result of examinations, it was found that determining variables (878, 23.1%), formulating hypothesis (612, 16.1%), interpreting data (474, 12.47%), and making operational definition (287, 7.55%) were the most measured process skills. Additionally, it was observed that designing-making experiments (184, 4.84%), designing study (156, 4.1%), measuring (138, 3.63%), predicting (123, 3.23%), classifying (115, 3.02%), and observing (104, 2.73%) were other frequently measured process skills.

When the SPS tests were examined, it was found that making predictions (6, 0.15%) and decision making (7, 0.18%) are the least measured process skills. Additionally, presenting (0.1%), guessing (0.12%), describing (0.15%), defining the problem (0.23), asking questions (0.3%), comparing (0.31%), and using equipment (0.39%) are other least measured process skills in SPS tests. Distinctly, in one of the SPS tests, it was found that socio-scientific issues (1, 0.02%) were stated as process skills.

Table 6. *Process skills measured in science process skills tests and the frequency of measurement.*

Process Skills	Frequency	Percentage (%)
Observing	104	2.73
Classifying	115	3.02
Measuring	138	3.63
Data Recording	71	1.86
Founding Space and Number Relationship	74	1.94
Making Predictions	6	0.15
Determining Variables	878	23.1
Interpreting Data	474	12.47
Inferring	71	1.86
Formulating Hypothesis	612	16.1
Using Data and Formulating Models	103	2.71
Designing-Making Experiments	184	4.84
Controlling Variables	124	3.26
Decision Making	7	0.18
Comparing	12	0.31
Predicting	123	3.23
Communicating	38	1
Using Equipment	15	0.39
Designing Study	156	4.1
Making Operational Definition	287	7.55
Deducing	79	2.02
Guessing	5	0.12
Presenting	4	0.1
Describing	6	0.15
Using Numbers	21	0.53
Associating	22	0.56
Asking Questions	12	0.3
Logical Thinking	29	0.74
Making Study	20	0.51
Defining the Problem	9	0.23
Socio-scientific Issues	1	0.02
Total	3800	100

4. DISCUSSION and CONCLUSION

Within the scope of the study, the originality of the SPS tests was reviewed by the researchers. Among the SPS tests in master's theses and doctoral dissertations, it was revealed that to measure process skills of the students, an SPS test found the literature was used by the authors of the graduate studies. It was observed that the SPS test which was developed by Burns et al. (1985) and translated into Turkish by Geban et al. (1992) was used in master's theses and doctoral dissertations directly or through making revisions on the form according to the sample and the aim of the corresponding studies. For instance, in his study Aydoğdu (2006) pointed out that this test is for eighth graders and because his sample consisted of seventh graders, he used the adapted version of the tests by removing some questions. At that point, it can be emphasized that direct use of an SPS test found in the literature might not be always possible since the aim, sample, and the target group and the necessities of revising the SPS test retrieved from the literature might be a matter of discussion. On the other hand, among 36.12% of the

SPS tests used in the graduate theses, it was found that an original test was developed for the corresponding study. When the procedure of the development of SPS tests was examined, it was realized that the tests were developed by considering the aim, the sample, the success level of the sample, the subject matter, etc. All in all, in nearly more than half of the theses, adapted version of the SPS tests were used instead of developing a new SPS tests.

One of the other aims of the study is to search for the question type of the SPS tests used in master's theses and doctoral dissertations. Within the scope of this aim, every question of the SPS tests was examined in detail. It was observed that most of the questions (89.91%) of the SPS tests have a multiple-choice format. In addition, it is possible to figure out that SPS tests also have questions in mixed type question format or multiple-choice and open-ended question format at the same time although the number of those tests are less. It is emphasized that multiple-choice questions are objective and easy to score in which students are required to select the right choice among different alternatives (Tan, 2009). In this kind of test, students are restricted with the alternatives given to them and not free to answer the questions in their own ways (Tan et al., 2002). It is emphasized that multiple-choice tests measure the knowledge about facts, however, they do not provide students with the opportunity to regulate the knowledge they constitute and share the constituted knowledge (Yıldırım, 1983). Additionally, Tan (2009) emphasized that while multiple-choice tests can measure knowledge, comprehension and application behavior level, they are not enough to measure the behavior of creativity, producing idea and product, and synthesis. When SPSs are examined, it is possible to say that those skills have behavioral skills from knowledge to synthesis and creativity. Thus, the SPS tests which have only multiple-choice questions will not be enough to measure all the process skills. At that point, constructing SPS tests by using different type of questions comes to the forefront since by this way, SPSs and different type of behavior can be measured more easily.

It was aimed to find proportionally which process skills were measured in all of the SPS tests used in master's theses and doctoral dissertations. After examinations, it was revealed that causative and experimental SPSs were measured more frequently when compared to basic SPSs. It was pointed out that the most frequently measured SPSs were determining variables and formulating hypothesis. Kılıç et al. (2016) pointed out that determining variables is an important SPS since determining the variables of a study is an important factor which affects the research. In addition, hypothesis is a propositional statement which contains the knowledge about the effect of independent variables on the dependent variable (Kılıç et al., 2016) and thus, formulating hypothesis has an important place among SPSs. Although it has measured with fewer questions when compared to causative and experimental SPSs, there are still enough questions which measured basic SPSs. Çepni et al. (1997) emphasized that basic SPSs are useful in developing higher-order SPSs in students and pointed out the importance of teaching students the basic SPSs to teach them complex SPSs. In the present study, it was found that observing, classifying, and measuring are the process skills which are measured frequently through SPS tests.

SPSs are emphasized as the basic skills which are used to constitute scientific knowledge by using scientific way (Tan & Temiz, 2003) and make science learning easier and students active during lessons and help them to take the the responsibilities of their own learning (Çepni et al., 1997). From these perspectives, the importance of developing SPSs in students, deciding whether they gain those skills or not and the necessity of valid and reliable SPS tests to measure the outcomes come into prominence.

As a result of the examination, it was revealed that process skills were generally measured by using multiple-choice questions in the SPS tests. At that point, the requirement of a new SPS test in which each process skill was measured with appropriate question type comes to the forefront and preparing and using a new SPS test is suggested by the researchers.

Thanks to this study, it was observed that 21st century skills have been measured less frequently by existing SPS tests. Among the SPSs, only logical thinking and defining problem can be shown as 21st century skills which were used in the tests. However, the frequency of measuring those skills is very rare. At that point, constructing a new SPS test by using items which measure 21st century skills more frequently can be pointed out as another suggestion.

Acknowledgments

This study was presented at ERPA International Congress which was held in 2018 in İstanbul.

Declaration of Conflicting Interests and Ethics

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

Authorship Contribution Statement

Okan Sibic and **Burcin Acar Sesen** work collaboratively in the process of Investigation, Sampling, Examiniton, Writing of the Manuscripts. Burçin Acar Şeşen also supervised the writing process.

Orcid

Okan Sibic https://orcid.org/0000-0001-7241-274X
Burcin Acar Sesen https://orcid.org/0000-0002-1585-0441

REFERENCES

AAAS. (1993). *Benchmarks for science literacy, a project 2061 report.* Oxford University Press.

Anagün, Ş.S., & Yaşar, Ş. (2009). Developing scientific process skills at science and technology course in fifth grade students. *İlköğretim Online, 8*(3), 844-865.

Aslan, S., Kılıç, H.E., & Kılıç, D. (2016). *Bilimsel süreç becerileri [Science Process Skills]*. Pegem Akademi.

Bair, C. R. (1999, November). *Meta-synthesis*. The annual meeting of the Association for the Study of Higher Education, San Antonio, TX.

Burns, J.C., Okey, J.R., & Wise, K.C. (1985). Development of an integrated process skill test: TIPS II. *Journal of Research in Science Teaching, 22*(2), 169-177.

Charlesworth, R., & Lind, K.K. (2012). *Math and science for young children*. Wadsworth Cengage Learning.

Çepni, S., Ayas, A., Johnson, D., & Turgut, M.F. (1997). *Physics teaching*. YÖK/Dünya Bankası Milli Eğitimi Geliştirme Projesi, Hizmet Öncesi Öğretmen Eğitimi.

Daşdemir, İ. (2012). *The effect of using of animation on students? academic achievement, retention of learned knowledge and scientific process skills* [Unpublished doctoral dissertation]. Atatürk University.

Demirörs, F. (2018). *Effect of the 7e learning model enriched with self regulated cognitive strategies on the student's achievment in the subject matter energy and on their science process skills* [Unpublished doctoral dissertation]. Hacettepe University.

Enger, S.K., & Yager, R.E. (1998). *The Iowa assessment handbook*. Iowa University Science Education Center.

Erkol, S., & Ugulu, I. (2014). Examining biology teachers candidates' scientific process skill levels and comparing these levels in terms of various variables. *Procedia-Social and Behavioral Sciences, 116*, 4742-4747.

Geban, Ö., Aşkar, P., & Özkan, İ. (1992). Effects of computer simulations and problem-solving approaches on high school students. *The Journal of Educational Research, 86*(1), 5-10.

Guba, E.G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *ECTJ, 29*(2), 75-91.

Gültekin, B.G. (2018). *Examination of the impacts of activity based on science process skills on the problem solving abilities of students at 4th grade in primary school.* [Unpublished master's thesis]. Karadeniz Technical University.

Kazeni, M.M.M. (2005). *Development and validation of a test integrated science process skills for the further education and training learners* [Unpublished master's thesis]. University of Pretoria, South Africa.

Lincoln, Y.S., & Guba, E.G. (1985). Establishing trustworthiness. *Naturalistic inquiry, 289*(331), 289-327.

Ministry of National Education. (2018). *Elementary and middle school (3, 4, 5, 6, 7, and 8th grades) science curriculum*. Board of Education and Training.

Ongowo, R.O., & Indoshi, F.C. (2013). Science process skills in the Kenya certificate of secondary education biology practical examinations. *Procedia-Social and Behavioral Sciences, 4*(11), 713-717.

Ostlund, K.L. (1992). *Science process skills assesing hands-on student performance*. Addison-Wesley.

Rezba, R.J., Sprague, C.R., McDonnough, J.T., & Matkins, J.J. (2007). *Learning and assessing science process skills*. Kendall/Hunt Publishing Company.

Smith, K.A., & Welliver, P.W. (1994). *Science process assessments for elementary and middle school students*. Smith and Welliver Educational Services. http://www.scienceprocesstests.com/

Tan, Ş. (2009). *Öğretimde ölçme ve değerlendirme KPSS el kitabı (3rd* ed.) *[Assessment and evaluation in Teaching KPSS handbook (3rd ed.]*. Pegem Akademi Yayıncılık.

Tan, Ş., Kayabaşı, Y., & Erdoğan, A. (2002). *Planning and evaluation of teaching*. Anı Yayıncılık.

Tan, M., & Temiz, B.K. (2003). *Fen öğretiminde bilimsel süreç becerilerinin yeri ve önemi [The importance and role of the science process skills in science teaching]. Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 13*(13), 89-101.

Tatar, N. (2006). *The effect of inquiry-based learning approaches in the education of science in primary school on the science process skills, academic achivement and attitude.* [Unpublished doctoral dissertation]. Gazi University, Ankara.

Tobin, K.G., & Capie, W. (1982). Development and validation of a group test of integrated science processes. *Journal of research in Science Teaching, 19*(2), 133-141.

Walsh, D., & Downe, S. (2005). Meta-synthesis method for qualitative research: a literature review. *Journal of Advanced Nursing, 50*(2), 204-211.

Yıldırım, C. (1983). *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education]. ÖSYM Eğitim Yayınları*

APPENDIX

Evaluation Form for Science Process Skills Tests.

				Measured Science Process Skills and Question Numbers												
Thesis ID	Adaptation	Development	Question Type	Observing	Measuring	Data Recording	Founding Space and Number Relationship	Making Prediction	Determining Variables	Interpreting Data	Inferring	Formulating Hypothesis	Using Data and Formulating Models	Designing-Making Experiments	Controlling Variables	Decision-Making

Table 7. *The reviewed master's theses and doctoral dissertations.*

No	Percentage (%)
1	Acar, E.N. (2011). *The effect of project-based learning on scientific skill processes and attitudes towards biology of science teacher candidates*. [Unpublished master's thesis]. Çanakkale Onsekiz Mart University.
2	Akar, Ü. (2007). *The relationship between student teachers' scientific process skills and critical thinking*. [Unpublished master's thesis]. Afyon Kocatepe University.
3	Aktamış, H. (2007). *The effects of scientific process skills on scientific creativity: the example of primary school seventh grade physics*. [Unpublished doctoral dissertation]. Dokuz Eylül University.
4	Aktaş, S. (2016). *The effect of middle school 6th, 7th and 8th science teaching curriculum programs on the students' cognitive styles, emotional intelligent, science process skills and academic achievement*. [Unpublished master's thesis]. Mustafa Kemal University.
5	Altunsoy, S. (2008). *The effect of the inquiry-based learning approach on students? science process skills, academic achievements and attitudes in secondary biology teaching*. [Unpublished master's thesis]. Selçuk University.
6	Arslan, A. (2013). *The examination of pre-service teachers' science process skills and conceptual change in inquiry and model based inquiry environment.* [Unpublished master's thesis]. Marmara University.
7	Aydınlı, E. (2007). *Evaluation of science process skill study on the 6,7 and 8. Students*. [Unpublished master's thesis]. Gazi University.
8	Aydoğdu, B. (2006). *Identification of variables effecting science process skills in primary science and technology course*. [Unpublished master's thesis]. Dokuz Eylül University.
9	Aydoğdu, B. (2009). *The effects of different laboratory techniques on students' science process skills, views towards nature of science, attitudes towards laboratory and learning approaches in science and technology course*. [Unpublished doctoral dissertation]. Dokuz Eylül University.
10	Bahadır, H. (2007). *The effect of elementary science education based on scientific method process on science process skills, attitude, academic achievement and retention*. [Unpublished master's thesis]. Hacettepe University.
11	Başdaş, E. (2007). *The effect of hands-on science learning method in the education of science in primary school on the science process skills, academic achievement and motivation.* [Unpublished master's thesis]. Celal Bayar University.
12	Bayrak, B. (2011). *The effect of problem based instruction supported with web technology on the academic achievement, conceptual understanding, scientific process skills of 8th grade students in science and technology instruction: Acid base sample.* [Unpublished doctoral dissertation]. Marmara University.
13	Bilen, K. (2009). *The effects of a laboratory instruction designed based on the 'predict-observation-explain' strategy on preservice teachers? on conceptual achievement, science process skills, attitudes and views about the nature of science*. [Unpublished doctoral dissertation]. Gazi University.
14	Birinci, E. (2008). *The effect of using project-based learning in the adaptation and development of materials on teacher candidates? critical thinking, creative thinking and scientific-process skills*. [Unpublished master's thesis]. Karaelmas University.
15	Bodur, Z. (2015). *The effect of outdoor class activities in the solar system and beyond unit on seventh grade students' academic achievements, scientific process abilities and motivation.* [Unpublished master's thesis]. Marmara University.
16	Bozkurt, E. (2014). *The effect of engineering design based science instruction on science teacher candidates' decision making skills, science process skills and perceptions about the process*. [Unpublished doctoral dissertation]. Gazi University.
17	Cin, M. (2013). *Effects of concept cartoon activities based-argumentation method on students' conceptual understanding levels and scientific process skills*. [Unpublished master's thesis]. Dokuz Eylül University.

18	Çakar, E. (2008). *Determination of the level of students' achievement of the science process skills acquisition of 5th-grade science and technology program*. [Unpublished master's thesis]. Süleyman Demirel University.
19	Çelik, S. (2009). *The influence of project based learning approach on pre-service science teachers? conceptions of the nature of science and technology and scientific process skills*. [Unpublished doctoral dissertation]. Atatürk University.
20	Çelik, K. (2012). *The effect of inquiry based learning method for the teaching of reproduction, growth and development in the living things unit on the students' academic achievements, science process skills and attitudes toward science and technology course*. [Unpublished master's thesis]. Dokuz Eylül University.
21	Çelik, P. (2013). *The effect of problem based learning on pre-service teachers' physics course achievement, learning approaches and science process skills*. [Unpublished doctoral dissertation]. Dokuz Eylül University.
22	Çetin, A. (2013). *Mode-method interaction: The effects of inquiry vs. expository and blended vs. face-to-face instruction on 9th grade students' achievement in, science process skills in and attitudes towards physics*. [Unpublished doctoral dissertation]. Middle East Technical University.
23	Çınar, B. (2016). *The effect of enriching science and technology education by using stories which include historical process of scientific improvement on attitudes to science, the image of scientist, skills of scientific process and the academic achievement*. [Unpublished master's thesis]. Sakarya University.
24	Çoban, G.Ü. (2009). *The effects of model based science education on students' conceptual understanding, science process skills, understanding of scientific knowledge and its domain of existence: The sample of 7th grade unit of light*. [Unpublished doctoral dissertation]. Dokuz Eylül University.
25	Çümen, V. (2018). *The analysis effect of GEMS based learning program's on 6th grade student's achievement about concept of densities, conceptual changes and scientific progress*. [Unpublished master's thesis]. Uşak University.
26	Daşdemir, İ. (2012). *The effect of using of animation on students? academic achievement, retention of learned knowledge and scientific process skills*. [Unpublished doctoral dissertation]. Atatürk University.
27	Demir, M. (2007). *The factors affecting the pre-service primary teachers' adequacies on science process skills*. [Unpublished doctoral dissertation]. Gazi University.
28	Demirçalı, S. (2016). *The effects of model based science education on students' academic achievement,scientific process skills and mental model development: the sample of 7th grade unit of "The Solar System and Beyond: The Puzzle of Space"*. [Unpublished doctoral dissertation]. Gazi University.
29	Demirezen, S. (2010). *The effect of 7e model to students achievement, development of scientific process skills, conceptual achievement and retention levels in electrical circuits subject*. [Unpublished doctoral dissertation]. Gazi University.
30	Duran, M. (2008). *The effects of scientific process skills in science teaching on students' attitudes towards science*. [Unpublished master's thesis]. Muğla Sıtka Koçman University.
31	Elbistanlı, A. (2012). *Investigation of the effect of problem based learning approach on the achievement, attitude and scientific process skills of 11. grade students through chemical equilibrium subject*. [Unpublished master's thesis]. Mustafa Kemal University.
32	Elmacı, S. (2015). *Investigation of class teachers' process skills in scope of a number of variables*. [Unpublished master's thesis]. Dumlupınar University.
33	Ercan Ö.T. (2010). *İlköğretim yedinci sınıf fen ve teknoloji dersinde 5E öğrenme halkası ve bilimsel süreç becerileri doğrultusunda uygulanan etkinliklerin, öğrencilerin akademik başarıları, bilimsel süreç becerileri ve derse yönelik tutumlarına etkisi*. [Unpublished doctoral dissertation]. Ege University.
34	Erdoğan, M. (2010). *Effect of experiment techniques of group and demonstration to students' scientific process abilities, achievement and the ability of recalling*. [Unpublished master's thesis]. Selçuk University.

35	Erentay, N. (2013). *The effect of nature based outdoor activities upon the science knowledge levels, scientific process skills and attitudes towards environment of the fifth grade students*. [Unpublished master's thesis]. Akdeniz University.
36	Eroğlu, G. (2015). *Determination of science process skills of teacher candidates in the field of science.* [Unpublished master's thesis]. Gazi University.
37	Ertek, Y. (2014). *Investigation of the relationship between scientific process skills and problem solving skills stated in the physics curriculum*. [Unpublished master's thesis]. Ankara University.
38	Erten, N. (2013). *An investigation of primary teacher's science process skills in terms of some variables.* [Unpublished master's thesis]. Afyon Kocatepe University.
39	Gençoğlan, D.M. (2017). *The effects of argumentation based science learning (ABSL) approach based on authentic case studies on the success, attitude and scientific process skills of 8th grade students in the acids and bases lesson*. [Unpublished master's thesis]. Kahramanmaraş Sütçü İmam University.
40	Gök, G. (2014). *The effect of 7e learning cycle instruction on 6th grade students' conceptual understanding of human body systems, self-regulation, scientific epistemological beliefs, and science process skills*. [Unpublished doctoral dissertation]. Middle East Technical University.
41	Güçlüer, E. (2012). *The effect on success, attitude and scientific process skills of the use of scientific literacy developing activities in the unit named "systems of our body" in science and technology course*. [Unpublished doctoral dissertation]. Dokuz Eylül University.
42	Güden, C. (2015). *Examining secondary school students' cognitive process skills and attitudes towards science and technology course (Çanakkale sample)*. [Unpublished master's thesis]. Çanakkale Onsekiz Mart University.
43	Gülay, A. (2012). *Effect of self regulated learning on 5th grade students' academic achievement and scientific process skills*. [Unpublished master's thesis]. Recep Tayyip Erdoğan University.
44	Güler, Z. (2010). *The relationship among elementary students' test scores of level determination exam, course achievements, science processing skills and logical thinking skills*. [Unpublished master's thesis]. Abant İzzet Baysal University.
45	Güler, B. (2013). *The effect of blended learning method on preservice elementary science and technology teachers' attitudes toward technology and self-regulation and science process skills*. [Unpublished master's thesis]. Dokuz Eylül University.
46	Gültekin, Z. (2009). *The effect of project based learning applications on the students' views about the nature of science, science process skills and the attitude of students in science education*. [Unpublished master's thesis]. Marmara University.
47	Güney, T. (2015). *The effect of simulation aided science laboratory applications based on inquiry on science process skill: An example of the force and motion unit*. [Unpublished master's thesis]. Kırıkkale University.
48	Güngör, S.N. (2016). *The influence of teaching biological subjects and concepts to pre-science teachers through predict-observe-explain (POE) method on achievement, permanence, and scientific process skills*. [Unpublished doctoral dissertation]. Uludağ University.
49	Hazır, A. (2006). The fifth-grade primary school students' the level of acquisition of science process skills. [Unpublished master's thesis]. Afyon Kocatepe University.
50	İpek, Y. (2010). *Investigating scientific process skills development at science and technology course*. [Unpublished master's thesis]. Yüzüncü Yıl University.
51	Kandemir, E.M. (2011). *Determination of the level of teachers' understanding of integrated scientific process skills*. [Unpublished master's thesis]. Ege University.
52	Kanlı, U. (2007). *The effects of a laboratory based on the 7e model with verification laboratory approach on students? development of science process skills and conceptual achievement.* [Unpublished doctoral dissertation]. Gazi University.

53	Kaplan, M. (2016). *The effect of the differentiated method on seventh graders' conceptual learning, scientific process skills and academic achievement in the science unit "Force and movement"*. [Unpublished master's thesis]. Dokuz Eylül University.
54	Kara, E. (2017). *Research of the effects of science education based on predict - observe - explain strategy on 5th grade middle school students' science process skills and success*. [Unpublished master's thesis]. Marmara University.
55	Karaca, D. (2011). *Effect of the use of science writing heuristic (SWH) in General Physics Laboratory-I lesson on teacher candidates' achievement and scientific process skills*. [Unpublished master's thesis]. Mehmet Akif Ersoy University.
56	Karademir, E. (2009). *The effect of computer supported education towards students' academic success levels, scientific process skills and attitudes in the electric unit of science and technology lesson*. [Unpublished master's thesis]. Eskişehir Osmangazi University.
57	Karapınar, A. (2016). *The impact of inquiry-based learning environment on scientific process skills, inquiry skills and scientific reasoning skills of pre-service teachers*. [Unpublished master's thesis]. Celal Bayar University.
58	Karar, E.E. (2011). The study of science process skills of eighth graders with regard to several variant. [Unpublished master's thesis]. Adnan Menderes University.
59	Karatay, R. (2012). *Developing a science process skills test regarding the units of the 7th grade Science and Technology education program*. [Unpublished master's thesis]. Çanakkale Onsekiz Mart University.
60	Karatekin, P. (2012). *Effect of candidate teachers of Science and Technology at biology laboratories on TGA method on students? success, attitude and scientific process abilities*. [Unpublished master's thesis]. Celal Bayar University.
61	Karslı, F. (2011). *The effect of enriched laboratory guide materials on improving science process skills and conceptual change of prospective science teachers*. [Unpublished doctoral dissertation]. Karadeniz Technical University.
62	Kartal Taşoğlu, A. (2009). *The effect of problem based learning on students? Achievements, scientific process skills and attitudes towards problem solving in physics education*. [Unpublished master's thesis]. Dokuz Eylül University.
63	Keçeci, G. (2014). *The effects of inquiry-based science teaching on students' science process skills and attitudes*. [Unpublished doctoral dissertation]. Fırat University.
64	Keskinkılıç, G. (2010). *The effect of reflective thinking based learning activities in 7th class science and technology lesson on the students' achievements and their scientific process skills*. [Unpublished doctoral dissertation]. Selçuk University.
65	Kılıç, A.S. (2015). *The impact of the activities prepared through the integration of science and mathematics on the critical thinking and science process skills of the gifted 6th grade secondary school students.* [Unpublished doctoral dissertation]. Gazi University.
66	Kırıktaş, H. (2014). *The effect of inquiry based science teaching on pre-servive science teachers' academic achievement, science process skills and attitudes towards biology laboratory practice*. [Unpublished master's thesis]. Dokuz Eylül University.
67	Kırılmazkaya, G. (2014). *The effects of web based inquiry science teaching development on preservice teachers concept learning and scientific process skills*. [Unpublished doctoral dissertation]. Fırat University.
68	Kırtak Ad, V.N. (2016). *The effect of full studio model on pre-service primary science teachers' conceptual understanding, social emotional learning, inquiry and science process skills: An example of fluid mechanics*. [Unpublished doctoral dissertation]. Balıkesir University.
69	Kocagül, M. (2013). *The effect of inquiry based professional development activities on elementary Science and Technology teachers science process skills and self-efficacy and inquiry based teaching beliefs*. [Unpublished master's thesis]. Dokuz Eylül University.
70	Korucuoğlu, P. (2008). *Evaluation of correlation between scientific process skills' usage level of physics teacher candidates with the attitudes towards physics, gender, class level and high school type which they graduated from*. [Unpublished master's thesis]. Dokuz Eylül University.

71	Kozcu Çakır, N. (2013). *The science process skills of pre-service science teachers' qualitative and quantitative analysis*. [Unpublished doctoral dissertation]. Gazi University.
72	Köksal, E.A. (2008). *The acquisition of science process skills through guided (teacher-directed) inquiry*. [Unpublished doctoral dissertation]. Middle East Technical University.
73	Kula, G. (2011). *The effect of pre-school education on 1st, 2nd and 3rd grade high school students' science process skills: the sample of Polatlı district.* [Unpublished master's thesis]. Gazi University.
74	Kula, Ş.G. (2009). *The effect of inquiry-based science learning on the students' science process skills, achievement, concept learning and attitude*. [Unpublished master's thesis]. Marmara University.
75	Kurtuluş, N. (2012). *The effect of instructional applications based on creative thinking on scientific creativity, scientific process skills and academic achievement*. [Unpublished master's thesis]. Karadeniz Technical University.
76	Meşeci, B. (2013). *The gaining of scientific skills and the effectiveness the learning process of the unit particulated structure of the matter*. [Unpublished master's thesis]. Amasya University.
77	Mutlu, S. (2012). *The effects of science and technology education based on scientific process skills on scientific process skills, motivation, attitude, and achievement of elementary school students*. [Unpublished master's thesis]. Trakya University.
78	Önol, M. (2013). *The impact of creative problem solving activities on scientific process skills and success*. [Unpublished master's thesis]. Balıkesir University.
79	Özahioğlu, B. (2012). *The effect of project-based learning in Science and Technology at primary school on scientific process skills, success and attitude*. [Unpublished master's thesis]. Çanakkale Onsekiz Mart University.
80	Özdemir, H. (2009). *The level of having scientific process skills of 5th class primary school's students (Afyonkarahisar sample)*. [Unpublished master's thesis]. Afyon Kocatepe University.
81	Özdemir, G. (2017). *An action research about enriched curriculum towards the contribution to scientific process skills and achievement for gifted students*. [Unpublished master's thesis]. Hacettepe University.
82	Özdoğru, E. (20139. *The effect of Lego programme based science and technology education on the students' academic achievement, science process skills and their attitudes toward Science and Technology course for physical facts learning field*. [Unpublished master's thesis]. Dokuz Eylül University.
83	Özer, D.Z. (2011). *The effect of project based learning approach on both the academic achievement and the development of science process skills of prospective teachers of science education department towards biology lesson*. [Unpublished doctoral dissertation]. Uludağ University.
84	Özkan, D.O. (2011). *The effect of using V-diagrams in 'living things and energy relations' unit's experiments in eighth grade science and technology lessons on students' achievements, science process skills and attitudes*. [Unpublished master's thesis]. Gazi University.
85	Öztürk, N. (2008). *Primary seventh-grade students' level of gaining science process skills in science and technology course*. [Unpublished master's thesis]. Eskişehir Osmangazi University.
86	Öztürk, Ç. (2008). *The effects of 5e model on the scientific process skills, academic achievement and attitude towards the geography course*. [Unpublished doctoral dissertation]. Gazi University.
87	Öztürk, A. (2014). *The effects of curricula at Mevlana Public and Science Center on students' science process skills and attitudes toward science*. [Unpublished master's thesis]. Ege University.
88	Parim, G. (2009). *The effects of inquiry on the concept learning, achievement and development of scientific process skills of 8th grade students as related to photosynthesis and respiration*. [Unpublished doctoral dissertation]. Marmara University.

89 Recepoğlu, B. (2012). *The effect of the open- ended experimental technique on science process skills, scholar success and attitude toward biology*. [Unpublished master’s thesis]. Çanakkale Onsekiz Mart University.

90 Sabır, A. (2016). *The investigate of affecting factors on the science process skills of 4th and 5th grade students*. [Unpublished master’s thesis]. Mustafa Kemal University.

91 Savaş, E. (2011). *Effect of the science process skills laboratory approach supported with peer instruction on science process skills of teacher candidates*. [Unpublished master’s thesis]. Balıkesir University.

92 Sedef, A. (2012). *The effect of creative drama activities on elementary school seventh grade students scientific process skills, scientific creativity and self-regulation*. [Unpublished master’s thesis]. Pamukkale University.

93 Serin, G. (2009). *The effect of problem based learning instruction on 7th grade students' science achievement, attitude toward science and scientific process skills*. [Unpublished doctoral dissertation]. Middle East Technical University.

94 Sevinç, E. (2008). *The effects of the 5E model on the students' conceptual understanding, the development of their scientific process skills and their attitude in the organic chemistry laboratory course*. [Unpublished master’s thesis]. Gazi University.

95 Şahbaz, Ö. (2010). *The effects of different methods on students' science process skills, problem solving skills, academic achievements and retentions in primary school fifth grade science and technology lessons*. [Unpublished doctoral dissertation]. Dokuz Eylül University.

96 Şardağ, M. (2013). *A study of test development to measure science process skills of 8th grade students.* [Unpublished master’s thesis]. Balıkesir University.

97 Şen, A.Z. (2011). *Examining 12th grade high school students' science process skills levels*. [Unpublished master’s thesis]. Balıkesir University.

98 Şencan, D. (2013). *The effects of real-world problems on the 7th grade students' scientific process abilities, academic achievement and scientific literacy: Force and motion*. [Unpublished master’s thesis]. Marmara University.

99 Tavukçu, F. (2008). *The effect of computer-assisted learning environment in science education on the success, science process skills and attitudes towards computer use of students*. [Unpublished master’s thesis]. Karaelmas University.

100 Temiz, B.K. (2001). *Investigation of the appropriateness of the 9 th grade physics curriculum on the progression of the students scientific process skills*. [Unpublished master’s thesis]. Gazi University.

101 Temiz, K.B. (2007). *Assessing Science Process Skills in Physics Teaching*. [Unpublished doctoral dissertation]. Gazi University.

102 Tezcan, G. (2011). *Developing a science process skills test regarding the units of the 6th grade science and technology education program*. [Unpublished master’s thesis]. Çanakkale Onsekiz Mart University.

103 Topkara, F. (2010). *Anatolian high school students, high school of science and net for the entrance examination, their attitudes towards physics course, the relationship between academic achievement and the scientific process skills: the case of district in Ankara, Elmadağ*. [Unpublished master’s thesis]. Gazi University.

104 Toprak, F. (2011). *The effects of 3E and 5E teaching models practiced in general chemistry laboratory of science education on students' academic success scientific process skills and their attitude to the course*. [Unpublished master’s thesis]. Ondokuz Mayıs University.

105 Türker, E. (2011). *To investigate how scientific process skills approach based on model using affect the students? Success, development of process skills and motivations*. [Unpublished master’s thesis]. Karadeniz Technical University.

106 Usta Gezer, S. (2014). *The effects of reflective inquiry based general biology laboratory activities' on preservice science teachers' laboratory self-efficacy perceptions, critical thinking tendencies and scientific process skills*. [Unpublished doctoral dissertation]. Marmara University.

107	Uzun, F. (2013). *The effect of the general physics-I laboratory course based on context-based approach on preservice science teachers' achievement, scientific process skills, motivation and recall*. [Unpublished master’s thesis]. Marmara University.
108	Ünaldı, Ö. (2012). *Effect of the scientific process skills-based science education on students' science attitudes and scientific process skills*. [Unpublished master’s thesis]. Ankara University.
109	Yalçın, T. (2014). *The effect of inquiry based learning method on students' scientific process skills and conceptual understanding*. [Unpublished master’s thesis]. Dokuz Eylül University.
110	Yaprakdal, A.B. (2013). *The effect of learning objects design on the critical and creative thinking capacities and science process skills of prospective teachers*. [Unpublished doctoral dissertation]. Marmara University.
111	Yavuz Şahin, S. (2009). *The contribution of development science process skills that been consist at implementation process in the unit of human and environment grade 7 in the primary science and technology curriculum*. [Unpublished master’s thesis]. Balıkesir University.
112	Yıldırım, M. (2011). *Interrelationships of scientific process skills*. [Unpublished master’s thesis]. Atatürk University.
113	Yıldırım, A. (2012). *Effect of guided inquiry experiments on the acquisition of science process skills, achievement and differentiation of conceptual structure*. [Unpublished master’s thesis]. Middle East Technical University.
114	Yıldız, N. (2010). *The effect of experiment applications on the success, attitude and scientific process abilities of the students in the solution of the learning scenarios based on problems in science educatio*n. [Unpublished master’s thesis]. Marmara University.
115	Yılmaz, F.N. (2015). *The effect of project based learning method on the 6th graders' achievement and scientific process skills in science education*. [Unpublished master’s thesis]. Pamukkale University.
116	Yırtıcı, Z. (2014). *Impact of optional courses of scientific application on students’ scientific process skills and motivations towards science*. [Unpublished master’s thesis]. Gazi University.

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 1, 39–60

https://doi.org/10.21449/ijate.1058302

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

Investigation of the Effectiveness of the Research Skills Teaching Program

Betul Polat [1,*], **Omer Kutlu** [2]

[1]Nigde Omer Halisdemir University, Faculty of Education, Department of Educational Sciences, Nigde, Turkiye
[2]Ankara University, Faculty of Educational Sciences, Department of Educational Sciences, Ankara, Turkiye

ARTICLE HISTORY

Received: Jan. 06, 2021
Revised: Nov. 12, 2021
Accepted: Jan. 13, 2022

Keywords:
Research skill,
Effectiveness,
Primary school,
Teaching program,
Evaluation

Abstract: In this study, it was investigated whether the Research Skills Teaching Program (RSTP) prepared for elementary school 4th grade students was effective in imparting the sub-skills required for conducting the research process to students. To this end, pretest-posttest control group design; one of the quasi-experimental designs; was employed. The study was conducted on the students attending two classes of a state elementary school. In the experimental group, the program developed on the basis of the Big Six Research Skills Model was administered, while no such special application was conducted in the control group. In both of the groups, the Research Skills Test was administered as pretest and posttest and Monitoring Tests. The collected data were analyzed by using Independent Samples *t*-test and Paired Samples *t*-test. The responses given by the experimental and control group students to each of the open-ended items in the Monitoring Tests were separately examined. It was found that the activities developed within the context of the teaching program had comprehensive effect on the development of the skills needed to direct the research process. When the number of the students giving the most correct answers in the items in the monitoring tests was examined, it was seen that although there are students finding the most correct answers to the items in the experimental group, their levels of using these strategies were not found to be at the desired level and there are students in the control group finding the most correct answers only for four skills.

1. INTRODUCTION

The developments in technology lead to occurrence of rapid and important changes in social life. This reality caused our age to be called the information age and forced individuals to be involved in lifelong learning. In this era where knowledge is considered a product of wealth, the perspective of the nature of learning has changed and the concept of lifelong learning has come to the fore (Berber, 2003).

Lifelong learning is a process that stimulates and activates individuals to use their knowledge, skills, values and understanding when needed in their real life (World Initiative on Lifelong Learning [WILL], cited in Candy, 2003). According to Cambridge (2010), lifelong learning is a process that includes all kinds of learning actions that allow individuals to develop their knowledge, skills and strategies by recognizing themselves and exhibiting their talents throughout their lives. In this connection, lifelong learning includes skills and features that

*CONTACT: Betul Polat ✉ betul.polat6006@gmail.com Nigde Omer Halisdemir University, Faculty of Education, Department of Educational Sciences, Nigde, Turkiye

e-ISSN: 2148-7456 **/© IJATE** 2022

enable the individual to gain new knowledge and acquire new skills both in personal and business life (Demirel, 2007).

Changing perspective of the phenomenon of learning has also changed the characteristics of the individuals needed by the information societies. In this regard, today's information societies are not in need of individuals who store information that will lose its actuality in a short time; yet, they need individuals who can have access to new information properly and quickly and use the information they have learned effectively in every area of their lives (Kutlu et al., 2017).

Changing desired characteristics of individuals in information societies forced student characteristics to be reviewed and changed. Students' being lifelong learners and successful in life depends on their gaining the ability to use the basic knowledge and skills they have acquired during their school life in the real life (Berberoğlu, 2006). In this regard, in developed societies, not the individuals who use information as it is rather the individuals who do research, think critically, creatively, solve problems, know themselves, are confident, can use what they have learned in real life; that is, the individuals having higher order mental skills are considered to be successful (Kumandaş & Kutlu, 2013). As a result of the handling of the qualities that a successful individual should have in different dimensions, the concept of student achievement has also been expressed differently. The concept of student achievement, defined as the level of achieving the goals and behaviours in the curriculum for many years, has been defined as the power to learn the basic knowledge and skills and to use them in the life situations they encounter by associating them with their own individual characteristics since 1980s. Therefore, higher order mental skills that individuals need to acquire in order to be successful in life are the unity of cognitive, affective and kinetic characteristics that the individual uses while demonstrating his/her ability (Kutlu et al., 2017).

Higher order mental skills require learning by understanding rather than rote-learning, using information, solving problems related to new situations, making explanation, synthesis and generalization and using the hypothesis formulation skill (Üstünoğlu, 2006). Higher order mental skills include the skills of asking questions, doing research, conducting critical, reflective, logical, systematic and creative thinking, solving problems, thinking analytically, making evaluations and producing new information (King et al, 1998; Zoller, 2000). These thinking skills become activated when students face problems, uncertainties, questions or dilemmas. For this reason, in order to impart higher order mental skills to students, appropriate learning environments should be prepared at schools (Aksu, 2005), students should be provided with activities based on thinking and questions that require students to think should be posed (Beyer, 1987).

Inculcation of higher order mental skills in students in elementary school, which is one of the most important steps of education for application individuals who are equipped with the higher order mental skills required by societies and who can adapt to developments, has become one of the most important goals of educational institutions (Kutlu et al, 2010). It is only possible for students to continue learning on their own if they are equipped with higher order mental skills (Doğanay, 2008). In this context, it has been attempted to include higher order mental skills in curriculums from the elementary education onward and to impart them to students (Milli Eğitim Bakanlığı [MEB], 2006).

Rapid increase in the amount and dissemination of information in today's information age has given rise to the question "What is the way of having access to information?" In order to find an answer to this question, more emphasis has been placed on the necessity of gaining research skills which are seen as one of the main characteristics of contemporary societies and which are one of the higher order mental skills (Alkan, 1989; Shuman et al., 2005). Research skills are considered as one of the most important basic life skills that 21st century learners should have. Research skills include skills that require access to reliable and qualified information from different sources, to present this information effectively by bringing it together, and to direct the research process accurately and systematically (Polat-Demir & Kutlu, 2016).

The research process, which consists of many sub-skills, is defined in different ways in the literature. The American Association of School Librarians [AASL] (2007) defines them as skills that enable students to create new insights, achieve results and produce new information by guiding the research process so that they can understand, learn and master the topics. Abston et al. (2004), on the other hand, defined them as the ability to research information carefully and systematically to investigate and identify a phenomenon or principle.

According to Bird (2000), students will learn most of the skills that make up the research process at school. Therefore, there are many important duties to be assumed by teachers and parents in this process. Bird (2000) discusses the basic skills of the research process under three headings described below. These are researching, evaluating and note taking.

Researching: Doing research is not just finding and extracting information, but knowing where and how to find it. For example, the student needs to know how to obtain information from books.

Evaluating: When students read a book, they need to finds answers to such questions as "How correct is the information included in the book?", "Does the book present the desired information?", "Does the book include some biases?" Elementary school students may find it difficult to make such an inquiry, but they can make a start on it. Teachers and parents should encourage students to ask these questions about the use of resources. Encouraging students to ask questions will help them gain research skills.

Note taking: Students should be taught simple note-taking skills such as drawing pictures, maps and plans, writing descriptions, noting measurements, instructions and plans, taking pictures, and writing notes from the book.

When these three basic skills are examined, it will be seen that students can learn these skills easily at school. When students are given tasks that require the use of research skills, it is important that parents support their children in the development of these skills. In this context, it can be possible to train individuals equipped with the skills that make up the research processes through a qualified education approach from an early age (İlter, 2013; Numanoğlu, 1999). Although a great emphasis is put on the importance of imparting research skills particularly to elementary school students and they are included in curriculums, it has been reported that students are fall short in conducting research (Alkan-Dilbaz, 2013; Chu et al., 2008; Polat-Demir & Kutlu, 2016; 2017). This inadequacy also affects students' later learning process and reduces students' tendency and interest in doing research (Knutson, Dozier & Migotsky, 1995). Students who do not have sufficient knowledge and skills in conducting research find measurement activities such as projects and performance tasks that require the use of these skills in different situations harder than they really are. Therefore, such activities are seen as a waste of time by teachers, parents and students.

The importance of imparting research skills to students from a young age and the necessity of preparing appropriate learning environments for this to happen are known (Alkan-Dilbaz, 2013; Chu et al, 2008; Güneş, 2011; Polat-Demir, 2016; Wu & Hsieh, 2006; Yıldırım, 2007). Nevertheless, it is noteworthy that the studies for the development of research skills are not sufficient and that there is no application program that will contribute to the development of these skills. This study was carried out to contribute to the elimination of this deficiency. To this end, it has been investigated whether the Research Skills Teaching Program (RSTP) prepared for elementary school 4th grade students is effective in imparting the sub-skills that make up the research process to the students. In this connection, answers to the following questions were sought:

1. Is there any significant difference between the pretest and posttest mean scores taken from the Research Skills Test by the experimental group and the control group?
2. Is there any significant difference between the scores taken from the monitoring tests by the experimental group students subjected to the Research Skills Teaching Program and the control group students not subjected to this program?
3. What is the number of the most correct answers obtained by the experimental and control group students for the open-ended items in the monitoring tests used in the Research Skills Teaching Program?

The teaching program prepared within the context of the current study is thought to guide teachers, experts and families about how the skills that make up the research processes should be imparted to students. In this context, activities and worksheets prepared within the context of the program are important in terms of providing examples on how to impart research skills to students. In the current research, the level of the students' use of research skills, which are one of the higher order mental skills, was determined with tests consisting of open-ended items. These items prepared within the scope of the current research are thought to provide examples for both practitioners and researchers for measuring higher order mental skills. For this reason, this research is also important in terms of providing examples about the tools that can be used to measure higher order mental skills and the preparation of these tools.

2. METHOD

2.1. Research Model

In the current study, it was attempted to reveal the cause and effect relationship between the independent variable and dependent variable. For this purpose, the Research Skills test was applied as a pre-test to six different classes in the study, and two classes that were equivalent in terms of these skills were included in the study. One of the groups was randomly assigned as the experimental group and the other as the control group. In the study, the pre-test - post-test paired control group design, which is one of the semi-experimental designs, was used because the groups were studied on ready groups and group matching was made (Büyüköztürk et al., 2009).

2.2. Study Group

The study group of the current research is comprised of the 4th grade students attending a state school in the city of Nigde. A great care was taken to match students in terms of their socio-economic status and mother and father's education level and to do so, opinions of the school principle and classroom teachers were sought. The Research Skills Test developed by the researchers and explained in detail below was administered to the students in these classes as pretest. Independent samples *t*-test was run to determine whether there is a significant difference between the scores taken from the test by the students in these classes. As no significant difference was found between their scores, these two classes were selected as the study group. One of the two randomly selected classes was assigned to the experimental group and the other to the control group. The findings of the independent samples *t*-test are presented in Table 1

Table 1. *Results of the Independent Samples t-test conducted on the pretest scores*

Test	Group	N	$\bar{X}$	S_x	*df*	*t*	*p*
Pretest	Experimental group	34	15.12	7.13	68	.322	.748
	Control group	36	14.67	4.32			

The experimental group students' mean score taken from the pretest is $\bar{X}$= 15.12, while that of the control group students is $\bar{X}$= 14.67. The pretest mean scores of the students calculated out of 100 points show that the students in both groups have very little knowledge and very few skills on the research process. The results of the independent samples *t*-test revealed that there is no significant difference between the pretest mean scores of the experimental and control group students ($t_{(68)}$=.322; p>.05). This proves that the knowledge and skills possessed by the groups are equal to each other.

The current study was conducted with the participation of 70 students; 34 students in the experimental group, 36 students in the control group. Distribution of the students to experimental and control groups by gender is given in Table 2.

Table 2. *Distribution of the students to experimental and control groups by gender.*

Group	Gender	N	Total
Experimental group	Female	15	34
	Male	19	
Control group	Female	17	36
	Male	19	

As can be seen in Table 2, the number of male students in both the experimental and control groups is higher than that of the female students. Of the experimental students, 44.1% are females and 55.9% are males while in the control group, 47.2% are females and 52.8% are males.

2.3. Research Skills Teaching Program (RSTP)

The Research Skills Teaching Program (RSTP) was prepared on the basis of the Big Six Research Skills Model developed by Eisenberg and Berkowitz in 1987. Through this model, it was aimed to make the students conduct research by directing their research process (Eisenberg & Berkowitz, 1990). In the development process of the RSTP, the sub-skills of the research skills were determined considering the six processes in the Big Six Research Skills Model *(target – task definition, information seeking strategies, finding and having access, using knowledge, organizing knowledge, evaluation)*. A great care was taken for the objective statement to represent the mental level corresponding to both a specific content and the relevant skill. The stages in Bloom's updated cognitive classification were used in defining the “mental level” dimension regarding the objectives (Anderson et al., 2001).

A great care was taken for the compliance of the objectives with the sub-skills of the research skills, for them to be measurable and to be scientifically correct. Moreover, special importance was attached for expressions to be clear and understandable and unnecessary expressions were avoided. The clarity and understandability of the objectives, their compliance with the sub-skills constituting the research skills and their measurability were submitted to the review of three experts specialized in the fields of measurement and evaluation, curriculum development and language. The sub-skills and objectives of RSTP are given in Table 3. While creating the learning area / content, the literature was reviewed taking into account the relevant grade level. Teaching activities suitable for each objective have been designed by considering the learning area. Activities are associated with the objectives and topics of the 4th grade social studies course. The examples in the activities were prepared in accordance with the content of the social studies course. The social studies course was selected because its content is easy to relate to the objectives of the prepared program and research skills are handled within the scope of the curriculum of this course. While designing classroom activities, mainly *lecturing, question-answer, demonstration and allowing students to practice* and *brainstorming* teaching methods and techniques were used.

Table 3. *The Sub-skills and objectives of the Research Skills Teaching Program.*

Sub-skills	Objectives
Target (Task) Definition	1. Selects the research topic by dividing the topic into sub-topics. (Analysis) 2. Narrows the subject of the research and writes research questions in such a way as to allow him/her to investigate different aspects of the subject. (Creating)
Information Seeking Strategies	3. Selects suitable types of resources to reach the information he/she needs. (Analysis) 4. Determines key words suitable for the research questions. (Analysis)
Finding and Having Access	5. Has access to information by using search indices in the library catalogue. (Application) 6. Reaches the information he/she needs by using the contents list, directories list, guides and keywords in printed sources. (Application) 7. When searching the internet with keywords, it reaches the information by using different techniques (Boolean operators, +, -, "....." (quotation marks). (Application)
Using Information	8. Prepares a note card to record the information he/she reaches. (Application) 9. Prepares a bibliography card. (Application) 10. Apply the rules of quoting and referencing. (Application)
Organizing Information	11. Determines the outline of a research. (Analysis) 12. Organizes the information in a manner suitable for the outlines in the research report. (Creating) 13. Prepares the cover page of the research report. (Application)
Evaluating	14. Evaluates the research report in terms of compliance with the report writing rules. (Evaluating)

Ten different teaching materials were prepared in accordance with the learning content and teaching activities in order to provide resources for the students about research skills and were distributed to the students in the experimental group along the process. In the teaching materials, examples and exercises related to the skills are also included. These examples and exercises were created taking into account the content of the social studies course.

A lesson plan was prepared for each sub-skill taking into account the objectives in the RSTP, learning area and teaching activities. Lesson plans were evaluated by four experts (two measurement and evaluation experts, one curriculum development expert and one social studies teaching expert) in terms of the suitability of the RSTP for the level of the students and its scientific accuracy.

The RSTP was applied to the experimental group students by researchers in order to improve their research skills throughout the research process. No such a special application was made to the control group. The application in the experimental group lasted five weeks, three class hours a week, thus a total of 15 class hours.

2.4. Data Collection Tools

The data in the current study were collected by using the Research Skills Test as pretest and posttest and the monitoring tests used along the process. The features of these tests are discussed below.

2.4.1. Research Skills Test

The Research Skills Test (RST) was developed by the researchers to measure students' research conducting skills. In the RST, there are 11 open-ended items to measure 13 objectives in the first five sub-skills addressed in the teaching program *(target-objective definition, information seeking strategies, finding and having access, using information, organizing information)*. Since the research skills include integrated skills, two objectives were tried to be measured together in some items. The RST includes a text and questions based on this text. In the text, the subject of "Technological Products" in the "Science, Technology and Society" learning area in the elementary school 4th grade social studies curriculum was taken into consideration. Information from different resources was collected and a text called "*Technology and Technological Products*" was prepared (Başdoğan, 2013; Daştan & Gürler, 2016; Mısırlı, 2007; Turam, 1999). The students were asked to design a research using the information provided in the text and respond to the items given in the test.

A great care was taken to ensure that open-ended items would be suitable for the student level. The items were expressed plainly and clearly, and unnecessary expressions were avoided. In order to prevent students from giving long and unrelated answers to the items, some restrictive guidance was given at the root of the item and a place was left for the response of each item.

The reliability of the items was determined through expert review. The opinions of four experts (two measurement and evaluation experts, one curriculum development expert and one social studies teacher) were sought to determine whether the items are related to the relevant objectives, whether there is any mistake from a scientific point of view, whether the items are suitable for the grade level and whether there is any mistake in terms of expression. In light of their feedbacks, required corrections were made and final form of the RST was given.

In the writing of open-ended items, care was taken to express the item root in a simple, short and clear manner and to be suitable for the level of the student. Necessary space is left for the student to write his/her answer under each item.

The validity of the items was determined based on expert opinion, and for this purpose, two assessment and evaluation, one Turkish language and one Social Studies teaching experts were consulted. The experts were asked whether the items were related to the relevant acquisition, whether they had a scientific error, whether they were appropriate for the grade level, and whether they contained errors in terms of language and expression characteristics. Experts were asked to evaluate whether each item was appropriate in line with these criteria and the percentages of agreement were calculated. The percentages of agreement for each item were calculated using the following reliability formula proposed by Miles and Huberman (1994).

$$Percentage\ of\ Reliability = \frac{Number\ of\ Consensus}{Number\ of\ Consensus + Number\ of\ Disagreement} x100$$

It was determined that the calculated agreement percentage was 75% for the 10th item and 100% for the other items. Accordingly, it was concluded that the validity of the RST was high and the RST was evaluated out of 100 points.

2.4.2. Monitoring Tests

Monitoring tests were prepared to be consisted of open-ended items to monitor the learning in each of the basic skills that make up the research process. The items were developed in such a way as to allow students to associate each skill with case studies that correspond to their real-life situations. In order to determine the suitability of the items to the student level, their relation with the objectives in the RSTP and their power to measure the related skill, the opinions of two classroom teachers and two measurement and evaluation experts were consulted and the items were finalized.

The items forming the monitoring tests were applied to the experimental and control groups after the completion of the learning phase related to each sub-skill. After the items were applied, they were immediately evaluated and the students were given feedback before proceeding to learning related to the next skill. The items were prepared based on the objectives in the relevant skill. Information about the characteristics of these items is given below:

“Target (Task) Definition” Monitoring Test 1: It consists of two items. Through these items, the students’ skills of selecting the research topic and writing research questions suitable for the topic they have selected were measured. The highest score to be taken from these items is 12.

“Information Seeking Strategies” Monitoring Test 2: It consists of two items. Through these items, the students’ skills of selecting the types of resources suitable for findings answers to the research questions and determining proper key words were measured. The highest score to be taken from these items is 15.

“*Finding and Having Access*” *Monitoring Test 3*: It consists of 4 items. Through these items, the students’ skills of using the indices in the library catalogue to have access to the information they need, using Boolean operators and different techniques while conducting an internet search and making use of the content list, directory list and guiding and key words in the printed resources were measured. The highest score to be taken from these items is 13.

“Using Information” Monitoring Test 4: It consists of three items. Through these items, the students’ skills of preparing bibliography and note card and citing the references within the text in accordance with the rules were measured. The highest score to be taken from the items is 15.

“Organizing Information” Monitoring Test 5: It consists of two items. Through these items, the students’ skills of determining the outline of the research report, organizing the information in compliance with the outline and preparing the cover page were measured. The highest score to be taken from these items is 29.

A rubric was prepared for each item in order to score responses to the open-ended items in the monitoring tests. While preparing the rubrics, five dimensions were defined as “*the Most Correct Answer*”, “*Distant Correct Answers*”, “*No Answer*” “*Wrong Answers*” and “*Other Answers*” for each item. *The most correct answer* is the sample answer that fully covers the characteristic measured by the item and does not have any deficiencies. *Distant correct answers* contain partial accuracy and are scored depending on their proximity to the most correct answer. *Wrong answers* are accurate in themselves, but they represent another characteristic, not the characteristic measured by the item. *Other responses* are those that do not have anything to do with the basic information learned within the context of the characteristic measured by the item. These answers can sometimes be fictitious and absurd.

2.5. Data Analysis

The data were analyzed by using SPSS 21.0 software. In the analysis of the data, the significance level was accepted to be α =0.05. In the analysis of the data, first it was checked whether the distributions of the scores taken by the experimental and control groups are normal. For this purpose, the skewness and kurtosis coefficients calculated for the distribution of the scores taken from the pretest and posttest by the experimental and control group students were examined. The skewness and kurtosis coefficients were found to be between -1 and +1. Thus, it was concluded that the pretest and posttest scores exhibited a normal distribution (Büyüköztürk, 2009). Then, Levene *F* test was run to test the homogeneity of the variances. The pretest score (F= 7.364, p < .008) and posttest score (F= 15.173, p <.000) found as a result of the Levene *F* test have revealed that the variances are not equal forbetween goups. As normality assumption was satisfied but homogeneity of the variances was not satisfied, Paired Samples *t*-test, one of the parametric test statistics, was used to test whether there is a significant difference between the pretest and posttest scores of the experimental and control group students while Independent Samples *t*-test was used to test whether there is a significant difference between the posttest scores of the experimental and control group students. In order to determine how effective the RSTP in developing research skills, eta-squared (η^2) and Cohen d values were calculated. The Eta-squared value (η^2) named as effect size showing how much of the total variance in the dependent variable is explained by the independent variable varies between 0.00 and 1.00. If the eta-squared value is in the range of 0.01≤ η^2<0.06, then it indicates "a small effect", in the range of 0.06≤ η^2<0.14, then it indicates "a medium effect" and in the range of 0.14≤ η^2, it indicates "a large effect". In order to determine whether the sample selected in the study represents the universe and whether the power of the research is sufficient, the power of the statistical test results was calculated and the GPower3.1 analysis program was used.

It was also investigated whether there is a significant difference between the scores taken from the tests consisted of monitoring items by the experimental and control group students. To this end, the responses given to the open-ended items in the monitoring tests by the experimental and control group students were scored and the scores taken from each monitoring test were separately analyzed. First, it was investigated whether the distribution of the scores is normal. For this purpose, skewness and kurtosis coefficients were examined and they were found to be between -1 and +1. On the basis of this finding, it was concluded that the scores taken from the monitoring tests showed a normal distribution. As the normality assumption was satisfied, Independent Samples *t*-test; one of the parametric test statistics, was used.

The responses given by the experimental and control group students to each of the open-ended items in the monitoring tests used in the application process of the RSTP were separately analyzed. The scores and numbers of the students reaching the most correct answer, distant correct answers and wrong answer were separately calculated for the experimental and control groups and are presented in tables.

3. RESULT / FINDINGS

In this section, findings obtained from the analysis of the data are organized and interpreted considering the research questions.

3.1. Findings related to Pretest and Posttest Scores of the Experimental and Control Group Students

In order to determine whether the pretest and posttest scores of the experimental and control group students varied significantly, Paired Samples *t*-test was used. Findings are presented in Table 4.

Table 4. *Paired Samples t-test results of the experimental and control groups.*

Group	Test	*N*	$\bar{X}$	S_x	*df*	*t*	*p*
Experimental Group	Pretest	34	15.12	7.13	33	10.277	.000*
	Posttest	34	42.62	17.24			
Control Group	Pretest	36	14.67	4.32	35	-.142	.880
	Posttest	36	14.53	6.31			

* p<0.05

As can be seen in Table 4, while there is a significant difference between the pretest and posttest mean scores taken from the RST by the experimental group students ($t_{(33)}$= 10.277, p<.05), there is no significant difference between the pretest and posttest mean scores of the control group students ($t_{(35)}$= -.142, p> .05). While the experimental group students' mean score was $\bar{X}$= 15.12 before the application, it increased to $\bar{X}$= 42.62 after the application. On the other hand, while the control group students' mean score before the application was $\bar{X}$= 14.67, it was found to be $\bar{X}$= 14.53 after the application, thus, no increase was observed. These findings show that there is a significant increase in the experimental group students' posttest mean score compared to their pretest mean score and that their level of using research skills significantly improved after the application.

Independent Samples *t*-test was used to check whether there is a significant difference between the posttest mean scores of the experimental and control group students. Findings are presented in Table 5.

Table 5. *Independent Samples t-test results for the posttest scores.*

Test	Group	*N*	$\bar{X}$	S_x	*df*	*t*	*p*	η^2
Posttest	Experimental group	34	42.62	17.24	68	9.150	.000*	0.552
	Control group	36	14.53	6.31				

*p<0.05

As can be seen in Table 5, there is a significant difference between the posttest mean scores of the experimental and control group students in favour of the experimental group students ($t_{(68)}$= 9.150, *p*<. 05). In the posttest, the students in the experimental group were found to be more successful than the students in the control group. These findings show that the application given had a positive effect on the development of the students' research skills. Eta-squared effect size value was found to be η^2= 0.552. These value show that the RSTP had a "large effect" on the development of the students' research skills. The power of the study was found to be 0.909. This finding shows that the power of the study conducted on 70 people is 90.9%.

3.2. Findings related to Monitoring Tests Scores of the Experimental and Control Groups

Independent-samples *t*-test was used to test whether there is a significant difference between the scores taken from the monitoring tests by the experimental and control group students. Findings are presented in Table 6.

Table 6. *Results of the Independent Samples t-test for the scores taken from the monitoring tests.*

Monitoring Test (MT)	Group	*N*	$\bar{X}$	S_x	*df*	*t*	*p*
MT1	Experimental group	33	8.88	2.83	64	6.629	.000*
	Control group	35	4.11	3.09			
MT2	Experimental group	33	8.42	3.08	65	6.979	.000*
	Control group	34	4.27	1.58			
MT3	Experimental group	32	8.75	2.45	61	14.296	.000*
	Control group	31	2.03	1.02			
MT4	Experimental group	32	10.44	3.75	62	13.240	.000*
	Control group	34	1.57	0.56			
MT5	Experimental group	33	14.70	4.38	66	10.434	.000*
	Control group	35	5.57	2.67			

*p<0.05

When the scores given in the table are examined, it is seen that the experimental group students received higher scores than the control group students in all the monitoring tests. When the results of the independent samples *t*-test are examined, it is seen that the differences between the mean scores are significant (*p*<0.05) in favour of the experimental group. These findings show that the teaching performed within the context of RSTP developed the students' level of using research skills. Moreover, Table 6 also shows that increasing mean scores were obtained from the monitoring tests in the advancing phases of the teaching process. This also shows that the students' learning of research skills improved.

The findings obtained from the experimental and control groups for the basic research skills aimed to be measured by each monitoring test are presented "between Table 7 to Table 11" In these tables, the "score" column shows the highest score to be taken from the relevant items in the monitoring test while the “number” column shows the number of students responding to the relevant item. In the “Total” column, the sum of the scores taken from all the items in the test is presented.

3.2.1. Findings obtained from the “Target (Task) Definition” sub-skill of the Monitoring Test 1

The distribution of the scores taken by the students in the experimental and control groups in the current study from the “*Target (Task) Definition*” sub-skill is given in Table 7. The "*Target (Task) Definition*" sub-skill included in the Monitoring Test 1 generally aims to measure the student's skills of *selecting a research topic* and *writing a research question for the selected topic*. To this end, two items to test the related skill are used. The characteristics measured by the items are as follows:

1st item: The student determines three research topics on the basis of the given information and selects one of them as the research topic.

2nd item: The student writes two research questions suitable for the research topic he/she has selected.

The distribution of the responses given to the two items in the “Target (Task) Definition” sub-skill is given in Table 7.

Table 7. *The distribution of the responses given by the experimental and control group students to the two items in the “Target (Task) Definition” sub-skill in the Monitoring Test 1.*

Monitoring Test 1*	Group	The Most Correct Answer		Distant Correct Answers		Wrong Answers		Received Minimum Score	Received Maximum Score
		Score	Frequency	Score Range	Frequency	Score	Frequency		
1a	Experimental	6	21	1-5	12	0	-	2	6
	Control	6	5	1-5	19	0	11	0	6
1b	Experimental	2	30	1	1	0	2	0	2
	Control	2	16	1	2	0	17	0	2
2a	Experimental	2	15	1	13	0	5	0	2
	Control	2	-	1	26	0	9	0	1
2b	Experimental	2	11	1	8	0	14	0	2
	Control	2	-	1	2	0	32	0	1
Total	Experimental	12	7	1-11	25	0	1	0	12
	Control	12	-	1-11	33	0	3	0	10

* *The monitoring test 1 was responded by 33 students in the experimental group and 35 students in the control group.*

As can be seen in Table 7, after the completion of the application given in relation to the “Target (Task) Definition” sub-skill in the RSTP, it is seen that the responses of the experimental and control group students are gathered in different response categories. The findings obtained from the analysis of the responses given by 33 students in the experimental group and 35 students in the control group to the items in the monitoring test 1 are as follows:

In the (1a) item, the students were asked to determine three research topics on the basis of the information given in the text and 6 points were defined for the most correct answer, 1-5 points were defined for the distant correct answers and 0 point was defined for the wrong answer. In the (1a) item, 21 students from the experimental group and 5 students from the control group were able to determine three research topics based on the information given in the text and reached the most correct answer; the remaining 12 students in the experimental group were grouped in the category of distant correct answers while the remaining students in the control group were grouped in the categories of distant correct answers and wrong answer. In the (1b) item, the students were asked to select one of the three research topics they had determined as the research topic and 2 points were defined for the most correct answer, 1 point for distant correct answers and 0 point for the wrong answer. In the (1b) item, 30 students in the experimental group and 16 students in the control group were able to select the research topic and to reach the most correct answer and the more students in the control group were grouped in the category of the wrong answer.

In the (2a) and (2b) items, the students were asked to write a research question for the research topic and 2 points were defined for the most correct answer, 1 for distant correct answers and 0 point for the wrong answer. In the (2a) item, 15 students in the experimental group were able to write the first research question suitable for the research topic and thus reached the most correct answer, 13 of them were gathered in the category of distant correct answers while 5 students in the category of the wrong answer, while none of the students in the control group reached the most correct answer and a high majority of them were gathered in the category of distant correct answers. In the (2b) item, while 11 of the students in the experimental group were able to write the second research question suitable for the selected research topic and reached the most correct answer, 8 of them were grouped in the category of distant correct answers and 14 were grouped in the category of the wrong answer, none of the students in the control group was able to reach the most correct answer and they were largely grouped in the category of the wrong answer.

When all the items in the monitoring test 1 are correctly answered, then the highest score to be taken is 12. When Table 7 is examined, it is seen that seven students in the experimental group gave the most correct answer and got this highest score yet none of the students in the control group reached the most correct answer and in both of the groups, the majority of the both experimental and control groups students were grouped in the category of distant correct answers.

3.2.2. Findings obtained from the "Information Seeking Strategies" sub-skill of the Monitoring Test 2

The distribution of the scores taken by the students in the experimental and control groups in the current study from the "*Information Seeking Strategies*" sub-skill is given in Table 8. The "Information *Seeking Strategies* " sub-skill included in the Monitoring Test 2 generally aims to measure the student's skills of *determining the types of resources suitable for having access to the needed information* and *determining the key words suitable for the research questions*. To this end, two open-ended items to test the related skill are used. The characteristics measured by the items are as follows:

1st item: The student selects three different resources suitable for having access to the information needed for the given three topics and writes their characteristics.

2nd item: The student determines three key words suitable for the research question.

The distribution of the responses given to the two items in the "Information Seeking Strategies" sub-skill is given in Table 8. As can be seen in Table 8, after the completion of the application given in relation to the "Information Seeking Strategies" sub-skill in the RSTP, it is seen that the responses of the experimental and control group students are gathered in different response categories. The findings obtained from the analysis of the responses given by 33 students in the experimental group and 34 students in the control group to the items in the monitoring test 2 are as follows.

Table 8. *The distribution of the responses given by the experimental and control group students to the two items in the "Information Seeking Strategies" sub-skill in the Monitoring Test 2.*

Monitoring Test 2*	Group	The Most Correct Answer		Distant Correct Answers		Wrong Answers		Received Minimum Score	Received Maximum Score
		Score	Frequency	Score Range	Frequency	Score	Frequency		
1a	Experimental	4	5	1-3	28	0	-	1	4
	Control	4	-	1-3	34	0	-	1	3
1b	Experimental	4	5	1-3	24	0	-	0	4
	Control	4	-	1-3	18	0	16	0	2
1c	Experimental	4	6	1-3	26	0	1	0	4
	Control	4	-	1-3	20	0	14	0	2
1	Experimental	12	2	1-11	31	0	-	1	12
	Control	12	-	1-11	34	0	-	2	9
2	Experimental	3	15	1-2	14	0	4	0	3
	Control	3	6	1-2	21	0	7	0	1
Total	Experimental	15	2	1-13	31	0	-	1	15
	Control	12	-	1-11	34	0	-	2	11

* *The monitoring test 2 was responded by 33 students in the experimental group and 34 students in the control group.*

In the (1a), (1b) and (1c) items, the students were given a research topic for each and they were asked to write which resources they would prefer while seeking information about these topics and the characteristics of these resources and 4 points were defined for the most correct answer, 1-3 points for distant correct answers and 0 point for the wrong answer. As can be seen in Table 8, only some students in the experimental group were able to determine three resources suitable for having access to the information needed for the research topic and wrote their characteristics correctly while all the control group students were grouped in the categories of distant correct answers and the wrong answer. In the first item, only two students in the experimental group were able to reach the most correct answer. In the second item, the students were asked to write three key words suitable for the given topics and 3 points were defined for the most correct answer, 1-2 points for distant correct answers and 0 point for the wrong answer. While 15 students in the experimental group and 6 students in the control group reached the most correct answer, the majority of the students in both the experimental and control groups were grouped in the category of distant correct answers.

When all the items in the monitoring test 2 are given the most correct answers, the highest score to be taken is 15. When Table 8 is examined, it is seen that two students in the experimental group gave the most correct answer and got this highest score yet none of the students in the control group reached the most correct answer and the remaining 65 students in the experimental and control groups were grouped in the category of distant correct answers; yet, the scores of the experimental group students are higher.

3.2.3. Findings obtained from the "Finding and Having Access" sub-skill of the Monitoring Test 3

The distribution of the scores taken by the students in the experimental and control groups in the current study from the "*Finding and Having Access*" sub-skill is given in Table 9. The "Finding and Having Access" sub-skill included in the Monitoring Test 3 generally aims to measure the student's skill of *reaching the resources needed by using printed resources and techniques to conduct an internet search in the library catalogue with key words*. To this end, four open-ended items to test the related skill are used. The characteristics measured by the items are as follows:

1st item: The student selects the suitable search index in the library catalogue to have access to the information needed for the topic given and explains its reason.

2nd item: The student prefers to search techniques in the search engine to have access to the information needed for the topic given and explains its reason.

3rd item: The student writes the search techniques he/she has preferred correctly in the search engine.

4th item: The student has access to the information needed by using the content and list of directory of the book.

The distribution of the responses given to the four items in the "Findings and Having Access" sub-skill is given in Table 9.

Table 9. *The distribution of the responses given by the experimental and control group students to the four items in the "Finding and Having Access" sub-skill in the Monitoring Test 3.*

Monitoring Test 3*	Group	The Most Correct Answer		Distant Correct Answers		Wrong Answers		Received Minimum Score	Received Maximum Score
		Score	Frequency	Score Range	Frequency	Score	Frequency		
1	Experimental	2	18	1	-	0	14	0	2
	Control	2	-	1	-	0	31	0	0
2a	Experimental	2	18	1	13	0	1	0	2
	Control	2	-	1	-	0	31	0	0
2b	Experimental	2	9	1	19	0	4	0	2
	Control	2	-	1	-	0	31	0	0
3a	Experimental	2	17	1	4	0	11	0	2
	Control	2	-	1	-	0	31	0	0
3b	Experimental	2	15	1	4	0	13	0	2
	Control	2	-	1	-	0	31	0	0
4	Experimental	3	26	1-2	6	0	-	2	3
	Control	3	13	1-2	15	0	3	0	3
Total	Experimental	13	2	1-12	30	0	-	5	13
	Control	13	-	1-12	28	0	3	0	3

* *The monitoring test 3 was responded by 32 students in the experimental group and 31 students in the control group.*

As can be seen in Table 9, after the completion of the application given in relation to the "Finding and Having Access" sub-skill in the RSTP, it is seen that while the responses of the experimental group are generally gathered in the categories of the most correct answer and distant correct answers, the responses of the control group students are generally gathered in the category of the wrong answer. The findings obtained from the analysis of the responses given by 32 students in the experimental group and 31 students in the control group to the items in the monitoring test 3 are as follows.

In the first item, the students were asked to determine which search index they need to use while conducting a resource search in the library catalogue and to explain its reason and 2 points were defined for the most correct answer, 1 point for distant correct answers and 0 point for the wrong answer. As can be seen in Table 9, in these items, only 18 students in the experimental group reached the most correct answer and the remaining 14 students were grouped in the category of the wrong answer while all of the control group students were gathered in the category of the wrong answer.

In the (2a) and (2b) items, the students were asked to write the search technique while conducting a search in the internet with its reason and 2 points were defined for the most correct answer, 1 for distant correct answers and 0 point for the wrong answer. When Table 9 is examined, it is seen that only some of the students in the experimental group were able to reach the most correct answer and almost all of the students in the experimental group were grouped under the categories of the most correct answer and the wrong answer while almost all of the students in the control group were gathered in the category of the wrong answer.

In the (3a) and (3b) items, the students were asked to write the search technique they preferred while conducting a search in the internet in the search engine and 2 points were defined for the most correct answer, 1 point for distant correct answers and 0 point for the wrong answer. When Table 9 is examined, it is seen that only some of the students in the experimental group were able to reach the most correct answer in these items and moat of the students in the experimental group were gathered in the categories of the most correct answer and the wrong answer while all of the students in the control group were gathered in the category of the wrong answer.

In the fourth item, the students were asked to determine three different page numbers from the content and list of directories of the book and 3 points were defined for the most correct answer, 1-2 points for distant correct answers and 0 point for the wrong answer. In this item, 26 students in the experimental group were able to correctly make use of the content and list of directories and reached the most correct answer and the remaining 6 students were gathered in the category of distant correct answers. In the control group, while 13 students were able to reach the most correct answer, 15 students were gathered in the category of distant correct answers and three were gathered in the category of the wrong answer.

When all the items in the monitoring test 3 are given the most correct answers, the highest score to be taken is 13. When Table 9 is examined, it is seen that two students in the experimental group gave the most correct answers to all the items and got this highest score yet none of the students in the control group reached the most correct answer and the majority of the students were gathered in the category of distant correct answers.

3.2.4. Findings obtained from the "Using Information" sub-skill of the Monitoring Test 4

The distribution of the scores taken by the students in the experimental and control groups in the current study from the "*Using Information*" sub-skill is given in Table 10. The "Using Information" sub-skill included in the Monitoring Test 4 generally aims to measure the student's skill of *preparing a note card to record the information reached, a bibliography card and referencing*. To this end, three open-ended items to test the related skill are used. The characteristics measured by the items are as follows:

1st item: The student prepares a note card using the information given.

2nd item: The student prepares a bibliography card using the information given.

3rd item: The student shows the information related to the reference in compliance with the citation rules.

The distribution of the responses given to the three items in the "Using Information" sub-skill is given in Table 10. As can be seen in Table 10, after the completion of the application given in relation to the "Using Information" sub-skill in the RSTP, while the responses of the experimental group are generally gathered in the categories of the most correct answer and distant correct answers, the responses of the control group students are generally gathered in the categories of distant correct answers and the wrong answer. The findings obtained from the analysis of the responses given by 32 students in the experimental group and 32 students in the control group to the items in the monitoring test 4 are as follows.

Table 10. *The distribution of the responses given by the experimental and control group students to the three items in the "Using Information" sub-skill in the Monitoring Test 4.*

Monitoring Test 4*	Group	The Most Correct Answer		Distant Correct Answers		Wrong Answers		Received Minimum Score	Received Maximum Score
		Score	Frequency	Score Range	Frequency	Score	Frequency		
1	Experimental	7	6	1-6	25	0	1	0	7
	Control	7	-	1-6	25	0	7	0	1
2	Experimental	6	8	1-5	23	0	1	0	6
	Control	6	-	1-5	25	0	7	0	1
3	Experimental	2	23	1	4	0	5	0	2
	Control	2	-	1	-	0	32	0	0
Total	Experimental	15	2	1-14	30	0	-	0	15
	Control	15	-	1-14	31	0	1	0	2

* *The monitoring test 4 was responded by 32 students in the experimental group and 32 students in the control group.*

In the first item, the students were asked to prepare a note card on the basis of the information given and 7 points were defined for the most correct answer, 1-6 points for distant correct answers and 0 point for the wrong answer. The rubric for this item is given in Appendix-1 as an example. When Table 10 is examined, it is seen that six students in the experimental group reached the most correct answer and 25 students were gathered in the category of distant correct answers while the students in the control group were gathered in the categories of distant correct answers and the wrong answer.

In the second item, the students were asked to prepare a bibliography card by using the information given and 6 points were defined for the most correct answer, 1-5 points for distant correct answers and 0 point for the wrong answer. While Table 10 is examined, it is seen that only eight students in the experimental group reached the most correct answer and 23 students were gathered in the category of distant correct answers, while the students in the control group were gathered in the categories of distant correct answers and the wrong answer.

In the third item, the students were asked to show the information related to the source in accordance with the citation rules and 2 points were defined as the most correct answer, 1 point for distant correct answers and 0 point for the wrong answer. When Table 10 is examined, it is seen that 10 students in the experimental group reached the most correct answer and the other students were gathered in the categories of distant correct answers and the wrong answer while 32 students in the control group were gathered in the category of the wrong answer.

When all the items in the monitoring test 4 are given the most correct answers, the highest score to be taken is 15. When Table 10 is examined, it is seen that two students in the experimental group gave the most correct answer to all the items and got this highest score while none of the students in the control group reached the most correct answer to the items and 61 students in both the experimental group and the control group were gathered in the category of distant correct answers yet the scores of the experimental group students are higher than those of the control group students.

3.2.5. Findings obtained from the "Organizing Information" sub-skill of the Monitoring Test 5

The distribution of the scores taken by the students in the experimental and control groups in the current study from the "*Organizing Information*" sub-skill is given in Table 11. The "Using Information" sub-skill included in the Monitoring Test 5 generally aims to measure the student's skill of *determining the outline of the research and organizing information in compliance with* *the outline and preparing a cover page*. To this end, two open-ended items to test the related skill are used. The characteristics measured by the items are as follows:

1st item: The student writes the research report by determining its main and sub headings.

2nd item: The student prepares the cover page of the research report.

The distribution of the responses given to the two items in the "Using Information" sub-skill is given in Table 11.

Table 11. *The distribution of the responses given by the experimental and control group students to the two items in the "Organizing Information" sub-skill in the Monitoring Test 5.*

Monitoring Test 5*	Group	The Most Correct Answer		Distant Correct Answers		Wrong Answers		Received Minimum Score	Received Maximum Score
		Score	Frequency	Score Range	Frequency	Score	Frequency		
1	Experimental	20	1	1-19	32	0	-	2	20
	Control	20	-	1-19	35	0	-	1	8
2	Experimental	9	22	1-8	11	0	-	6	8
	Control	9	-	1-8	22	0	13	0	6
Total	Experimental	29	1	1-28	32	0	-	8	29
	Control	29	-	1-28	35	0	-	1	13

* *The monitoring test 5 was responded by 33 students in the experimental group and 35 students in the control group.*

As can be seen in Table 11, after the completion of the application given in relation to the "Organizing Information" sub-skill in the RSTP, the students in the experimental and control groups were gathered in the category of distant correct answers yet the scores of the experimental group students are higher than those of the control group students. The findings obtained from the analysis of the responses given by 33 students in the experimental group and 35 students in the control group to the items in the monitoring test 5 are as follows.

In the first item, the students were asked to write the research report by determining the main and sub-headings and 20 points were defined for the most correct answer, 1-19 for distant correct answers and 0 point for the wrong answer. When Table 11 is examined, only one student in the experimental group reached the most correct answer in this item while the other students in both of the groups were gathered in the category of distant correct answers.

In the second item, the students were asked to prepare a cover page for the research report and 9 points were defined for the most correct answer, 1-8 points for distant correct answers and 0 point for the wrong answer. When Table 11 is examined, it is seen that 22 students in the experimental group reached the most correct answer in this item and the remaining 11 students were gathered in the category of distant correct responses. The students in the control group were gathered in the categories of distant correct answers and the wrong answer.

When all the items in the monitoring test 5 are given the most correct answers, the highest score to be taken is 29. When Table 11 is examined, it is seen that only one student in the experimental group gave the most correct answers to all the items and got this highest score while none of the students in the control group reached the most correct answer to all the items and that all the remaining students in both of the groups were gathered in the category of distant correct answers.

When the scores obtained by the students from both the posttest application of the RST and the monitoring tests are examined, it is seen that the activities prepared within the teaching program had comprehensive positive effects on the development of the skills directing the research process. When the number of the students having reached the most correct answer to the items in the monitoring tests is examined together with these scores, it is seen that there are students in the control group reaching the most correct answer in the sub-skills of "*selecting a research* *topic", "determining key words"* and *"having access to the information needed by students by making use of the content list and list of directories in printed resources and key words*". These findings also reveal the inadequacy of the existing programs in our education system in developing the skills that make up the research process. In the experimental group, although there were students who had the most correct answers in all of the skills, it was observed that all students' level of using these skills did not reach the desired level. These findings indicate that the prepared program had an important effect on developing students' research skills, but that a longer period is required to develop these skills.

4. DISCUSSION and CONCLUSION

When the findings obtained from the application of the pretest before the implementation of the Research Skills Teaching Program are examined, it is seen that the students' level of using skills related to research process is quite low. This finding is similar to the findings reported in the literature on research skills. In the existing research, it has also been reported that students' level of using research process skills is low (Alkan-Dilbaz, 2013; Chu et al., 2008; Polat-Demir & Kutlu, 2016; 2017). This shows that not enough importance is attached to these skills in Turkey.

At the end of the study however it was found that a significant improvement occurred in the experimental students' level of using the skills required to conduct the research process successfully when compared to the control group students. The findings of the current research show that the RSTP was effective in the inculcation of the skills constituting the research process in students. This finding concurs with the findings reported by the studies investigating the effect of special programs prepared to develop higher order thinking skills such as problem solving, critical thinking, conducting research, scientific process skills (Goudas & Giannoudis, 2008, Kurnaz & Kutlu, 2016; Ünal & Aral, 2014).

In their study, Goudas and Giannoudis (2008) examined the effectiveness of the Team Sports-Based Life Skills Program they developed for 6th and 8th grade students. The program was developed with a focus on goal setting, positive thinking and problem solving life skills. As a result of the study, it was concluded that this program was effective in developing students' life skills such as goal setting, positive thinking and problem solving. Kurnaz and Kutlu (2016) examined the effectiveness of the Scientific Process Skills Program they prepared for 4th grade elementary school students. As a result of the study, it was determined that this program was effective in developing students' scientific process skills. Ünal and Aral (2014), on the other hand, demonstrated that the Experiment Based Education Program was effective in developing the problem solving skills of 6-year-old children.

When the distribution of the scores taken from the monitoring tests used during the implementation of the program is examined, it is seen that the control group students couldn't reach the most correct answer in most of the items in these tests and that the number of control group students having reached the most correct answer is very few. On the other hand, in the experimental group, there are students having reached the most correct answer in all the items in the tests, yet some experimental group students' level of using these skills is quite low. In this respect, it is thought that a longer and circular teaching process is needed to impart the skills required for the successful conduct of the research process. Therefore, research skills should not only be included in the curriculum as a subject of a course, but activities that will enable students to gain these skills should be included in all courses from pre-school education onwards. In the literature, it is emphasized that more practice and time are required for students to acquire higher order mental skills (Beyer, 1991; Kurnaz & Kutlu, 2016; Kutlu et al., 2017). Beyer (1991) stated that it would take time to develop critical thinking skills and thinking activities should be done to impart these skills to students. In their study investigating the effectiveness of the Scientific Process Skills Program, Kurnaz and Kutlu (2016) emphasized that it should be taken into consideration that the development of scientific process skills could take a long time.

Based on the results of the current study, suggestions are made for both practitioners and researchers. A Research Skills Program prepared independently of the course content can be applied in schools from elementary school onwards. Thus, the development of students' skills that make up research processes can be positively affected. In schools, activities that can positively affect students' use of research skills can be included in all the subjects. Effective use of research skills by teachers and families can make it easier for students to develop these skills. For this reason, applications can be organized to increase the level of using research skills by teachers and parents. Researchers can develop a Research Skills Teaching Program and examine its effectiveness for different grade levels, regardless of the course objectives and content.

Declaration of Conflicting Interests and Ethics

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

Authorship Contribution Statement

Betul Polat: Investigation, Resources, Methodology, Visualization, Software, Formal Analysis, and Writing -original draft. **Omer Kutlu**: Investigation, Resources, Methodology, Visualization, Software, Formal Analysis, and Writing -original draft.

ORCID

Betul Polat https://orcid.org/0000-0002-1618-3118

Omer Kutlu https://orcid.org/0000-0003-4364-5629

REFERENCES

Abston, K., Stout, V.J., & Crowder, C. (2004). *Lessons learned in a virtual team: An integrative model for graduate student research skill development.* The Academy of Human Resource Development International Conference (AHRD). Austin.

Aksu, M. (2005,). *Eğitim fakültelerinin değişen rolleri ve Avrupa boyutu [Changing roles of education faculties and the European dimension]*. Eğitim Fakültelerinde Yeniden Yapılandırmanın Sonuçları ve Öğretmen Yetiştirme Sempozyumu, Eylül 22-24, Gazi Üniversitesi, Ankara.

Alkan, C. (1989). Eğitim Bilimlerinde araştırma [Research in educational sciences]. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, 22*(1), 23-27. https://doi.org/10.1501/Egifak_0000000860

Alkan-Dilbaz, G. (2013). *Araştırma temelli öğrenmenin tutum, akademik başarı, problem çözme ve araştırma becerilerine etkisi [The effects of inquiry-based learning on attitude, academic success, problem solving and inquiry skills]* [Unpublished master's thesis]. Mersin University.

American Association of School Librarians. (2007). *Standarts for the 21st century learner.* Retrieved October 9, 2014, from http://www.ala.org/aasl/sites/ala.org.aasl/files/content/guidelinesandstandards/learningstandards/AASL_LearningStandards.pdf

Anderson, L.W., Krathwohl, D.R., Airasian, P.W., Cruikshank, K.A., Mayer, R.E., Pintrich, P.R., Raths, J., & Wittrock, M.C. (2001). *Öğrenme öğretim ve değerlendirme ile ilgili bir sınıflama: Bloom'un eğitimin hedefleri ile ilgili sınıflamasının güncelleştirilmiş biçimi* (D. A. Özçelik, Çev.). Pegem Akademi.

Başdoğan, B. (2013). *Sosyal bilgiler 4. sınıf ders ve öğrenci çalışma kitabı [Social studies 4th grade course and student workbook]*. Evrensel İletişim Yayınları.

Berber, Ş. (2003). Bilgi çağında eğitim [Education in the information age]. *TSA, 7*(2), 39-50.

Berberoğlu, G. (2006). *Sınıf içi ölçme ve değerlendirme teknikleri [In-class measurement and evaluation techniques]*. Morpa Yayınları.

Beyer, B.K. (1987). *Practical strategies for the teaching of thinking*. Allyn and Bacon, Inc.

Beyer, B. (1991). *Teaching thinking skills: A handbook for elementary school teachers.* Allyn and Bacon.

Bird, P. (2000). *Help your child to learn at primary school: How to support your child and improve their learning potential*. How to Books Ltd.

Büyüköztürk, Ş. (2009). *Sosyal Bilimler için veri analizi el kitabı [Manual of data analysis for social sciences]*. Pegem Akademi.

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2009). *Bilimsel araştırma yöntemleri [Scientific research methods]*. Pegem Akademi.

Cambridge, D. (2010). *Eportfolios for lifelong learning and assessment.* Jossey-Bass.

Candy, P.C. (2003). *Lifelong learning and information literacy*. Report for U.S. National Commission on Libraries and Information Science and National Forum on Information Literacy. http://www.nclis.gov/libinter/infolitconf&meet/papers/candy-fullpaper.pdf

Chu, S., Chow, K., Tse, S., & Kuhlthau, C.C. (2008). Grade 4 students' development of research skills through inquiry-based learning projects. *Library, Information Science & Technology Abstracts (LISTA), 14*(1), 10-37.

Daştan, İ., & Gürler, C. (2016). Yeşil bilgi teknolojileri ürün tercihinde tüketici satın alma niyetlerini etkileyen faktörlerin tespiti [Factors affecting consumers' intentıon to buy green it goods]. *Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi, 30*(1), 173-186.

Demirel, Ö. (2007). *Eğitimde yeni yönelimler [New directions in education]*. Pegem Akademi.

Doğanay, A. (2008). *Öğretim ilke ve yöntemleri [Teaching principles and methods]*. Pegem Akademi.

Eisenberg, M.B., & Berkowitz, R.E. (1990). *Information problem solving: The Big Six Skills approach to library information skills instruction.* Ablex Publishing Corporation.

Goudas, M., & Giannoudis, G. (2008). A team-sports-based life skills programme in an physical education context. *Learning and Instruction, 18*(6), 538-546. https://doi.org/10.1016/j.learninstruc.2007.11.002

Güneş, P. (2011). *Dereceli puanlama anahtarının ilköğretim öğrencilerinin araştırma becerisi ve bilişsel alan düzeyine etkisi [The effect of rubric on primary students? research skills and cognitive level]* [Unpublished doctoral dissertation]. Hacettepe University,

İlter, İ. (2013). *Sosyal Bilgiler öğretiminde 5E öğrenme döngüsü modelinin öğrenci başarısına, bilimsel sorgulayıcı-araştırma becerilerine, akademik motivasyona ve öğrenme sürecine etkileri [The effects of 5E learning cycle model to students' achievement, scientific inquiry skills, academic motivation and learning process in the social studies teaching]*. [Unpublished doctoral dissertation]. Atatürk University.

King, F.J., Goodson, L., & Rohani, F. (1998). *Higher order thinking skills: Definitions, strategies, assessment*. Retrieved October 11, 2015 from: http://www.cala.fsu.edu/files/higher_order_thinking_skills.pdf

Knutson, D.S., Dozier, K.S., & Migotsky, S.C. (1995, July 12-15). *Meta Research: Researching student researchers' methods* [Conference presentation]. 14th, University Park, PA (ERIC ED393100)

Kumandaş, H., & Kutlu, Ö. (2013). Okulöncesi öğretmen adaylarının kendi sunum becerilerine ilişkin öz değerlendirmeleri ile eğitici değerlendirmesinin karşılaştırılması [The comparison of trainers and self assessments regarding presentation skills of preschool teachers candidate]. *Educational Science and Practice*, *12*(23), 43-55.

Kurnaz, B., & Kutlu Ö. (2016). İlkokul 4. sınıf için hazırlanan bilimsel süreç becerileri programının etkililiğinin belirlenmesi [Determining the effectiveness of science process skills program prepared for elementary school grade 4]. *İlköğretim Online, 15*(2), 529-547. http://dx.doi.org/10.17051/io.2016.36891

Kutlu, Ö., Doğan, D., & Karakaya, İ. (2017). *Ölçme ve değerlendirme: Performansa ve portfolyoya dayalı durum belirleme [Measurement and evaluation: Assessment based on performance and portfolio]*. Pegem Akademi.

Kutlu, Ö., Yalçın, S., & Pehlivan, E.B. (2010). İlköğretim programında yer alan kazanımlara dayalı soru yazma ve puanlama çalışması. [A Study on writing and scoring open-ended questions based on the primary school curriculum objectives]. *İlköğretim Online, 9*(3), 1201-1215.

MEB. (2006). *İlköğretim Sosyal Bilgiler Dersi 4. sınıflar öğretim programı ve kılavuzu [Primary Education Social Studies Course 4th grade curriculum and guide]*. MEB Talim Terbiye Kurulu Başkanlığı Devlet Kitapları Müdürlüğü.

Miles, M.B., & Huberman, A.M. (1994). *Qualitative data analysis*: *An expanded sourcebook* (2nd ed). Sage Publications, Inc.

Mısırlı, İ. (2007). *Genel ve teknik iletişim [General and technical communication]*. Detay Yayıncılık.

Numanoğlu, G. (1999). Bilgi toplumu - Eğitim - Yeni kimlikler- Bilgi toplumu ve eğitimde yeni kimlikler [Information society - Education - New identities - Information society and new identities in education]**.** *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, *32*(1-2), 341-350. https://doi.org/10.1501/Egifak_0000001170

Polat-Demir, B., & Kutlu, Ö. (2016). The effect of electronic portfolio applications on the 6th graders' research skills. [Elektronik portfolyo uygulamalarının ortaokul 6. sınıf öğrencilerinin araştırma becerilerine etkisi]. *Education and Science*, *41*(188), 227-253. http://dx.doi.org/10.15390/EB.2016.6724

Polat-Demir, B., & Kutlu, Ö. (2017, Nisan 6-8). *Ortaokul öğrencilerinin araştırma beceri düzeylerinin farklı değişkenlere göre incelenmesi [Examining the research skill levels of secondary school students according to different variables]* [Konferans sunumu]. Uluslararası Avrasya Sosyal Bilimler Kongresi (International Congress of Eurasian Social Sciences) Alanya, Antalya.

Shuman, L.J., Besterfield-Scare, M., & McGourty, J. (2005). The ABET "Professional skills" - can they be taught? Can they be assessed? *Journal of Engineering Education, 94*(1), 41-55. https://doi.org/10.1002/j.2168-9830.2005.tb00828.x

Turam, E. (1999). *Ekranaltı çocukları [Underscreen kids]*. İrfan Yayıncılık.

Ünal, M., & Aral, N. (2014). Deney yöntemine dayalı eğitim programının 6 yaş çocuklarının problem çözme becerilerine etkisinin incelenmesi [An investigation on the effects of experiment based education program on six years olds' problem solving skills]. *Education and Science, 39*(176), 279-291. http://dx.doi.org/10.15390/EB.2014.3592

Üstünoğlu, E. (2006). Üst düzey düşünme becerilerini geliştirmede bilişsel soruların rolü [The role of cognitive questions in developing higher-order thinking skills]. *Çağdaş Eğitim Dergisi*, *331*, 17-24.

Wu, H.K., & Hsieh, C.E. (2006). Developing sixth greders' inquiry skills to construct explanations in inquiry based learning environments. *International Journal of Science Education, 28*(11), 1289-1313. https://doi.org/10.1080/09500690600621035

Yelland, N. (2007). *Shift to the future: Rethinking learning with new technologies in education.* Routledge Taylor & Francis Group.

Yıldırım, S. (2007). *İlköğretim 4. Sınıf sosyal bilgiler dersinde proje tabanlı öğrenme modelinin araştırma becerilerinin gerçekleşme düzeyine etkisi [The influence of project-based learning model on the realization level of research skills in the social science lesson of the 4th grade primary school students]* [Unpublished master's thesis]. Marmara University.

Zoller, U. (2000). Interdisiplinary systematic HOCS development -the key for meaningful STES- oriented chemical education. *Chemistry Education: Research and Practice in Europe (CERAPIE), 1*, 189-200.

APPENDIX

Example Rubric
Monitoring Test 4 - Item 1

ANSWERS	Achievement Score
Most Correct Answer	
The student fills in the scorecard using all four pieces of information in the piece "Historical development". The student writes the information about "the change of village settlement centers" on the note card. The student writes down the note card and bibliography card number and the page numbers containing the information.	**7**
Distant Correct Answers	
The student fills in the note card by missing one of the four information in the "Historical development" text. The student writes the information on the "change of village settlement centers" on the note card, the note card and bibliography card number, but writes one of the page numbers where the information is found.	**6**
The student fills in the scorecard using the three pieces of information from the text "Historical development". The student writes the information on the "change of village settlement centers", the note card and bibliography card number on the note card, but does not write the page number where the information is found. ***Or*** The student writes the information on the "change of village settlement centers" on the note card, writes the note card number and the page numbers where the information is found.	**5**
The student fills in the note card by missing one of the three pieces of information in the "Historical development" text. The student writes the information on the "change of village settlement centers" on the note card, writes the note card number and one of the page numbers where the information is found.	**4**
The student fills in the scorecard using the two pieces of information from "Historical development" text. The student writes the information on the "change of village settlement centers" and the page number of the information on the note card. ***Or*** The student writes the information on the "change of village settlement centers" and the bibliography card number on the note card.	**3**
The student fills in the note card by missing one of the two pieces of information in the "Historical development" text. The student writes the information on the "change of village settlement centers" and one of the page number containing the information on the note card.	**2**
The student fills in the note card by writing only one piece of information in the "Historical development" text. The student writes briefly the information on the "change of village settlement centers" on the note card. **Or** The student either writes down the bibliography card number or only one of the page numbers containing the information.	**1**
Empty	**0**
Wrong Answer	
The student writes down information that is completely different from what needs to be written.	**0**

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 1, 61–79

https://doi.org/10.21449/ijate.946517

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

Enhancing teaching and learning in higher education through formative assessment: Teachers' Perceptions

Shamsiah Banu Mohamad Hanefar [1,*], Nusrat Zerin Anny [2], Md. Sajedur Rahman [3]

[1]Centre for Academic Partnerships & Engagement (CAPE), University of Nottingham Malaysia, Malaysia
[2]Department of Sociology, Rajshahi College, Rajshahi, Bangladesh
[3]Department of Economics, Rajshahi College, Rajshahi, Bangladesh

ARTICLE HISTORY

Received: June 01, 2021
Revised: Oct. 12, 2021
Accepted: Jan. 13, 2022

Keywords:
Formative assessment,
Teachers' perceptions,
Higher education.

Abstract: A good assessment system is one of the preconditions for quality education. Formative assessment is comparatively an emerging idea to assess the students throughout the academic year with the intention to identify and overcome the weaknesses of the students and enhance their learning outcome. Taking these into account, this study attempted to explore the teachers' perceptions of the use of formative assessment in enhancing teaching and learning in Bangladesh higher education. A mixed-method study was employed with survey and semi-structured interview as the data collection methods. 100 participants were randomly (simple random) selected for the survey, and 6 participants were purposively selected for the interviews. For analysing the data, descriptive analysis and content analysis were used. The findings of the study revealed majority of the participants agreed that formative assessment is crucial to enhance teaching and learning in Bangladeshi colleges. Nonetheless, there are some challenges like - teachers' biasness, shortage of teachers, large class, poor infrastructure, insufficient power supply, and heavy workload of the teachers. As a whole, this study will provide a fundamental ground for future research in formative assessment in Bangladeshi colleges specifically and for comparative study with other higher education institutions globally.

1. INTRODUCTION

Assessment has always been an important part of education that has a great impact on shaping the students' learning (Azim, 2014). One of the prerequisites for a good educational system is a good assessment system. (Bjornsrud & Engh, 2012). Assessment is a term used to describe a method in which teachers collect information about the teaching-learning environment, such as the students' level of knowledge and comprehension (Ngendahayo, 2014), is a key component of teaching and learning in higher education (Gikandi, Morrow, & Davis, 2011). According to William (2014), in many educational institutions, students are assessed at the completion of an academic year (summative assessment) even though evolving idea of a new assessment system (formative assessment) has been advocated by many scholars (Black & William, 2009; Sadler, 1998; Yorke, 2003). Formative assessment ensures information and feedback during the

*CONTACT: Shamsiah Banu Mohamad Hanefar ✉ Shamsiah.Banu@nottingham.edu.my, shbanu21@gmail.com 🖃 Centre for Academic Partnerships & Engagement (CAPE), University of Nottingham Malaysia, Malaysia

e-ISSN: 2148-7456 **/© IJATE** 2022

learning process, while summative assessment occurs after the learning process has been completed and provides feedback and information regarding the process (Figa, Tarekegne, & Kebede, 2020; Paul et al., 2016). The primary focus of this research is on formative assessment and its effect on teaching-learning practises.

In recent years, all over the world, formative assessment has become the preferred form of assessment compared to summative assessment (Ozan & Kıncal, 2018). Likewise, The World Bank directly links high-quality, formative assessment to better outcomes on standardised tests, and links better learning outcomes to increased national prosperity (Clarke, 2012). In-line with this need, Bangladesh is also experiencing a shift from summative to formative assessment for more than last one decade (Rahman et al., 2021), and formative assessment is gradually becoming more important in this context. The country has already started formative assessment in the junior secondary level named School Based Assessment in order to assess learners' holistic development since 2007 (Begum & Farooqui, 2008; Rahman et al., 2021). Having said that, the assessment system in higher education is still poor in Bangladesh (Mamun-ur-Rashid & Rhman, 2017; Rahman et al., 2019; Ahmed et al., 2021), where there is little or no place for effective on-going assessment and appropriate feedback system.

At present, Bangladesh is struggling to enhance the quality of education at every level (Mamun-ur-Rashid & Rhman, 2017). Its higher educational institutions are passing through a very hard time as their rankings are poor even in the context of South Asian Countries (Mahmud, 2019). In terms of assessment, higher educational institutions in the country still depend mostly on the traditional summative assessment as the main way of measuring the students' learning outcomes (Rahman et al., 2019). What is more, Muhammad et al.'s (2019) study has revealed that a large number of Bangladeshi colleges do not have a clear idea about what formative assessment is. As a result, students just memorise their course materials for the year-end examination and get their certificates (Muhammad et al., 2019). Here, it should also be noted that research on the use of formative assessment in Bangladeshi higher education is also scarce in number. This means that many issues related to formative assessment in higher education of Bangladesh are still unexplored. This study, therefore, has attempted to fill the knowledge gap regarding formative assessment in Bangladeshi higher education sector by exploring teachers' perception on the use of formative assessment to enhance teaching and learning practices.

The concept of formative assessment arose in the 1930s and 1940s as a result of cognitive and constructivist learning theories in line with the ideas of feedback and development (Roos & Hamilton, 2005). It is a pre-determined method for assessing students' learning status that is used by teachers to tailor their instructional strategies or by students to adapt their learning strategies (Barney & McCowans, 2009). Furthermore, it denotes a form of assessment that judges the students on the basis of their performance with the intention to give them constructive feedback aiming at the enhancement of learning outcomes (Sadler, 1998). This type of assessment is designed with the intention to explore students' intelligence or knowledge, permitting teachers to understand students' prior knowledge, and select the more appropriate teaching-learning strategy for them (Bransford et al., 1999). Clarck (2012) and Ngendahayo (2014) argued that formative assessment is considered as both 'learning for assessment' and 'learning as assessment' with the intention to promote students' learning and enhance their critical thinking respectively. However, in Bangladesh most teachers do not differentiate between formative and summative assessments for grading purpose, they occasionally use summative assessments for formative purposes. This supports William's (2000) findings, which claim that in most countries, few teachers are able or willing to use parallel assessment systems - one is for summative purposes, and the other is for formative purposes. As a result, teachers often replicate and duplicate the assessment process.

Many scholars claim that the effect of formative assessment on teaching and learning is positive (Winstone & Millward, 2012; Yorke, 2003; Young & Jackman, 2014). This type of assessment can help both teachers and students in a variety of ways. (López-Pastor & Sicilia-Camacho, 2017; Wang et al., 2006; Winstone & Millward, 2012). Rather than producing just a competent individual in a specific field, one of the current priorities of higher education is to develop highly skilled students with a variety of attributes such as problem-solving capacity, analytical thinking, stress management, and so on. (Dochy et al., 1999). In line with this, the Organization for Economic Cooperation and Development (OECD) recommended a shift from summative to formative evaluation in 2009, citing successful assessment as one of the most significant factors in improving learning outcomes (Bjornsrud & Engh, 2012). In their study, Harlen & James (1997) argue that the real learning is something that comes through human interaction or/and exchange of ideas and actions in the real-life situation. In this regard, a study by Umer and Omer (2015) found that formative assessment helps students to grow as an 'independent learner'. Dochy et al. (1999) also argue that formative assessment is a more effective way of assessing students in order to prepare them for the current competitive world.

In terms of its effectiveness, Black and William (1998) found that formative assessment is effective in almost all kinds of educational settings (Sadler, 1998; Umer & Omer, 2015), but some scholars argue that formative assessment is more effective for higher education (Bennett, 2011; Dochy et al., 1999). Ruiz-Primo (2011) found that informal formative assessment is more effective in terms of constructive learning. In fact, this type of assessment can be very helpful for both teachers and students (Bransford et al., 1999). In addition, several prior studies have also proven the effectiveness of different innovative designs or strategies of formative assessment in enhancing the students' achievement in terms of learning outcome (Black & Wiliam, 1998; Winstone & Millward, 2012; Yorke, 2003; Young & Jackman, 2014). In terms of teachers' perceptions on the effectiveness of formative assessment, several studies have found that most of the teachers perceive formative assessment positively (Dijksterhuis et al., 2013; López-Pastor & Sicilia-Camacho, 2017; Sach, 2012; Winstone & Millward, 2012; Young & Jackman, 2014). Based on the existing literature, it is clear that formative assessment has the potential to enhance the teaching-learning process.

In spite of a huge number of positive writings on the effectiveness of formative assessment, there are a considerable number of oppositions also (Antoniou & James, 2013). Problem of applying formative assessment starts from the issue of its definition as it is not a well-defined concept (Bennett, 2011). In this regard, the concept of formative assessment and one of its central elements, which is feedback, are used for different purposes in different papers (Antoniou & James, 2013; Bennett, 2011). This ambiguity affects the practical assessment scenario by creating misunderstanding as different people have different understandings about the same concept (Klenowski, 2009). Apart from that, the quality of feedback can also affect the whole process of assessment and learning (Dijksterhuis et al., 2013; Umer & Omer, 2015). Furthermore, the efficiency of this form of assessment in higher education is not clear (Yorke, 2003). In order to get the best result from formative assessment, a precise definition of the concept should be developed to reduce the ambiguities related to what formative assessment refers to (Bennett, 2011).

In addition, there are also some preconditions of implementing formative assessment like-motivated teacher; effective classroom practice; favourable classroom environment; long-term relation based on commitment between the learner and the educator, and the need for necessary tools (Antoniou & James, 2013; Dijksterhuis et al., 2013). As claimed by Black & Wiliam (2003), the lack of these necessary preconditions could become the major challenge in implementing formative assessment.

The existing tension between summative and formative assessment also affects the frequency of implementing formative assessment (Sach, 2012). Recent study by Figa et al. (2020) has listed several factors that may hinder the use of formative assessment. According to them, the lack of learning aids, teachers' competence and large class size are categorised as the most challenging factor while time constraint, impact of summative assessment and shortage of classroom facilities are considered as medium challenging factors. In addition, factors such as teachers' negative attitude, enhanced workload, absence of structured assessment guideline, and students' negative attitude have been found to be the moderate challenging factors in the implementation of formative assessment in higher education. Similar challenges have also been found in the papers of Awasthi, and Chaudhary (2015), Sach (2012), Sharma et al. (2015), and López-Pastor and Sicilia-Camacho (2017). Furthermore, some argue that formative assessment can create biasness (Antoniou & James, 2013).

In addition, Young and Jackman (2014) also found negative attitudes of the teachers due to the lack of confidence, incompetency, fear of workload, pressure from guardians of the students or administration, time, and resource constraints. The same study also suggests that in some cases, teachers consider formative assessment as a positive measure but do not apply this in their class because of the aforesaid reasons. In fact, some teachers refuse to apply formative assessment as they believe that their students should be assessed just as they were assessed by their teachers in their student life (López-Pastor & Sicilia-Camacho, 2017). In a nutshell, the above review suggests that there is a continuous debate on the effectiveness of formative assessment, which, in fact, is essential for the improvement of formative assessment practices. Overall, while some negative attitudes towards formative assessment have been documented in numerous past studies, it is also worth reiterating that the majority of teachers have demonstrated positive attitudes towards the approach.

The higher education of Bangladesh is passing through a lot of challenges like scarcity of efficient teachers, shortage of classroom, absenteeism, problems between the teachers and the students, and political pressure (Monem & Baniamin, 2010; Muhammad et al., 2019). Due to these issues, the country is struggling to ensure the quality of education in tertiary level (Rabiul, 2014). In most higher education institutions of Bangladesh, academic performance of a student is judged predominantly on the basis of a final examination (Shahidul, 2016). In their study, Bhuiyan & Hossain, (2017) recommended that the traditional classroom practice should be changed towards interactive learning in order to engage the students in different activities throughout the academic year. The teacher-student relations should also be improved through frequent interaction (Bhuiyan & Hossain, 2017). Within this kind of situation, it is not easy to implement effective formative assessment in the higher education of Bangladesh. Nonetheless, the government is implementing formative assessment in the secondary level institutions, where many teachers appreciate this as a way of getting better learning outcome (Begum & Farooqui, 2008; Harlen & James, 1997). Formative assessment has also been implemented successfully in the medical education of Bangladesh (Riaz et al., 2015).

As a whole, the above discussion shows that formative assessment is not totally absent in the Bangladeshi education system, but more issues need to be explored so that formative assessment can be implemented more effectively in this context. Drawing on this need, we have taken the initiative to explore teachers' perceptions toward formative assessment and its role in enhancing the teaching and learning in Bangladeshi higher education.

2. METHOD

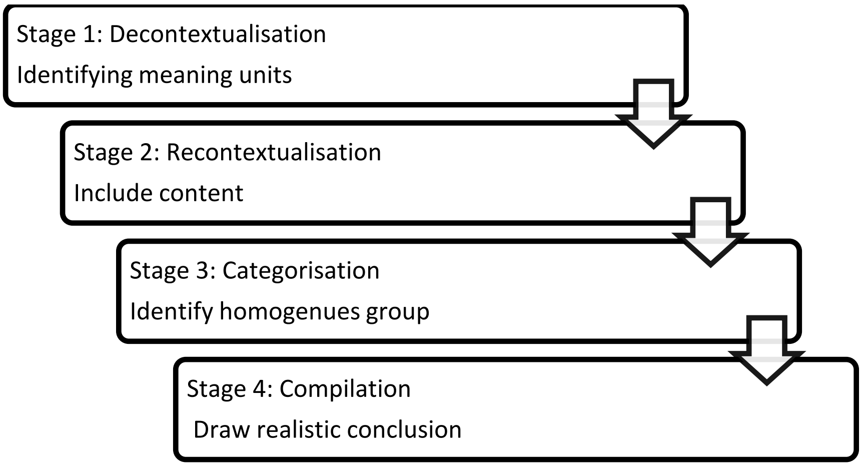
A mixed-method approach was adopted by combining qualitative and quantitative approaches with the intention to get a complete picture of the use of formative assessment to enhance teaching and learning based on a group of Bangladeshi college teachers' perspectives. This mixed - method has allowed the authors to triangulate the data and validate the data collected. For the data collection process, the study began with its quantitative strand, where a survey was administered. Next, the study proceeded with its qualitative strand, where a series of semi-structured interview sessions aiming to elicit in-depth information about teachers' perceptions on the use of formative assessment, were conducted. Prior to data collection, an information sheet describing the background and purpose of the study was presented and the consent was obtained from all the participants and the principal of a selected government college (RJ College).

This study has been carried out within the context of RJ College (one of the researchers' workplace) which is a well-known government college in Bangladesh. There are 250 teachers teaching in this college. Majority of the teachers are still using summative assessment to asses their students. Nonetheless, many of them have gradually started to use formative assessment tools such as class test, questioning, quiz, in-class competition, presentation, and workshop.

A total of 106 teachers from RJ College, Bangladesh have participated in this study. For the selection process, 100 teachers (among the 250 teachers teaching at the respective college) were first selected randomly using simple random sampling technique to participate in the survey. Then, six teachers from the same college were purposively selected for the subsequent semi-structured interviews. Teachers who participated in the survey were excluded from the interviews to avoid biasness in their answers. In terms of working experience, 78% of the survey respondents had worked in at least one other college before their service at this participating college. Among the interview participants, five respondents had worked in at least one other college prior to joining this participating college.

For the survey, a structured questionnaire containing 12 items were developed. To measure the teachers' perceptions and current practices of formative assessment, a 5-point Likert scale (from strongly disagree to strongly agree) was used. A pilot study was conducted for this survey with 40 participants and the Cronbach's Alpha test was performed through SPSS v 25, where the alpha value is 0.746. All quantitative data were analysed with SPSS v 25 (descriptive analysis was used). For the semi-structured interviews, an interview schedule containing 10 questions was used to guide the data collection process. To ensure its validity and reliability, two interview sessions were conducted to pilot the instrument (interview schedule) that had been developed for this study. Apart from that, the interview schedule was checked by two experts. In analysing the qualitative data, a content analysis had been conducted. Four main stages of qualitative content analysis have been used in this study (Figure 1): the decontextualisation, the recontextualisation, the categorisation, and the compilation (Bengtsson, 2016).

Figure 1. *A qualitative content analysis adapted from Bengtsson (2016).*


Stage 1: Decontextualisation
Identifying meaning units
Stage 2: Recontextualisation
Include content
Stage 3: Categorisation
Identify homogenues group
Stage 4: Compilation
Draw realistic conclusion

3. RESULTS / FINDINGS

3.1. Quantitative results and analysis

The results of the quantitative study depicted a mixed outcome. Majority of the respondents strongly agreed and agreed that formative assessment is important but currently it is not being emphasised as a main form of assessment throughout the academic year. Besides, there are some who still believe that formative assessment would not be able to provide a good quality feedback. Nonetheless, majority of them perceived formative assessment as a form of assessment that enhance students' engagement, attentiveness and interest. Furthermore, it allows students to inquire in-depth knowledge and decrease their absenteeism, and provides teachers to assess the gradual development of the students' learning outcome. Details of the results are provided in Table 1 below:

Table 1. *Quantitative results on teachers' perception on the role of formative assessment to enhance teaching and learning in higher education.*

No	Statement	1	2	3	4	5	Mean	Standard Deviation
1	Class performance based formative assessment makes the students more attentive in their class.	0	3	14	38	45	4.25	0.81
2	Formative assessment helps to engage the students with their study.	0	4	6	52	38	4.24	0.74
3	Formative assessment allows the teachers to assess gradual development status of the learning outcome.	0	0	10	56	34	4.24	0.62
4	Formative assessment allows the students to acquire in-depth knowledge.	3	3	11	42	41	4.15	0.95
5	Percentage of marks in final grading through formative assessment contributes to decrease absenteeism.	1	5	15	37	42	4.14	0.92
6	Formative assessment strategies make their study interesting to the students.	0	9	10	51	30	4.02	0.88
7	Formative assessment is the most important type of assessment.	9	12	13	54	12	3.48	1.13
8	Formative assessment happens all through the academic year.	4	26	13	35	22	3.45	1.21
9	Political/other pressure affects the system of grading through formative assessment.	7	22	18	27	26	3.43	1.29
10	Formative assessment provides quality feedback.	9	21	9	49	12	3.34	1.20
11	The current practice of formative assessment is satisfactory.	2	25	29	32	12	3.27	1.03
12	Teacher's biasness is a problem in case of grading through formative assessment.	15	29	24	16	16	2.89	1.30
Average mean score: 3.74								

(The items were re-arranged in a descending order to ease the discussion.)

The average mean score for all the 12 items is 3.74. Among all the 12 items, for 6 items, the mean value is less than the average mean score (Item 7 to 12). These results indicate that the teachers put less emphasis on the importance of the formative assessment and the practice is limited in the context of higher education in Bangladesh as they maybe feel more comfortable in the traditional method of assessment which is the summative assessment, i.e. final examination. This statement is supported by findings of other researchers (Muhammad et al., 2019; Rahman et al., 2019; Rabiul, 2014; Shahidul, 2016).

For the other remaining 6 items (Item 1 to 6), the mean value is higher than the average mean score. This clearly shows that teachers were aware about the benefits of formative assessment and they demonstrated positive perceptions on this type of assessment. Studies done by Dijksterhuis et al (2013), López-Pastor and Sicilia-Camacho (2017), Sach (2012), Winstone and Millward (2012), and Young and Jackman (2014) also indicate that teachers perceived formative assessment positively as it leads to many positive outcomes such as enhancing students' engagement and interest in learning, allows students to acquire in-depth knowledge, and increase class attendance.

3.2. Qualitative Findings and Analysis

Findings of qualitative analysis have revealed different themes related to the teachers' perceptions on formative assessment. The themes are: 1) Teacher's prior understanding of formative assessment in higher education, 2) current practice of formative assessment, 3) teachers' attitudes towards formative assessment, 4) importance of implementing formative assessment, and 5) challenges in implementing formative assessment. Table 2 outlined the details of these findings.

Table 2. *Qualitative finding on teachers' perception on the role of formative assessment (FA) to enhance teaching and learning in higher education.*

Meaning Unit	Condensed Meaning Unit	Code	Category	Theme
"Students can easily get prepared for their final examination through this type of assessment prior to the examination. As a result, they don't feel very much pressured before their final examination" (Teacher 1)	Students able to prepare for final examination with less pressure	Less pressured exam preparation	Benefit of FA	Teacher's prior understanding
"Students get the opportunity to enhance their problem-solving skills through effective application of formative assessment" (Teacher 4)	Student able to enhance their problem solving skills	Enhancement of problem solving skill	Benefit of FA	
"It enhances the learning of the students, as it makes them more confident and finally make them able prepare them to cope with the competitive real world" (Teacher 6)	Enhance student learning, build student's confident; coping with competitive real world	Enhancement of student learning To build student confident level To cope with real life scenario	Benefit of FA Benefit of FA	
"I use different types of FA such as class test, questioning, quiz, competition, presentation, workshop, and symposium" (Teacher 5)	Different types of FA are being used; class test, questioning, quiz, competition, presentation, workshop, and symposium.	Different types of FA in-use	Types of FA in practice	Current practice
"I believe many of us use formative assessment as a mean for grading and adding the marks rather than providing feedback to the students" (teacher 2)	FA is used as a mean for grading and adding the marks rather than providing feedback to the students	The use of FA to grade rather than providing feedback	Assessment of learning rather than assessment for learning	
"The main form of assessment is still summative assessment (SA), which is that is, a final examination of 80% of total marks for each course after ending an academic year. The remaining 20% of the marks comes from attendance (5%) and in-course assessment (15%) such as class test, written assignment, oral presentation, poster presentation, and etcetera" (Teacher 3)	SA contributes 80% of the grading meanwhile 20% is allocated for FA.	The purpose of SA and FA are for grading	Assessment of learning rather than assessment for learning	

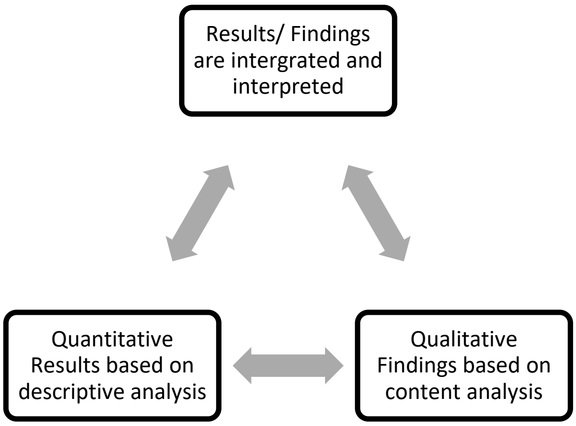
"On the spot class test is arranged whenever necessary to keep the learning momentum and attention of the students in class" (Teacher 1)	On the spot class test to keep the learning momentum and attention of the students	Formative assessment as a tool for student engagement	Assessment as learning	
"Actually, we work in a rigid system structure where there is a little chance to change the system. At present, we are adding the marks from in-course examination to the final marks of our students" (Teacher 6).	The work place system is rigid, little room for improvement. The marks allocated for in class activities are added to final exam marks for the final grading.	The purpose of FA is for grading	Assessment of learning	
"Some of us, we try to upgrade improve our common practice by observing the current assessment needs by other advanced countries" (Teacher 2)	Teachers are taking their own initiative to learn from best practices of other advanced countries.	Teacher initiation to learn from the best practices	Continuous learning and improvement	
"We are always improving the formative assessment practices of my department according to the demands of this era of globalization. We analyse the assessment systems of the developed countries and upgrade our assessment system" (Teacher 5)	The formative assessment practices are being improved by reviewing and analysing the best practices of other developed countries.	The practice of formative assessment is improving by reviewing and analysing the best practices	Continuous learning and improvement	
"The progressive and open-minded teachers are very positive and try to apply formative assessment in their classes as much as possible" (Teacher 6)	Progressive and open minded teachers apply formative assessment frequently.	Progressive and open minded teacher	Positive attitude	Teacher's attitude
"Some teachers are more comfortable with traditional method of assessment as they are not ready to make a drastic change that differs from the norm and they resist to a new system of assessment" (Teacher 5)	Some teachers are habituated with SA and hesitate to make any changes.	Preference towards SA	Status quo	
In some cases, teachers lose their interest in applying formative assessment in their classes because of various difficulties like infra-structural problem, power cut, large classes etcetera that they face while doing activities" (Teacher 1)	Teachers lost interest to practice FA because of issues related to infrastructural problems, power cuts, large classes etc.	Disadvantages – Environmental factors	Negative attitude	
"Both the standard of education and students' results in the final examination was were much better than before. (Teacher 4)	Standard of education and students' results were improved.	FA contribute to enhance the quality of instruction and student outcome	Quality of instruction	Importance

"FA allowed us to engage and assess our students throughout the academic year" (Teacher 2)	FA allowed teachers to engage and assess students throughout the academic year"	Enhancement of student engagement Continuous assessment	Student outcome Student engagement	
"It also allowed the students to get in-depth knowledge on their course content and overcome their weaknesses" (Teacher 3)	FA allowed students to gain in-depth knowledge of the course content, improve their understanding	Student able to improve their understanding on a course content	Continuous assessment In-depth unders tanding of a course content	
"I like to apply FA in my class because I can identify different needs of the students with different abilities and fulfil them as much as possible" (Teacher 5)	FA allows teachers to identify different needs and abilities of the students; provide opportunity for teacher to address them and take the necessary actions	Teacher able to identify the different students' needs and abilities; take necess ary actions	Identification and rectification of Students' learning needs and abilities	
"There is a need to increase the weighting of FA in the overall assessment measure; more quality feedback, and some preconditions for implementing FA successfully in Bangladeshi colleges had to be fulfilled prior to that" (Majority of the teachers)	The weighting of FA in the overall assessment measure should be increased and quality feedback is needed. Some pre-conditions to implement FA should be fulfilled	To increase the weighting of FA and provide quality feedback Pre-conditions to improve the implementation of FA	Priority on FA with quality feedback Pre-conditions to improve the implementation of FA	
"The institution should enhance a higher level of transparency among the teachers, and to gain a better support from superior and support staffs are essential for a successful implementation of FA" (Teacher 5 and 6)	A higher level of transparency among teachers, support from superior, and assistance from support staff (administration staff) contribute to the successful implementation of FA.	A higher level of transparency among teachers, support from superior, and assistance from support staff (administration staff	Pre-conditions to improve the implementation of FA	
"It is difficult to implement FA in a large class within a short period of a particular lesson" (Teacher 2)	Difficult to implement FA in a large class within a short period of a particular lesson	Difficulty in implementation due to large class size and short time period	Large class size Limited time period	Challenges

“Many times when I plan for a FA via multimedia, there will be a power cut or problems with the internet connectivity” (Teacher 4)	FA is not possible when there is a power failure and problems with internet connectivity.	Inability to conduct formative assessment due to power failure and poor internet connectivity.	Power failure and poor internet connectivity.
“Some of the problems that we faced that hinder us to implement FA successfully in our class are such as resource scarcity, infrastructural issues, shortage of teachers, and heavy work load of the teachers” (Majority of the teachers)	Some of the problems faced by the teachers to implement FA were resource scarcity, infrastructural issues, shortage of teachers, and heavy work load.	Resource scarcity, infrastructural issues, shortage of teachers, and teacher’s heavy work load hinder the successful implementation of in-class formative assessments.	Resource scarcity, infrastructural issues, shortage of teachers, and teacher’s heavy work load
“We also faced some challenges when the issue of cultural challenges such as favouritism and biasness happened while grading the students through formative assessment” (Teacher 3 and 4)	Issue related to cultural challenges; favouritism and biasness while grading the students through formative assessment	Favouritism, Biasness	Cultural challenges
“Sometimes the teachers of Bangladesh tend to over-grade favourite students or under-grade unpopular students. ... I find problems like political pressure, administrative pressure and nepotism as the possible reasons behind teacher’s biasness” (Teacher 6)	Bangladeshi teachers tend to over-grade or under-grade students based on their popularity and involvement in political activity.	Political influence	Cultural challenges
“In this country, students of different higher educational institutions are largely engaged in political activities. As a result, in many cases teachers of Bangladeshi colleges have to face huge political pressure while grading a politically active student. Sometimes the pressure comes from the college administration, which can also be the result of due to higher political pressure from a higher authority” (Teacher 6)	Bangladeshi teachers are indirectly pressured to give a good grade to students who are engaged in political activity	Political influence	Cultural chall enges

The quantitative results and qualitative findings show a similar pattern of teachers' perceptions on formative assessment. This allows for triangulation to happen to enhance the validity and reliability of this study. Figure 2 shows the triangulation between the two which is discussed in detail in the next section.

Figure 2. *Triangulation of quantitative results and qualitative findings based on descriptive analysis and content analysis.*


Results/ Findings are intergrated and interpreted
Quantitative Results based on descriptive analysis
Qualitative Findings based on content analysis

4. DISCUSSIONS

4.1. Teacher's Prior Understanding of Formative Assessment in Higher Education

From the quantitative results, the three items; formative assessment is the most important type of assessment, it happens all through the academic year, and it provides quality feedback indicate that majority of the teachers have the prior knowledge about what formative assessment is. The findings from the qualitative study also revealed the same. Overall, all the participants considered formative assessment as a very important and useful tool in higher education, where most of them regarded it as the dominant form of assessment. This, as a whole, mirrors the findings of Dochy et al. (1999) and Bennett (2011). Both quantitative and qualitative studies show that the teacher participants acknowledged the benefits that students can gain from formative assessment. Some of the benefits outlined by the participants are enhancement of students' interest in learning and class engagement, and allows students to acquire in-depth knowledge. In line with what had been found in a study by Dochy et al. (1999), the participants also related the use of formative assessment with the development of various qualities among the students such as problem solving skills through effective application of formative assessment. Besides, it also enhances the learning of the students, as it makes them more independent and confident learners, and finally prepare them to cope with the competitive real world (Harlen & James, 1997; Dochy et al., 1999).

However, despite the positive perceptions of formative assessment, none of the participants believed that it should be used as the only assessment system for higher education in Bangladesh, particularly in the current socio-political condition of the country. All respondents also acknowledged the necessity of summative assessment with some regarded formative assessment as a complementary for summative assessment.

4.2. Current Practice of Formative Assessment

From the survey, 44% of the respondents strongly agreed and agreed that the current practice of formative assessment is satisfactory. This finding is supported by the qualitative findings. In the interviews, all respondents claimed that formative assessment was commonly used in their classrooms even though summative assessment was still the main form of assessment in the participating college. Common examples of formative assessment strategies used by majority of them are class test, questioning, quiz, competition, presentation, and workshop. In this context, teachers implemented different strategies depending on the class content, but the most common form of formative assessment was class test, which was practised by all the six interviewees. Muhammad et al. (2019) also found class test as a popular form of formative assessment in their study in Bangladeshi colleges.

Having said that, it is worth mentioning here that the implementation of formative assessment was mainly due to the instructions of the National University of Bangladesh. The practice of formative assessment by the majority of the participants is merely for grading their students, which indicates the assessment of learning. Nonetheless, the essence of formative assessment is not about grading per se, but to provide quality feedback so that students can continue to learn and improve (assessment of learning). According to Crooks (1988, p.468), "Too much emphasis has been placed on the grading function of evaluation and too little on its role in assisting students to learn". For a formative assessment to be effective especially for students' learning, providing feedback is necessary. This is supported by Black and William (1998). Hence, it is essential for Bangladeshi teachers to incorporate feedback in any forms of formative assessment. Formative assessment should provide information about the learning process that teachers can use for instructional decisions and students can use in improving their performance, which motivates students (William, 2011).

Some teachers also try to improve their practice by observing the current teaching and learning needs of other advanced countries. They take initiative to improve the formative assessment practices of their department according to the demands of this era of globalization. Some of the teachers analyse the assessment systems of the developed countries and upgrade their own assessment system. Most of the participants reported that students enjoyed the implemented formative assessment tasks and it enhanced students' attention in classrooms. However, few teachers understood the right approach to practice formative assessment and the practical impact is very minimal.

4.3. Teachers' Attitude towards Formative Assessment

The quantitative findings indicate that majority of the respondents portrayed a positive attitude towards the practice of formative assessment. However, from the qualitative findings, the results were mixed. Some of them were positive about the use of formative assessment while some were more cautious about it. The interviewees also revealed that most of the teachers at the participating college were positive about formative assessment, but some were negative due to some reasons which had been highlighted in prior studies by Dijksterhuis et al. (2013), López-Pastor and Sicilia-Camacho (2017), Sach (2012); Winstone & Millward (2012), and Young & Jackman (2014). The progressive and open-minded teachers are very positive and try to apply formative assessment in their classes as much as possible. On the other hand, some teachers were being negative toward formative assessment because there are more comfortable with traditional method of assessment as they are not ready to make a drastic change that differs from the norm and they resist to a new system of asessment. Besides they do not have a clear idea of the ways to implement formative assessment appropriately and as a result they lack confidence to apply a new assessment format. The earlier-reviewed study by Young & Jackman (2014) also indicated this issue.

In some cases, teachers lose their interest in applying formative assessment in their classes because of various difficulties like infrastructural problem, power cut, large classes, etc. These issues are also acknowledged in the study of Muhammad et al. (2019). Nonetheless, research conducted by Sadler (1998), and Umer and Omer (2015) claimed that formative assessment can be implemented successfully in any kind of educational setting.

4.4. Importance of Implementing Formative Assessment

All the interviewees considered formative assessment as a useful tool for enhancing the quality of teaching and learning in higher education especially in Bangladeshi colleges as being agreed by Bennett (2011) and Dochy et al. (1999). They claimed that the standard of assessment and students' results in the final examination were better than before. Interviewees also acknowledged some benefits of formative assessment. During the interviews, it was mentioned that formative assessment allowed them to engage and assess their students throughout the academic year and it also allowed the students to get in-depth knowledge on their course content and overcome their weaknesses. These findings supported the quantitative results. In addition, the interviewed teachers also claimed that it had helped to enhance their students' learning outcomes. Besides, teacher were able to identify different needs of the students with different abilities and fulfil them as much as possible.

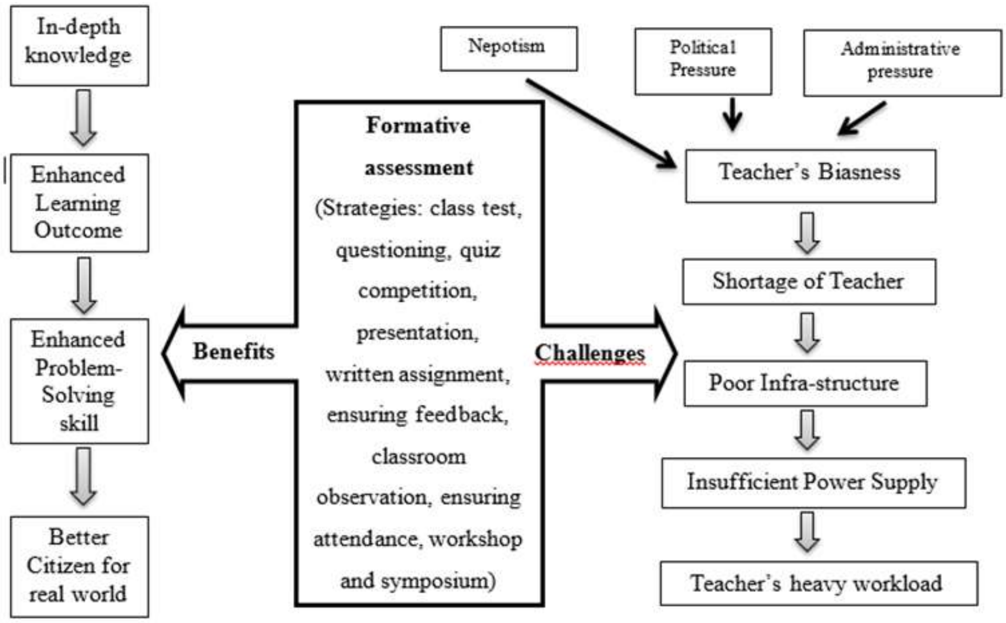
Although all the participants were positive towards the implementation of formative assessment, they defer on the way of grading through this type of assessment. Majority of the participants believed that it was necessary to increase the weighting of formative assessment in the overall assessment measure and it must be accompanied with quality feedback. Besides, they also believed that some preconditions for a successful implementation of formative assessment in Bangladeshi colleges had to be fulfilled prior to that. They also suggested that higher levels of transparency among the teachers, and better support from superior and support staff were also essential for a successful implementation of formative assessment. In contrast, some others claimed that the current weighting of assessments were appropriate in the context of Bangladesh.

4.5. Challenges in Implementing Formative Assessment

As frequently highlighted in the literature, there are some challenges in implementing formative assessment in Bangladesh higher education institutions. In line with what had been found in prior studies by Alam et al. (2014) and Muhammad et al. (2019), problems such as resource scarcity, infrastructural issues, insufficient power supply, shortage of teachers, large size class and heavy work load of the teachers were cited as some of the challenges that the interviewed teachers were facing when implementing formative assessment in their classrooms. In addition, these teachers also faced some challenges when grading their students using formative assessment (the marks of in-course examination for the Bangladeshi colleges). In this regard, sometimes teachers tend to over-grade favourite students or under-grade unpopular students because of the problems like political pressure, administrative pressure and nepotism. This might hinder the implementation of formative assessment successfully in a larger scale in Bangladeshi colleges.

In Bangladesh, students of different higher educational institutions are largely engaged in political activities. As a result, in many cases teachers have to face huge political pressure while grading a politically active student. Sometimes the pressure comes from the college administration, which can also be due to political pressure from a higher authority. Based on the responses given by teacher participants, the following conceptual framework has emerged:

Figure 3. *Conceptual framework for formative assessment in Bangladeshi Colleges.*


In-depth knowledge
Enhanced Learning Outcome
Enhanced Problem-Solving skill
Better Citizen for real world
Benefits
Formative assessment (Strategies: class test, questioning, quiz competition, presentation, written assignment, ensuring feedback, classroom observation, ensuring attendance, workshop and symposium)
Challenges
Nepotism
Political Pressure
Administrative pressure
Teacher's Biasness
Shortage of Teacher
Poor Infra-structure
Insufficient Power Supply
Teacher's heavy workload

Overall, the above findings reveal that formative assessment is being practised by majority of the respondents. Since the respondents have working experience in other colleges, it could be said, there is a high possibility that other higher educational institutions are also practising some kind of formative assessment but it only carries a small weighting in the overall. To some extent, responses from the interview sessions strengthen the survey findings, which suggest that teachers in this study believe that formative assessment is beneficial for the students and it has a great potential to positively enhance the teaching and learning in Bangladeshi higher education. Nonetheless the effective use of formative assessment is very limited because of the related challenges as depicted in Figure 3.

5. CONCLUSION

This study is a mixed-method study exploring the use of formative assessment in a higher education institution in Bangladesh. After analysing the quantitative and qualitative data, it is found that, formative assessment, has become increasingly more common although not being widely used in the context of higher education. In this regard, many teachers take this form of assessment positively in most of the cases. Authors believe that the practice of formative assessment could be enhanced if the highlighted challenges are overcome. These issues and challenges must be addressed accordingly and systematically.

Drawing on the findings of this study, some measures to enhance the formative assessment practice in Bangladeshi higher education institutions should be considered. First, the pressures on teachers have to be minimised to get the best result from formative assessment. Instead, teachers should be empowered or given security from any kind of harm. Secondly, the highlighted resource problems should also be addressed, and a smaller class size is also recommended. In addition, more teachers have to be recruited to reduce the workload of the teachers.

This research is only an initial study to explore the teachers' perception on the formative assessment practice in Bangladeshi higher education institution which will open up a greater avenue for further study on the assessment system. It is our hope that our research will contribute positively to the development of a better assessment system in higher education in

Bangladesh, which will ultimately contribute to the enhancement of the teaching and learning in Bangladesh higher 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 authors. Ethics Committee Number: The University of Nottingham, EDUC-4227-UNMC (Practice Based Inquiry).

Authorship Contribution Statement

Shamsiah Banu Mohamad Hanefar: Spervision, Methodology, validation, proof-reading, resources, writing **Nusrat Zerin Anny:** Data collection, investigation, resources writing. **Md. Sajedur Rahman:** Resources investigation.

Orcid

Shamsiah Banu Mohamad Hanefar https://orcid.org/0000-0002-3393-1640
Nusrat Zerin Anny https://orcid.org/0000-0001-8559-7157
Md. Sajedur Rahman https://orcid.org/0000-0003-4525-0663

REFERENCES

Ahmed, P.M., Rahman, P.D.M., Islam, P.D.S. M., Nayeem, P.D.A. I., Khan, P.D. Z.R., Rahman, P.M.H., Islam, M.Z., & Chackrabartty, S. (2021). *Quality and Relevance of Higher Education in Colleges Affiliated with National University Bangladesh: A Bachground Study*. https://cedp.gov.bd/wp-content/uploads/2021/03/Quality and Relevance.pdf

Alam, G.M., Mishra, P.K., & Shahjamal, M.M. (2014). Quality assurance strategies for affiliated institutions of HE : A case study of the affiliates under National University of Bangladesh. *Higher Education*, (68), 285–301. https://doi.org/10.1007/s10734-013-9712-y

Antoniou, P., & James, M. (2013). Exploring formative assessment in primary school classrooms: Developing a framework of actions and strategies. *Educational Assessment, Evaluation, and Accountability*, 25(4). https://doi.org/10.1007/s11092-013-918

Azim, F. (2014). Learning promoted by classroom assessment in higher education of Bangladesh: A case study. *The International Journal of Social Sciences*, *26*(1), 165–171.

Begum, M., & Farooqui, S. (2008). School-based assessment: Will it really change the education scenario in Bangladesh? *International Education Studies*, 1(2), 45–53. https://doi.org/10.5539/ies.v1n2p45

Bengtsson, M. (2016). How to plan and perform a qualitative study using content analysis. *NursingPlus Open*, 2, 8-14.

Bennett, R.E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy, and Practice*, *18*(1), 5-25. https://doi.org/10.1080/0969594X.2010.513678

Bhuiyan, A.K.M.Z.H., & Hossain, D.M.L. (2017). Quality education at college level institutions: Bangladesh perspective. *NAEM Journal*, *12*(24), 103–112.

Bjornsrud, H., & Engh, R. (2012). Teamwork to enhance adapted teaching and formative assessment. *Policy Futures in Education*, *10*(4), 402-410. https://doi.org/10.2304/pfie.2012.10.4.402

Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7-74. https://doi.org/10.1080/0969595980050102

Black, P., & Wiliam, D. (2003). In praise of educational research: Formative assessment. *British*

Educational Research Journal, *29*(5), 623-637. https://doi.org/10.1080/0141192032000133721

Boston, C. (2002). The Concept of Formative Assessment. *ERIC Clearinghouse on Assessment and Evaluation College Park MD.*, *1*(1), 1–8. https://doi.org/ED47026, 2002-10-100

Bransford, J.D., Brown, A.L., & Cocking, R. R. (Eds.). (1999). *How people learn: Brain, mind, experience, and school*. National Academy Press.

Cagasan, L., Care, E., Robertson, P., & Luo, R. (2020). Developing a formative assessment protocol to examine formative assessment practices in the Philippines. *Educational Assessment*, 1–17. https://doi.org/10.1080/10627197.2020.1766960

Clark, I. (2012). Formative assessment: A systematic and artistic process of instruction for supporting school and lifelong learning. *Canadian Journal of Education*, *35*(2), 24–40. https://doi.org/10.1007/s10648-011-9191-6

Clarke, M. (2012). What matters most for student assessment systems: A framework paper. *Systems Approach for Better Education Results (SABER) student assessment working paper*, no. 1, World Bank. http://documents.worldbank.org/curated/en/216631468149691772/What-matters-most-for studentassessment-systems-a-framework-paper

Crooks, T.J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, 58 (4), 438-481. https://doi.org/10.3102%2F00346543058004438

Dijksterhuis, M.G.K., Schuwirth, L.W.T., Braat, D.D.M., Teunissen, P.W., & Scheele, F. (2013). A qualitative study on trainees' and supervisors' perceptions of assessment for learning in postgraduate medical education. *Medical Teacher*, *35*(8). https://doi.org/10.3109/0142159X.2012.756576

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education*, *24*(3), 331–350. https://doi.org/10.1080/03075079912331379935

Figa, J.G., Tarekegne, W. M., & Kebede, M. A. (2020) The practice of formative assessment in Ethiopian secondary school curriculum implementation: The case of West Arsi Zone Secondary Schools. *Educational Assessment*, 25(4), 276-287. https://doi.org/10.1080/10627197.2020.1766958

Gikandi, J.W., Morrow, D., & Davis, N.E. (2011). Online formative assessment in higher education: A review of the literature. *Computers and Education*, *57*(4), 2333–2351. https://doi.org/10.1016/j.compedu.2011.06.004

Haque, M., Yousuf, R., Abu Bakar, S.M., & Salam, A. (2013). Assessment in undergraduate medical education: Bangladesh perspectives. *Bangladesh Journal of Medical Science*, *12*(4), 357–363. https://doi.org/10.3329/bjms.v12i4.16658

Harlen, W. & James, M. (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice*, *4*(3), 365–379. https://doi.org/10.1080/0969594970040304

Jeffrey A. Barney & Robert McCowens. (2009). Review of transformative assessment by W. James Popham. *Association for Supervision and Curriculum Development*, *6*(12), 137–138. https://doi.org/ISBN-978-1-4166-0667-3

Klenowski, V. (2009). Assessment for learning revisited : An Asia- Pacific perspective. *Assessment in Education : Principles, Policy and Practice*, *16*(3), 263–268. https://doi.org/10.1080/09695940903319646

López-Pastor, V., & Sicilia-Camacho, A. (2017). Formative and shared assessment in higher education. Lessons learned and challenges for the future. *Assessment and Evaluation in Higher Education*, *42*(1), 77–97. https://doi.org/10.1080/02602938.2015.1083535

Mahmud, M. (2019). Quality and Relevance of Higher Education in Bangladesh. *BIDS Critical Conversations 2019*. Bangladesh Institute of Development Studies.

Mamun-ur-Rashid, M., & Rhman, M.Z. (2017). Quality of higher education in Bangladesh:

Application of a modified SERVQUAL model. *Problems of Education in the 21st Century*, *75*(1), 72–91. https://doi.org/10.33225/pec/17.75.72

Monem, M., & Baniamin, H.M. (2010). Higher education in Bangladesh : Status, issues, and prospects. *Pakistan Journal of Social Sciences (PJSS)*, *30*(2), 293–305.

Al Faruki, M., J., Haque, M.A., & Islam, M.M. (2019). Student-centered learning and current practice in Bangladeshi college education. Journal of Education and Practice, 10(13), 95–107. https://doi.10.7176/JEP/10-13-11

Ngendahayo, E. (2014). Rethinking Rwandan higher education assessment system and approaches. *Rwandan Journal of Education*, *2*(2), 31–47.

Nicol, D., & MacFarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199–218. https://doi.org/10.1080/03075070600572090

Ozan, C., & Kıncal, R.Y. (2018). The effects of formative assessment on academic achievement, attitudes toward the lesson, and self-regulation skills. *E*ducatıonal Scıences: Theory & Practıce, *18*(1), 85–118. https://doi.org/10.12738/estp.2018.1.0216

Paul, B.K., Sarkar, S., Nandi, S., Alim, M.A., Biswas, S.K., & Rahman, M.A. (2016). Changing teaching through formative assessment: A review. *Bangladesh Medical Journal*, *45*(1), 47–53. https://doi.org/10.3329/bmj.v45i1.28968

Pintrich P.R., & Zusho A. (2002) Student motivation and self-regulated learning in the college classroom. In: Smart J.C., & Tierney W.G. (eds). Higher Education: Handbook of Theory and Research, 17. Springer. https://doi.org/10.1007/978-94-010-0245-5_2

Rabiul, I. (2014). Higher education in Bangladesh : Diversity, quality and accessibility. *First National Education Conference on Whither Policy Reform in Education: Lessons and Challenges*, (November), 1–16.

Rahman, K.A., Hasan, M.K., Namaziandost, E., & Ibna Seraj, P. M. (2021). Implementing a formative assessment model at the secondary schools: attitudes and challenges. *Language Testing in Asia*, *11*(1). https://doi.org/10.1186/s40468-021-00136-3

Rahman, T., Nakata, S., Yoko, N., Mokhlesur, R.M., Uttam, S., & Rahman Asahabur, M. (2019). *Bangladesh Tertiary Education Sector Review*. World Bank.

Riaz, F., Yasmin, S., & Yasmin, R. (2015). Introducing regular formative assessment to enhance learning among dental students at Islamic International Dental College. *Journal of the Pakistan Medical Association*, *65*(12), 1277–1282.

Roos, B., & Hamilton, D. (2005). Formative assessment: A cybernetic viewpoint. *Assessment in Education: Principles, Policy & Practice*, *12*(1), 7-20. https://doi.org/10.1080/0969594042000333887

Ruiz-Primo, M.A. (2011). Informal formative assessment: The role of instructional dialogues in assessing students' learning. *Studies in Educational Evaluation*, *37*(1), 15–24. https://doi.org/10.1016/j.stueduc.2011.04.003

Sach, E. (2012). Teachers and testing: An investigation into teachers' perceptions of formative assessment. *Educational Studies*, *38*(3), 261-276. https://doi.org/10.1080/03055698.2011.598684

Sadler, D.R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–144. https://doi.org/10.1007/BF00117714

Sadler, D. R. (1998). Formative Assessment: revisiting the territory. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 77–84. https://doi.org/10.1080/0969595980050104

Shahidul, H.M.M. (2016). Rethinking higher education. *The Daily Star*.

Sharma, S., Sharma, V., Sharma, M., Awasthi, B., & Chaudhary, S. (2015). Formative assessment in postgraduate medical education - Perceptions of students and teachers. *International Journal of Applied & Basic Medical Research*, 5(Suppl 1), S66–S70. https://doi.org/10.4103/2229-516X.162282

Umer, M., & Attayib Omer, A.M. (2015). An investigation of Saudi English: Major learners' perceptions of formative assessment tasks and their learning. *English Language Teaching*, *8*(2), 109–115. https://doi.org/10.5539/elt.v8n2p109

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. https://doi.org/10.1111/j.1365-2729.2006.00166.x

Wiliam, D. (2011). What is assessment for learning?. *Studies in Educational Evaluation*. 37(1), 3-14. https://doi.org/10.1016/j.stueduc.2011.03.001

William, D. (2014). Formative assessment and contingency in the regulation of learning processes. *Toward a Theory of Classroom Assessment as the Regulation of Learning*, (April). Philadelphia, USA.

Winstone, N., & Millward, L. (2012). Reframing Perceptions of the Lecture from Challenges to Opportunities: Embedding Active Learning and Formative Assessment into the Teaching of Large Classes. *Psychology Teaching Review*, *18*(2), 31–41.

World Bank. (2013). *Systems Approach for Better Education Results (SABER) reports*.

Yorke, M. (2003). Formative assessment in higher education : Moves towards theory and the enhancement of pedagogic practice. *Higher Education*, *45*, 477–501.

Young, J.E.J., & Jackman, M.G.A. (2014). Formative assessment in the Grenadian lower secondary school: teachers' perceptions, attitudes and practices. *Assessment in Education: Principles, Policy and Practice*, *21*(4), 398-411. https://doi.org/10.1080/0969594X.2014.919248

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 1, 80–97

https://doi.org/10.21449/ijate.1059304

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

Standard Setting in Academic Writing Assessment through Objective Standard Setting Method

Fatima Nur Fisne[1,*], Mehmet Sata[2], Ismail Karakaya[3]

[1]Gazi University, Gazi Faculty of Education, English Language Teaching Program, Turkiye
[2]Agri Ibrahim Cecen University, Faculty of Education, Department of Measurement and Evaluation in Education, Turkiye
[3]Gazi University, Gazi Faculty of Education, Department of Measurement and Evaluation in Education, Turkiye

ARTICLE HISTORY

Received: Mar. 19, 2021
Revised: Nov. 19, 2021
Accepted: Jan. 13, 2022

Keywords:

L2 academic writing assessment,
Many-facet Rasch measurement model,
Standard setting,
OSS

Abstract: Performance standards have important consequences for all the stakeholders in the assessment of L2 academic writing. These standards not only describe the level of writing performance but also provide a basis for making evaluative decisions on the academic writing. Such a high-stakes role of the performance standards requires the enhancement of objectivity in standard setting procedure. Accordingly, this study aims to shed light upon the usefulness of Objective Standard Setting (OSS) method in specifying the levels of proficiency in L2 academic writing. On the basis of the descriptive research design, the sample of this research includes the examinees and raters who were student teachers at the university level. Essay task and analytical writing scoring rubric were employed as the data collection tools. In data analysis, OSS method and two-step cluster analysis were used. The analysis results of OSS method based on many-facet Rasch measurement model (MFRM) outline the distribution of the criteria into the levels of proficiency. Also, the main findings in OSS method were validated with two-step cluster analysis. That is, OSS method may be practically used to help the stakeholders make objective judgments on the examinees' target performance.

1. INTRODUCTION

As a multidimensional field, assessing writing in an academic context has been the focus of attention in second language (L2) assessment in recent years. As an essential component of this field, standard setting serves as a basis for defining the levels of language attainment. Also, it provides evidence for decision makers to make instructional judgments on target performance. For this reason, there has been a growing interest in setting L2 academic writing standards over the years.

In broad terms, standard setting is viewed as "the process of determining cut-scores for a test" (Davies et al., 1999, p. 186). These cut-off scores may be single (e.g. pass/fail) or multiple (e.g. level of achievement) (Khatimin et al., 2013). In other words, "setting standards on educational assessments sometimes requires a single level" or "more than two stages or degrees of performance" (Cizek, 1993, p. 92-93). These single or multi-level standards have important

*CONTACT: Fatima Nur FISNE ✉ fatimanurfisne@gazi.edu.tr Gazi University, Gazi Faculty of Education, English Language Teaching Program, Turkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

consequences on stakeholders such as test-takers, instructors, and policy-makers. For example, standard setting is taken into account in making judgments on the placement of the examinees into the appropriate levels (Shin & Lidster, 2017). Furthermore, standard setting procedure may directly influence the whole decision-making process in an educational system (Sondergeld et al., 2020). This aspect of standard setting is primarily related to the decision validity that represents the quality and consistency of the educational decisions (Erkus et al., 2017). Such a significant role of standard setting requires the use of objective methods in setting cut-off scores because the assessment results are open to discussion when the cut-off scores are not set properly (Bejar, 2008). However, Sireci et al. (1997) state that "the most popular methods for setting passing scores and other standards on educational tests rely heavily on subjective judgment" (p. 3). Likewise, Davis-Becker et al. (2011) reveal that standard setting is generally viewed as "one of the most subjective and judgmental components" in spite of the pivotal importance of standard setting "in the test development and validation process" (p. 25) Accordingly, there is a need for more objective methods to determine more valid standards in the educational measurement. In addition, the performance levels should be objectively defined in L2 writing assessment to help the stakeholders reach a valid decision.

In order to meet the needs of objectivity in standard setting procedures and ensure the decision validity in L2 writing assessment, this research mainly utilizes OSS method in defining the levels of target performance objectively, determining a valid cut-off score, and then making objective decisions about the students' performance in L2 academic writing.

1.1. Review of Literature

Standard setting basically refers to "setting cutscores" in assessment (Sireci et al., 1997, p. 3). More specifically, "performance standards specify what level of performance on a test is required for a test taker to be classified into a given category" and the process of defining these levels is called standard setting (Cizek, 2012, p. 4). As it functions as a benchmark to define target performance levels and provides a basis for performance-related decisions, it is an essential part of the educational assessment and evaluation.

There are various standard setting methods that are used to determine performance standards. These methods are basically grouped within two categories: test-based and examinee-based standard setting (Yudkowsky et al., 2009). In test-based methods like Angoff (1971) and Ebel (1972), judges examine the test itself and test items and predict the level of the target performance. On the other hand, in examinee-based methods like the Borderline Group and Contrasting-Groups, judges mainly focus on test takers' performance, gather evidence on the performance levels and then estimate the standards. Livingston and Zieky (1982) provide a comprehensive overview of the commonly used standard setting methods. To illustrate, Nedelsky method (1954), which is one of the earliest methods, is used to determine the passing score for multiple-choice tests. In this method, judges attempt to define the wrong answers that a borderline test taker would recognize. Calculations are carried out through the elimination of the possible wrong answers. In Angoff method (1971), unlike Nedelsky, judges examine each item holistically without considering the possible wrong answers and make estimations on whether borderline test takers would be able to give a correct answer to each item. Based on the probability of the correct answers, passing score is calculated. In Ebel method (1972), judges make decisions by considering the difficulty and relevance levels of the items. In this method, a matrix including the dimensions of the difficulty (i.e. easy, medium, hard) and relevance (essential, important, acceptable, questionable) is constructed, and test items are placed into the appropriate cells. Following that, judges predict the possible correct answers. Passing score is defined on the basis of the calculations including the percentage of correct answers. As for the examinee-based methods, the Borderline Group method focuses on test takers' performance and requires judges to identify the borderline test takers in terms of target knowledge and skills. Passing score is set according to the median of the scores that are assigned to the borderline test takers. In the Contrasting-Group method, test takers are divided into two groups in consideration of their level of knowledge and skills, and passing score depends on the degree at which there is almost equal number of the test takers from both groups.

Sondergeld et al. (2020) assert that there are some concerns on the use of traditional standard settings methods, and to tackle these concerns, modern methods mainly based on item response theory (IRT) have been introduced. OSS method is one of these methods that aim to minimize the problems faced in setting standards like subjectivity and rater agreement/disagreement. It is basically established on test content rather than the direct expert opinions (Stone, 2001). Expert judgments are also used in this method, but the goal is not to specify the ratio/number of the correct responses; instead, experts discuss the essential content that might indicate the test takers' achievement (Sondergeld et al., 2020). As one of the modern criterion-based standard setting methods (Bichi et al., 2019), OSS method based on Rasch measurement model and Wright and Grosse's (1993) standard setting principles considers the expert opinions, test takers' performance and test/item difficulty at the same time (Khatimin et al., 2013). Through Rasch model, the measurement outputs are displayed on the logit scale, and the raw score can be analyzed on this scale in regard to the task/content achievement (Sondergeld et al., 2020). There are three important steps in OSS method: "defining the criterion set", "refining the criterion point", and "expressing the error" (Stone et al., 2011, p. 950). Hence, OSS method enables the examiners to analyze the level of performance in consideration of the standard error of measurement.

In the relevant literature, some research studies use and compare the standard setting methods, and examine the effectiveness and utility of these methods. For example, Davis-Becker et al. (2011) examined the Bookmark method in terms of item-ordering. Stone et al. (2011) compared OSS method and the Angoff approach on a longitudinal basis. In another study, Shin and Lidster (2017) discussed the comparative effectiveness of the Bookmark method, the Borderline group method, and cluster analysis in ESL (English as a Second Language) placement context. MacDougall and Stone (2015) emphasized the strengths of OSS method in standard setting procedure. In the research context of L2 writing assessment, some standard setting studies are related to the alignment of examinations to the Common European Framework of References (CEFR) that attempts "to describe the levels of proficiency required by existing standards, tests and examinations" (CoE, 2001, p. 21). In these studies, it is intended to link some language exams to the CEFR levels. For example, Tannenbaum and Wylie (2008) aimed to define the cut scores for two large-scale tests in accordance with the CEFR levels. Green (2018) investigated the English for Academic Purposes (EAP) context in terms of relating EAP testing to the CEFR. Fleckenstein et al. (2020) put emphasis on the writing profiles of students and tried to link EFL writing competences to the CEFR.

As a productive skill, writing encompasses different social, cultural and cognitive dynamics (Weigle, 2002). Owing to its dynamic structure, the assessment of L2 academic writing skills should be constructed on the systematic basis that entails the operationalization of the task characteristics and underlying dimensions. Harsch and Rupp (2011) call attention to the use of open-tasks in writing assessment and its advantages in enabling the examinees to produce a broad variety of written output. Use of these tasks in L2 writing assessment requires the attribution of levels or numerical values to target writing performance by the raters. In this respect, the rater-related issues are scrutinized in L2 writing assessment research. For example, Schaefer (2008) focused on the rater bias patterns in EFL writing assessment. The study findings pointed out that some criteria were severely rated whereas raters were lenient in some other criteria. Also, severity and leniency behaviours changed according to the students' level of ability in writing. Goodwin (2016) analyzed the rater behaviours in an academic language test aiming at both reading and writing skills and found differences between the attributions of the scores in admission and placement tests. Trace et al. (2017) underlined the importance of rater negotiation and explicated its effect on reducing rater bias in writing performance assessment. Elder et al. (2007) paid attention to the rater subjectivity and bias. Along with the rater behaviours in L2 writing assessment, the presentation of objective and valid performance standards to the raters is another crucial issue. In some settings, assessing writing has large-scale outcomes like "promotion, placement, and admission" (Wind & Engelhard, 2013, p. 297), and therefore objective performance standards should be given to the raters in order to provide absolute and credible evidence for decision-makers. In this regard, the importance of standard setting in L2 writing assessment comes into prominence. From this perspective, this research aims to investigate the usefulness of OSS method in determining objective and valid cut-off scores and performance standards in L2 academic writing assessment. The following research questions guide the researchers to explore the utility of OSS method in setting objective standards in L2 writing assessment:

1. What are the procedures of setting objective standards in L2 academic writing assessment through OSS method?
2. To what extent are OSS method-based decisions validated?

2. METHOD

2.1. Research design and participants

This study adopts a descriptive research design that presents the researchers with opportunities to elaborate on the variables to be examined (Best & Khan, 2006). Within the framework of the descriptive research, this study attempts to explain how useful OSS Method is in defining the cut-off scores and standards of L2 academic writing proficiency. The subject group includes 64 raters and 39 examinees who were the student teachers in the department of English language teaching (ELT) at the tertiary level. The raters were the third graders taking Educational Measurement course, and they were familiar with not only the process of academic writing but also the assessment of writing performance. Since the number of raters plays an essential role in standard setting procedures, the 3rd grade student teachers (n = 64) were selected as the participants with the practical purposes. With respect to the demographics of the raters, the mean age was 21.84. While 12 raters (18.78%) were male, 52 of the raters (81.25%) were female. As for the examinees, they were the first graders attending Advanced Reading and Writing course II at the same department. They completed Advanced Reading and Writing course I in the fall term and reinforced their knowledge and skills on how to develop outlines, specify topic, thesis and supporting sentences, sequence their ideas in a logical way, and ensure task achievement.

2.2. Data Collection

In this research, there are two sequential steps in the data collection. First, a sample writing task of IELTS (The International English Language Testing System)* was used. This task requires the examinees to write an essay by expressing agreement or disagreement on the given topic. This sample task was selected owing to the authenticity of the topic in which the examinees might address their real experiences. After they completed the task, in the second step, the essays were anonymously distributed to the raters, and each rater scored all the essays individually. With the aim of scoring the academic essays, Analytical Scoring Rubric for Academic Essays (ASRAE) was developed by the researchers (see Appendix A). This rubric

* This task was taken from the section of Sample Test Questions/Academic Writing on the official web page of IELTS (https://www.ielts.org/)

was also used in a different study conducted by the researchers (Sata & Karakaya, 2021). As explained in this study, ASRAE includes seven main criteria and 16 sub-criteria that intend to measure the components of academic writing. The main logic behind the development of this rubric is the characteristics of the target participants. Since the examinees were highly proficient in L2 and attained a mastery level in academic writing, the researchers felt the necessity to develop such kind of a rubric. The content validity of ASRAE was ensured through analyzing the essays, reviewing the literature review and calculating content validity index and ratio proposed by Lawshe (1975). For construct validity, exploratory factor analysis (EFA) was conducted, and EFA results indicated that the explained total variance was .73 with one-factor structure. In what follows the validation of the content and construct, the reliability of the rubric was determined. For this calculation, the reliability coefficient (ω) suggested by McDonald (1999) was employed, and the results show that McDonald ω coefficient was .97 (95% reliability interval: .96-.98) as elucidated in detail in Sata and Karakaya (2021). To sum up, validity and reliability results point out that ASRAE can be used as a reliable and valid tool to measure L2 academic writing proficiency.

2.3. Data Analysis

In order to analyze the rater scoring, set objective standards for academic writing proficiency, and validate these standards, OSS method based on MFRM and two-step cluster analysis were used in the current research. There were three main facets used in the Rasch analysis for OSS method: raters, examinees, and criteria. The raters scored each individual essay by considering the criteria given in ASRAE. So, fully crossed design was employed in this analysis. In line with three steps given in OSS method (Stone et al., 2011), four steps were followed in this analysis: (1) ensuring content validity, (2) specification of the performance levels, (3) difficulty level and standard errors, and (4) determination of proficiency levels.

OSS method requires meeting some assumptions of MFRM such as unidimensionality, local dependence, and model-data fit. With the aim of ensuring these assumptions, some analyses were conducted. Firstly, EFA was employed to test the unidimensionality, and the EFA results display that the factor structure is unidimensional. Following that, G2 statistics (Chen & Thissen, 1997) was used to test the local dependence. The results point out that LD χ2 values estimated for each criterion pairs are below 10. This result could be viewed as the indicator of the local dependence. As for the assumption of the model-data fit, the standardized values were examined. Linacre (2017) states that in order to ensure model-data fit, the number of the standardized values which are not between -2 and +2 should not exceed 5% of all the data. In this research, the number of total data was 37396, and the number of the standardized values that are not between -2 and +2 is 1547 [%4.14]). It is seen that there is a fit between model and data. Besides that, it is also crucial to examine the fit values of the target items (Khatimin et al., 2013). Accordingly, biserial correlation, outfit values, and standards of outfit values were examined to identify the misfit items (see Appendix B). According to the results, biserial correlations(x) are between .17 and .51, outfit values (MNSQ) are between 0.71 and 1.38, and the standards of outfit values (ZSTD) are between -9.00 and 9.00 (fit values: $0.4<x<0.8$, $0.5<MNSQ<1.5$ and $-2.0<z<2.0$). That is to say, there is no misfit in the dataset except for the standards of the outfit values. Holistically speaking, all the assumptions of OSS method were tested and ensured. It is noteworthy to state that the criterion of "Title of Essay" was excluded from the analysis of the OSS method since most examinees did not write a title for their essays unintentionally, and this exceptional case might cause an invalid standard setting. On the other hand, "Title of Essay" is still the component of the rubric (see ASRAE in Appendix A).

As the second data analysis method, two-step cluster analysis was conducted to provide evidence on the validity of the performance standards to be set through OSS method because Khalid (2011) explains that clustering analysis is based on less subjective process. The avoidance of subjectivity is the primary rationale to choose cluster analysis as the validation tool of OSS-method results. Cokluk et al. (2012) explain the main function of cluster analyzing as the identification of the similarities among the items/examinees and classification of them according to these similarities. To be more specific, cluster analysis can categorize target groups according to "distance" and "similarity" (Violato et al., 2003, p. 62). That is why this technique was selected to clarify and confirm OSS method results. The important assumptions to be met in the cluster analysis are the representativeness of the universe and avoidance of multicollinearity problem and outliers (Kayri, 2007). In this study, there are not any multicollinearity problems between variables and outliers in the datasets. However, larger samples may be required to offer more representativeness for the universe. So, the analysis results will be discussed in the target sample of the participants in this research.

3. RESULTS

This section elaborates on the standard setting procedures in L2 writing assessment through OSS method, identification of the cut-off score, and then presents the consistency between the results of OSS Method and two-step cluster analysis.

3.1. Standard Setting through OSS Method

In line with the first step given in OSS method, the content validity of ASRAE or definition of the criteria/content was ensured through expert opinions. 11 experts evaluated the appropriateness of the criteria on the basis of the rubric construct and components. They were asked to decide whether or not the criteria are essential. According to the expert judgments, content validity ratio (CVR) and index were calculated (Lawshe, 1975). CVR was reported as .75 and this value is above .59 that indicates the evidence for content validity with 11 experts (Wilson et al., 2012). The results show that the rubric has the content validity at the expected level.

Table 1. *Distribution of the criteria to the specified criterion points*

Criterion Points	Criteria	Logit Value	Standard Error
Criterion Point 1	Syntactic Complexity	0.43	0.02
	Idea Development	0.35	0.02
	Topic Sentence	0.34	0.02
	Lexical Range	0.33	0.02
Criterion Point 2	Thesis Statement	0.30	0.02
	Supporting Sentence	0.28	0.02
	Linking	0.24	0.02
Criterion Point 3	Accuracy of Grammatical Forms	0.01	0.03
	Coherence	-0.01	0.03
	Introduction-Body-Conclusion	-0.07	0.03
	Word Choice	-0.07	0.03
Criterion Point 4	Topic Relevance	-0.34	0.03
	Appropriate Length	-0.49	0.03
	Punctuation	-0.59	0.03
	Spelling	-0.70	0.03

The second step guides the specification of the performance levels in L2 academic writing. In this respect, the field experts suggested five levels of academic writing in consideration of the CEFR as A2, B1, B2, C1, and C2. The reason why the level of A1 is not included in this specification is that the examinees, who were student teachers in ELT department, had L2 writing experiences and had been practicing English language writing for a long time. In accordance with these five proficiency levels, four criterion points were defined for the analysis through OSS Method. When the number of the criteria was divided by the number of the criterion points (15/4 = 3.75), the number of the criteria to be assigned to each level was found as 3.75. That is, each level requires the proficiency almost in four criteria. Table 1 illustrates the distribution of the rubric criteria to the criterion points that are specified above. In the third step of OSS method, the mean difficulty levels and mean standard errors were calculated for each criterion point. These difficulty levels and standard errors are given in Table 2. In this table, negative logit values represent the criteria that are relatively easy for the examinees to achieve. On the other hand, positive logit values indicate relatively more difficult criteria in L2 academic writing assessment.

Table 2. *Mean difficulty and mean standard errors of the criterion points.*

Criterion Point	Mean Logit Value	Mean Standard Error
Criterion Point 1	+0.36	0.02
Criterion Point 2	+0.27	0.02
Criterion Point 3	-0.04	0.03
Criterion Point 4	-0.53	0.03

After the calculation of difficulty and standard errors, the cut-off score was estimated in the relevant data set. Khatimin et al. (2013) put forward that the examinees will be accepted as successful if they complete at least 60% of the task. The value of sixty-percent means the achievement of 9 criteria in ASRAE (15 x (60 / 100) = 9). In ASREA, when all the criteria are successively ordered in terms of the difficulty level, the logit value of the ninth criterion corresponds this value; in other words, the logit value .24 is accepted as the cut-off score (see Table 1).

When Table 3 is examined, it is seen that cut-off score (+0.24 logit) and criterion point 2 are in the same order. To find out the confidence interval of the cut-off score, standard error (0.03) was multiplied by ±1.96. It is seen that the confidence interval with 95% is between +0.18 and +0.30. This proves that calculated cut-off score is at confidence interval. Also, as given in Table 3, there is no more option for the cut-off score apart from +0.24 logit value because there is no logit value at confidence interval except for +0.24 logit. Table 3 provides information about the cut-off score, criterion points and the levels of the academic writing proficiency. Accordingly, out of 39 examinees, 15 examinees had high proficiency in L2 academic writing whereas 24 examinees had low proficiency in the same skill. In the final step, the examinees' proficiency levels were determined in consonance with the performance standards. Table 4 illustrates these levels and descriptive statistics. It can be seen that Level 5 and Level 2 have high frequencies. That is to say, the examinees can be holistically divided into two main groups in academic writing proficiency.

Table 3. *Estimation of the cut-off score and levels of academic writing proficiency.*

	Examinee	Observed	Fair-M	Logit	Standard	Infit	Infit	Outfit	Outfit	
		Average	Average	Measure	Error	MnSq	ZStd	MnSq	ZStd	
	S16	3.58	3.61	1.45	0.06	1.02	0.40	1.05	0.90	
	S21	3.47	3.51	1.15	0.05	1.19	3.50	1.26	4.60	
	S24	3.44	3.48	1.07	0.05	1.01	0.10	1.09	1.70	
	S28	3.39	3.43	0.95	0.05	0.88	-2.60	0.88	-2.50	
	S29	3.36	3.40	0.89	0.05	0.88	-2.50	0.87	-2.70	
	S15	3.34	3.38	0.85	0.05	0.90	-2.20	0.98	-0.40	
	S17	3.32	3.36	0.81	0.05	1.06	1.20	1.12	2.40	
	S37	3.29	3.33	0.74	0.05	1.01	0.20	1.04	0.90	
	S19	3.29	3.33	0.74	0.05	0.96	-0.80	1.01	0.10	
	S03	3.21	3.25	0.58	0.04	1.24	4.80	1.27	5.30	
	S34	3.19	3.23	0.56	0.04	0.94	-1.20	0.97	-0.50	
	S26	3.17	3.21	0.51	0.04	0.95	-1.00	0.97	-0.70	
	S01	3.11	3.15	0.41	0.04	0.99	-0.20	1.01	0.30	
CP-1	S13	3.11	3.15	0.40	0.04	0.86	-3.10	0.86	-3.20	
CP-2	S12	3.05	3.09	0.30	0.04	1.09	1.90	1.13	2.70	CS
	S31	2.88	2.92	0.03	0.04	0.68	-8.00	0.69	-7.70	
	S07	2.88	2.91	0.02	0.04	1.34	6.90	1.40	8.10	
	S25	2.87	2.90	0.00	0.04	0.76	-5.70	0.78	-5.40	
	S23	2.84	2.88	-0.03	0.04	0.74	-6.50	0.74	-6.30	
CP-3	S02	2.84	2.87	-0.04	0.04	0.93	-1.60	0.92	-1.90	
	S22	2.83	2.86	-0.05	0.04	0.84	-3.80	0.86	-3.20	
	S32	2.83	2.86	-0.05	0.04	0.97	-0.60	0.96	-0.80	
	S27	2.81	2.84	-0.08	0.04	0.72	-6.80	0.73	-6.80	
	S14	2.80	2.84	-0.09	0.04	1.01	0.20	1.04	1.00	
	S11	2.73	2.76	-0.21	0.04	0.99	-0.20	0.98	-0.40	
	S35	2.68	2.71	-0.27	0.04	1.10	2.20	1.10	2.30	
	S20	2.67	2.70	-0.29	0.04	0.89	-2.50	0.92	-1.90	
	S36	2.53	2.56	-0.48	0.04	0.75	-6.20	0.77	-5.60	
	S30	2.52	2.55	-0.49	0.04	0.93	-1.50	0.96	-1.00	
	S18	2.51	2.54	-0.50	0.04	0.89	-2.50	0.89	-2.50	
	S33	2.51	2.54	-0.51	0.04	0.97	-0.60	0.99	-0.30	
CP-4	S05	2.49	2.52	-0.53	0.04	1.08	1.70	1.13	2.90	
	S09	2.33	2.36	-0.74	0.04	0.95	-1.00	0.95	-1.00	
	S10	2.23	2.25	-0.87	0.04	1.47	9.00	1.50	9.00	
	S08	2.07	2.08	-1.07	0.04	0.78	-5.60	0.78	-5.40	
	S06	2.02	2.03	-1.14	0.04	1.01	0.10	1.01	0.20	
	S38	1.89	1.89	-1.30	0.04	1.38	8.10	1.39	8.40	
	S39	1.86	1.86	-1.34	0.04	1.37	7.90	1.38	8.10	
	S04	1.83	1.83	-1.37	0.04	1.16	3.70	1.17	3.80	

CP-1 (Criterion Point 1); CP-2 (Criterion Point 2); CP-3 (Criterion Point 3); CP-3 (Criterion Point 3), CS (Cut Score) Cut score and Criterion Point 2 are at the same line.

Table 4. *The levels of proficiency and descriptive statistics.*

Achievement Levels	Frequency	Percentage	Mean	Std. Deviation
Level 5 (0.36 - the highest logit)	14	35.90	0.79	0.30
Level 4 (0.27 and 0.35 logit)	1	2.56	0.30	--
Level 3 (-0.04 and 0.26 logit)	5	12.82	-0.01	0.03
Level 2 (-0.53 and -0.05 logit)	12	30.77	-0.30	0.20
Level 1 (-0.54 the lowest logit)	7	17.95	-1.12	0.24

3.2. Two-step Cluster Analysis Results

The mean of the scores that the raters assigned for each examinee was used in two-step cluster analysis. The analysis results highlight the existence of two clusters (the quality of clustering: 0.67, and Silhoutte coefficient: 0.58). With respect to the placement of the examinees to these clusters, it can be understood that two clusters show consistency with two groups divided by the cut-off score in OSS method. Put it another way, the first cluster includes the examinees with high proficiency in academic writing, and the second cluster includes the examinees with low proficiency. Therefore, two-step cluster analysis confirms and validates the findings in OSS method. Table 5 shows the comparative results of two-step cluster analysis and OSS Method.

Table 5. *Comparison of OSS method and two-step cluster analysis.*

Two-step Cluster Analysis			OSS Method		
Silhoutte Coefficient	Rank of Cluster Analysis	Cluster	Rank in OSS Method	Logit Value	Proficiency
0.723	S16	1	S16	1.45	High
0.788	S21	1	S21	1.15	High
0.807	S24	1	S24	1.07	High
0.833	S28	1	S28	0.95	High
0.841	S29	1	S29	0.89	High
0.845	S15	1	S15	0.85	High
0.846	S17	1	S17	0.81	High
0.843	S19	1	S37	0.74	High
0.824	S37	1	S19	0.74	High
0.814	S03	1	S03	0.58	High
0.802	S34	1	S34	0.56	High
0.776	S26	1	S26	0.51	High
0.698	S01	1	S01	0.41	High
0.696	S13	1	S13	0.40	High
0.566	S12	1	S12	0.30	High
0.070	S31	2	S31	0.03	Low
0.108	S07	2	S07	0.02	Low
0.147	S25	2	S25	0.00	Low
0.227	S23	2	S23	-0.03	Low
0.257	S02	2	S02	-0.04	Low
0.276	S22	2	S22	-0.05	Low
0.278	S32	2	S32	-0.05	Low
0.321	S27	2	S27	-0.08	Low
0.337	S14	2	S14	-0.09	Low
0.465	S11	2	S11	-0.21	Low

Table 5. *Continues*

0.528	S20	2	S35	-0.27	Low
0.562	S35	2	S20	-0.29	Low
0.623	S30	2	S36	-0.48	Low
0.628	S18	2	S30	-0.49	Low
0.629	S33	2	S18	-0.50	Low
0.631	S05	2	S33	-0.51	Low
0.631	S36	2	S05	-0.53	Low
0.624	S09	2	S09	-0.74	Low
0.612	S10	2	S10	-0.87	Low
0.586	S08	2	S08	-1.07	Low
0.575	S06	2	S06	-1.14	Low
0.545	S38	2	S38	-1.30	Low
0.538	S39	2	S39	-1.34	Low
0.529	S04	2	S04	-1.37	Low

4. DISCUSSION and CONCLUSION

“Standard setting involves judgments about the ideal performance standard and test score that reflect this standard” (Hsieh, 2013). It has important consequences for the stakeholders such as students, teachers, and policy-makers in different areas (Fulcher, 2013; Shin & Lidster, 2017; Sondergeld et al., 2020; Stone et al., 2011). However, standard setting methods may include subjective evaluation and judgments (Davis-Becker et al., 2011). Considering this perspective, the current research study employed OSS method in order to provide objective and valid cut-off scores and performance standards for the stakeholders. In this way, it was attempted to establish a basis for making credible and valid decisions on L2 academic writing.

OSS method is based on item response theory and Rasch model and analyzes the data at item/rater/difficulty level. Stone et al. (2011) put emphasis on three important points in OSS method: “defining criterion set”, “refining criterion point”, and “expressing error” (p. 950). This study adopted these perspectives and set performance standards in four steps. Firstly, the content of criterion set was defined in line with the review of literature and validated in light of the expert opinions. 7 main criteria and 16 sub-criteria were specified in accordance with the feedback received from the experts. In the second step, the performance levels were determined with reference to the CEFR, and criterion points were defined in consonance with these levels. In the following step, mean difficulty levels and standard errors were calculated. Then the cut-off score was estimated by considering the task achievement level accentuated in Khatimin et al. (2013). It was found that the cut-off score (+0.24 logit) and the criterion point 2 were at the same line. Besides that, the confidence interval at which the cut-off score could be placed was found. The cut-off score divided the examinees into two groups with high proficiency (n = 15) and low proficiency (n = 24) in L2 academic writing. Finally, L2 academic writing proficiency, levels of the examinees and basic descriptive statistics were presented. The validity of OSS method results, especially the cut-off score, was confirmed by two-step cluster analysis which is based on less subjectivity (Khalid, 2011). Two-step cluster analysis results pointed out the emergence of two clusters, and the same examinees at high proficiency level and low proficiency level were successively placed into these two clusters. Therefore, it can be concluded that OSS method facilitates the specification of objective performance standards and valid cut-off scores in L2 academic writing assessment. MacDougall and Stone (2015) and Stone et al. (2011) found that OSS method was effective in the construct development in the target area when compared to other standard setting methods. In this research, it can be seen that OSS method serves as an objective basis for setting performance standards in L2 academic writing, specifying the valid cut-off score, and making reliable and objective decisions about L2 writing performance.

With regard to limitations of the study, the results may appear to be lack of generalizability in L2 writing assessment context due to the fact that this research was carried out with a specific subject group. Further research may focus on larger samples including more raters in different educational settings. Thus, performance standards and cut-off scores may be validated in various contexts. Another limitation of the study is related to the rater training. The raters in this study did not have any professional training about how to rate language skills. So, this standard setting study may be replicable by proving the raters with sufficient training about how to rate academic writing performance. As further research studies, different standard setting methods may be compared in terms of the objectivity in L2 academic writing assessment. This comparison may give more concrete evidence on the utility of performance standards. This study also suggests the use of OSS method in standard setting studies for different language skills.

Acknowledgments

The preliminary findings of this study were presented at an international conference titled "6th International Congress on Measurement and Evaluation in Education and Psychology" in September, 2018. The extended abstract was published in the abstract booklet

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research and publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s). Ethics Committee Approval and its number should be given by stating the institution name which gave the ethical approval. Ethics Committee Number: Gazi University, 80287700-302.08.01-54466.

Authorship Contribution Statement

Fatima Nur Fisne: Introduction, Review of Literature, Methodology (Data Collection, Instrument Development), Discussion and Conclusion. **Mehmet Sata**: Methodology (Instrument Development, Data Collection, Data Analysis), Results. **Ismail Karakaya:** Supervision.

Orcid

Fatima Nur FISNE https://orcid.org/0000-0001-9224-2485
Mehmet SATA https://orcid.org/0000-0003-2683-4997
Ismail KARAKAYA https://orcid.org/0000-0003-4308-6919

REFERENCES

Bejar, I.I. (2008). Standard setting: What is it? Why is it important*? R&D Connections, 7*, 1-6.

Best, J.W., & Khan, J.V. (2006). *Research in Education (10th Edition)*. Pearson.

Bichi, A.A., Talib, R., Embong, R., Mohamed, H. B., Ismail, M. S., & Ibrahim, A. (2019). Rasch-based objective standard setting for university placement test. *Eurasian Journal of Educational Research*, *19*(84), 57-70. https://doi.org/10.14689/ejer.2019.84.3

Chen, W.H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265-289. https://doi.org/10.3102/10769986022003265

Cizek, G.J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement, 30*(2), 93-106. https://doi.org/10.1111/j.1745-3984.1993.tb01068.x

Cizek, G.J. (Ed.). (2012). An introduction to contemporary standard setting: concepts, characteristics, and concepts. *In Setting performance standards: Concepts, methods, and perspectives.* Mahwah, NJ: Lawrence Erlbaum Associates.

Council of Europe [CoE]. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge, England: Cambridge University Press.

Cokluk, O., Sekercioglu, G., & Buyukozturk, S. (2012). *Sosyal bilimler icin cok degiskenli istatistik: SPSS ve LISREL uygulamalari* (2nd edition) [Multivariate statistics for social sciences: SPSS and LISREL applications], Pegem Akademi.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Studies in language testing 7: Dictionary of language testing.* Cambridge University Press.

Davis-Becker, S.L., Buckendahl, C.W., & Gerrow, J. (2011). Evaluating the bookmark standard setting method: The impact of random item ordering. *International Journal of Testing, 11*(1), 24-37. https://doi.org/10.1080/15305058.2010.501536

Elder, C., Barkhuizen, G., Knoch, U., & Von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing, 24*(1), 37-64. https://doi.org/10.1177/0265532207071511

Erkus, A., Sunbul, O., Omur-Sunbul, S., Yormaz, S., & Asiret, S. (2017). *Psikolojide olcme ve olcek gelistirme-II* (1st edition) [Measurement in psychology and scale development-II], Pegem Akademi.

Fleckenstein, J., Keller, S., Krüger, M., Tannenbaum, R.J., & Köller, O. (2020). Linking TOEFL iBT® writing rubrics to CEFR levels: Cut scores and validity evidence from a standard setting study. *Assessing Writing, 43,* 1-15. https://doi.org/10.1016/j.asw.2019.100420

Fulcher, G. (2013). *Practical language testing*. Routledge. https://doi.org/10.4324/980203767399

Goodwin, S. (2016). A Many-Facet Rasch analysis comparing essay rater behavior on an academic English reading/writing test used for two purposes. *Assessing Writing, 30*, 21-31. https://doi.org/10.1016/j.asw.2016.07.004

Green, A. (2018). Linking tests of English for academic purposes to the CEFR: The score user's perspective. *Language Assessment Quarterly, 15*(1), 59-74. https://doi.org/10.1080/15434303.2017.1350685

Harsch, C., & Rupp, A.A. (2011). Designing and scaling level-specific writing tasks in alignment with the CEFR: A test-centered approach. *Language Assessment Quarterly, 8*(1), 1-33. https://doi.org/10.1080/15434303.2010.535575

Hsieh, M. (2013). An application of multifaceted Rasch measurement in the Yes/No Angoff standard setting procedure. *Language Testing, 30*(4), 491-512. https://doi.org/10.1177/0265532213476259

IELTS (The Internatinal English Language Testing System). https://www.ielts.org/

Kayri, M. (2007). Two-step clustering analysis in researches: A case study. *Eurasian Journal of Educational Research (EJER), 28*, 89-99.

Khalid, M. N. (2011). Cluster analysis-a standard setting technique in measurement and testing. *Journal of Applied Quantitative Methods, 6*(2), 46-58.

Khatimin, N., Aziz, A.A., Zaharim, A., & Yasin, S.H.M. (2013). Development of objective standard setting using Rasch measurement model in Malaysian institution of higher learning. *International Education Studies, 6(*6), 151-160. https://doi.org/10.5539/ies.v6n6p151

Lawshe, C.H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 563-575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Linacre, J.M. (2017). *A user's guide to FACETS: Rasch-model computer programs.* Chicago: MESA Press.

Livingston, S.A., & Zieky, M.J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests.* Educational Testing Service: New Jersey.

McDonald, R.P. (1999). *Test theory: A unified approach.* Mahwah, NJ: Erlbaum.

MacDougall, M., & Stone, G.E. (2015). Fortune-tellers or content specialists: Challenging the standard setting paradigm in medical education programmes. *Journal of Contemporary Medical Education, 3*(3), 135. https://doi.org/10.5455/jcme.20151019104847

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465-493. https://doi.org/10.1177/0265532208094273

Shin, S.Y., & Lidster, R. (2017). Evaluating different standard-setting methods in an ESL placement testing context. *Language Testing, 34*(3), 357-381. https://doi.org/10.1177/0265532216646605

Sireci, S.G., Robin, F., & Patelis, T. (1997). Using cluster analysis to facilitate standard setting. *Applied Measurement in Education, 12*(3), 301-325. https://doi.org/10.1207/S15324818AME1203_5

Sondergeld, T.A., Stone, G.E., & Kruse, L.M. (2020). Objective standard setting in educational assessment and decision making. *Educational Policy, 34*(5), 735-759. https://doi.org/10.1177/0895904818802115

Stone, G.E. (2001). Objective standard setting (or truth in advertising). *Journal of Applied Measurement, 2*(2), 187-201.

Stone, G.E., Koskey, K.L., & Sondergeld, T.A. (2011). Comparing construct definition in the Angoff and Objective Standard Setting models: Playing in a house of cards without a full deck. *Educational and Psychological Measurement, 71*(6), 942-962. https://doi.org/10.1177/0013164410394338

Sata, M. & Karakaya, I. (2021). Investigating the effect of rater training on differential rater function in assessing academic writing skills of higher education students. *Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 163-181. https://doi.org/10.21031/epod.842094

Tannenbaum, R.J., & Wylie, E.C. (2008). Linking English-language test scores onto the common European framework of reference: An application of standard-setting methodology. *ETS Research Report Series, 2008*(1), i-75. https://doi.org/10.1002/j.2333-8504.2008.tb02120.x

Trace, J., Janssen, G., & Meier, V. (2017). Measuring the impact of rater negotiation in writing performance assessment. *Language Testing, 34*(1), 3-22. https://doi.org/10.1177/0265532215594830

Violato, C., Marini, A., & Lee, C. (2003). A validity study of expert judgment procedures for setting cutoff scores on high-stakes credentialing examinations using cluster analysis. *Evaluation & The Health Professions, 26*(1), 59-72. https://doi.org/10.1177/0163278702250082

Weigle, S.C. (2002). *Assessing writing.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Wilson, F.R., Pan, W., & Schumsky, D.A. (2012). Recalculation of the critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development, 45*(3), 197-210. https://doi.org/10.1177/0748175612440286

Wind, S.A., & Engelhard Jr, G. (2013). How invariant and accurate are domain ratings in writing assessment? *Assessing Writing, 18(*4), 278-299.

Wright, B.D., & Grosse M. (1993). How to set standards. *Rasch Measurement Transactions, 7*(3), 315-316.

Yudkowsky, R., Downing, S. M., & Tekian, A. (2009). Standard setting. In R. Yudkowsky & S. Downing (Ed.), *Assessment in health professions education* (pp. 86-105). Routledge. https://doi.org/10.4324/9781315166902-6

APPENDIX

Appendix A

ANALYTIC WRITING SCORING RUBRIC FOR ACADEMIC ESSAYS

	ORGANIZATION						CONTENT	
Point	**Title of Essay**	**Introduction-Body-Conclusion**	**Thesis Statement**	**Topic Sentence**	**Supporting Sentences**	**Appropriate Length**	**Topic Relevance**	**Idea Development**
4	Title of essay *comprehensively* represents the focus of the written text. It is *highly* relevant to the task.	The organization of introduction, body, and conclusion paragraphs is *highly* appropriate to written genre.	Thesis statement is *noticeably* given in introduction paragraph. It *comprehensively* includes the specific idea(s) to be elaborated in the written text.	Topic sentence *comprehensively* addresses and supports the specific idea(s) given in thesis statement. It *extensively* demonstrates the main idea of the paragraph.	Supporting sentences *comprehensively* illustrate the main idea given in topic sentence.	There are *at least 250 words* in written text. It is constructed with *appropriate length*.	Written text is *highly* relevant to assigned topic in task. It *comprehensively* addresses all parts of the task.	*Extensive* details are provided to develop, support and illustrate information or ideas presented in written text.
3	Title of essay *adequately* represents the focus of the written text. It is relevant to the task.	The organization of introduction, body, and conclusion paragraphs is *largely* appropriate to written genre.	Thesis statement is *evidently* given in introduction paragraph. It *mostly* includes the specific idea(s) to be elaborated in the written text.	Topic sentence *mostly* addresses and supports the specific idea(s) given in thesis statement. It *largely* demonstrates the main idea of the paragraph.	Supporting sentences *adequately* illustrate the main idea given in topic sentence.	Text length is between *200 and 249 words*. It is *slightly* shorter than required length.	Written text is *mostly* relevant to assigned topic in task. It *adequately* addresses the basic parts of the task.	*Adequate* details are provided to develop, support and illustrate information or ideas presented in written text.

2	Title of essay *moderately* represents the focus of the written text. It is relevant to the task in *some respects.*	The organization of introduction, body, and conclusion paragraphs is *moderately* appropriate to written genre.	Thesis statement is *less explicitly* given in introduction paragraph. It *moderately* includes the specific idea(s) to be elaborated in the written text.	Topic sentence *moderately* addresses and supports the specific idea(s) given in thesis statement. It demonstrates the main idea of the paragraph in *some respects.*	Supporting sentences *moderately* illustrate the main idea given in topic sentence.	Text length is between 150 *and 199 words*. It is *seemingly* shorter than required length.	Written text is *moderately* relevant to assigned topic in task. It *partially* addresses the basic parts of task.	*Basic* details are provided to develop, support and illustrate information or ideas presented in written text.
1	Title of essay *slightly* represents the focus of the written text. It is *partially* relevant to the task.	There is *inadequate* organization of introduction, body, and conclusion paragraphs in the written text.	Thesis statement is *vaguely* given in introduction paragraph. It *slightly* includes the specific idea(s) to be elaborated in the written text.	Topic sentence *partially* addresses and supports the specific idea(s) given in thesis statement. It *slightly* demonstrates the main idea of the paragraph.	Supporting sentences *partially* illustrate the main idea given in topic sentence.	Text length is between 100 *and 149 words*. It is *considerably* shorter than required length.	Written text is *slightly* relevant to assigned topic in task. It lacks addressing the basic parts of the task.	*Some details are* provided but they are not enough to develop, support and illustrate information or ideas presented in written text.
0	Written text does not include a title or title of essay is *completely* irrelevant.	Written text lacks organization of introduction, body and conclusion paragraphs.	Thesis statement is not given in introduction paragraph or it does not include any specific idea(s) to be elaborated in the written text.	Topic sentence is not included in written text, or it does not address the thesis statement or demonstrate the main idea of the paragraph.	Written text does not include supporting sentences or they do not illustrate the main idea given in topic sentence.	Text length is *below 99 words.* It does not meet the requirement of appropriate length.	Written text is irrelevant to assigned topic in task. It fails to address the task adequately.	Information or ideas are not *thoroughly* developed, supported or illustrated in written text.

	COHERENCE	**COHESION**	**GRAMMAR**		**VOCABULARY**		**MECHANICS**	
Point	**Coherence**	**Linking**	**Accuracy of Grammatical Forms**	**Syntactic Complexity**	**Word Choice**	**Lexical Range**	**Spelling**	**Punctuation**
4	Information or ideas sequenced in paragraphs are *highly* consistent. There is a *considerably* logical progression between sentences in written text.	A *wide* range of cohesive devices used to connect ideas in written text provides a smooth transition between sentences.	All grammatical forms are *accurately* used in written text. The communication is *successfully* established.	Complex and sophisticated sentences are *extensively* used in written text in which syntactic structures are *highly* diverse.	All the words and phrases are *appropriately* used. The intended meaning is *clearly* conveyed in written text.	There is a *wide range* of vocabulary used in written text which includes *highly* sophisticated words and phrases.	All the needed spelling rules are *accurately* used in written text.	All the needed punctuation rules are *accurately* used in written text.
3	Information or ideas sequenced in paragraphs are *mostly* consistent. There is an *adequately* logical progression between sentences in written text.	An *adequate* range of cohesive devices used to connect ideas in written text provides an easy transition between sentences.	The use of the grammatical forms is *mostly accurate* in the written text. There are *few grammatical errors* which do not impede communication.	Complex and sophisticated sentences are *widely* used in written text in which syntactic structures are *adequately* diverse.	The use of words and phrases is *mostly appropriate*. There are *few* misused words or phrases which cannot obscure the intended meaning.	There is an *adequate range* of vocabulary used in written text which includes *largely* sophisticated words and phrases.	All the needed spelling rules are *mostly accurate* in written text but there are *few errors* which violate these rules.	All the needed punctuation rules are *mostly accurate* in written text but there are few errors which violate these rules.
2	Information or ideas sequenced in paragraphs are *moderately* consistent but there are some inconsistencies which *partially* interrupt logical progression between sentences.	The use of cohesive devices *at basic level* to connect ideas in written text provides a complete transition between sentences.	It is attempted to use the grammatical forms accurately in written text but there are *occasional grammatical errors* which slightly impede communication.	Complex and sophisticated sentences are *moderately* used in written text in which syntactic structures are *partially* diverse.	It is attempted to use the words and phrases appropriately but there are *occasionally* misused words or phrases which *slightly* obscure the intended meaning.	The *basic* vocabulary is used in written text which includes *moderately* sophisticated words and phrases.	It is intended to use the needed spelling rules *accurately* in written text but there are *occasional errors* which violate these rules.	It is intended to use the needed punctuation rules *accurately* in written text but there are *occasional errors* which violate these rules.

1	Paragraphs are constructed with *slightly* consistent information or ideas which interrupt logical progression and sequence between sentences.	A *limited* range of cohesive devices used to connect ideas in written text makes transition between sentences fragmentary.	The use of the grammatical forms is *generally inaccurate* in written text. There are *frequent grammatical errors* which largely impede communication.	Complex and sophisticated sentences are *slightly* used in written text in which syntactic structures are diverse to some extent.	The use of words and phrases is *generally inappropriate*. There are *frequently* misused words or phrases which *largely* obscure the intended meaning.	There is a *limited range* of vocabulary used in written text which includes *slightly* sophisticated words and phrases.	The use of the needed spelling rules is *largely* inaccurate. There are *frequent errors* which violate these rules.	The use of the needed punctuation rules is *largely* inaccurate. There are *frequent errors* which violate these rules.
0	Written text lacks consistency and logical progression between sentences.	There is an *inadequate* use of cohesive devices in written text which lacks transition between sentences.	The use of grammatical forms is *completely inaccurate* in the written text. This causes a breakdown in communication.	Written text lacks sentential complexity, sophistication and syntactic variety.	The use of vocabulary is completely inappropriate in written text. The intended message is obscured.	A repetitive vocabulary is largely used in written text which lacks sophistication.	All the needed spelling rules are *inaccurately* used in written text.	All the needed punctuation rules are *inaccurately* used in written text.

Appendix B

Fit values of the rubric criteria

Criteria	Logit value	Standard Error	Infit	ZStd	Outfit	ZStd	Correlation
Syntactic Complexity	0.43	0.02	0.72	-9.00	0.73	-9.00	0.40
Idea Development	0.35	0.02	0.71	-9.00	0.72	-9.00	0.48
Topic Sentence	0.34	0.02	1.16	5.60	1.13	4.70	0.47
Lexical Range	0.33	0.02	0.70	-9.00	0.71	-9.00	0.41
Thesis Statement	0.30	0.02	1.40	9.00	1.37	9.00	0.43
Supporting Sentence	0.28	0.02	0.80	-7.60	0.81	-7.30	0.48
Linking	0.24	0.02	0.86	-5.50	0.87	-4.90	0.45
Accuracy of Grammatical Forms	0.01	0.03	0.97	-1.00	1.00	-0.10	0.27
Coherence	-0.01	0.03	0.81	-7.30	0.83	-6.60	0.44
Introduction-Body-Conclusion	-0.07	0.03	1.38	9.00	1.26	8.40	0.51
Word Choice	-0.07	0.03	0.81	-7.00	0.84	-6.10	0.35
Topic Relevance	-0.34	0.03	1.10	3.40	1.12	4.00	0.39
Appropriate Length	-0.49	0.03	1.28	8.70	1.21	6.40	0.45
Punctuation	-0.59	0.03	1.18	5.70	1.24	7.40	0.25
Spelling	-0.70	0.03	1.30	9.00	1.38	9.00	0.17
Mean	0.00	0.03	1.01	-0.3	1.01	-0.2	0.40
S (Universe)	0.36	0.00	0.25	7.4	0.23	7.2	0.09
S (Sample)	0.37	0.00	0.26	7.7	0.24	7.4	0.10

Model, Universe: RMSE = 0.03 Adjusted S = 0.36 Separation Ratio = 14.18
Separation Index = 19.25 Reliability = 1.00
Model, Sample: RMSE = 0.03 Adjusted S = 0.37 Separation Ratio= 14.69
Separation Index = 19.91 Reliability = 1.00

Model, Chi-square (Fixed Effect) : 2818.10 *sd* = 14 *p* = .00

Model, Chi-square (Normal) : 13.90 *sd* = 13 *p* = .38

Biserial correlations(x) are between .17 and .51, outfit values (MNSQ) are between 0.71 and 1.38, and the standards of outfit values (ZSTD) are between -9.00 and 9.00 (fit values: $0.4<x<0.8$, $0.5<MNSQ<1.5$ and $-2.0<z<2.0$). As a consequence, it is understood that there is no misfit in the dataset except for the standards of the outfit values.

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 1, 98–117

https://doi.org/10.21449/ijate.799754

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

School characteristics mediating the relationship between school socioeconomic status and mathematics achievement

Ozlem Albayrakoglu[1,*], Selda Yildirim[2]

[1]Bolu Abant Izzet Baysal University, Bolu, Turkiye.
[2]Bolu Abant Izzet Baysal University, Faculty of Education, Department of Mathematics and Science Education, Bolu, Turkiye.

ARTICLE HISTORY

Received: Sep. 24, 2020
Revised: Nov. 04, 2021
Accepted: Dec. 04, 2021

Keywords:
Equity,
Multilevel Multiple Mediation,
Quality of Mathematics Teaching,
School Climate,
School SES,
TIMSS 2015.

Abstract: While numerous studies have reported the effect of school socioeconomic status (SES) on achievement, the factors that can cause this relationship are not well established. This study is, therefore, an attempt to understand school SES and students' mathematics achievement relationship by assuming that this relationship occurs through a correlation between school SES and school characteristics. Identifying these school characteristics is crucial to reduce the relation between SES and achievement for educational equity. Focusing on the 8th-grade mathematics data from Trends in International Mathematics and Science Study (TIMSS) 2015, this study aimed to identify school characteristics (quality of mathematics teaching at school, discipline at school, sense of school belonging, and school academic emphasis) that can mediate the relationship between school SES and students' mathematics achievement. The results of multilevel regression analyses showed that controlling school characteristics reduced the relationship between school SES and students' mathematics achievement in most of the educational systems. However, the results of multilevel multiple mediation analysis showed that the relationship between school SES and students' mathematics achievement were mediated through discipline at school, school academic emphasis, or sense of school belonging in some educational systems. In addition, the results indicated that the quality of mathematics teaching at school was not a mediator in this relationship. These results suggest the need for eliminating the effect of school SES on some school characteristics to improve equity in education.

1. INTRODUCTION

Equity in mathematics education means providing individualized support to students as much as they need (National Council of Teachers of Mathematics, 2000). This support addresses the barriers that students face to achieve their potential in mathematics. Socioeconomic status (SES) is one of these barriers because researchers point out that both student SES and school SES determine student achievement in many countries (Perry & McConney, 2010; Sirin, 2005). Besides, Chmielewski's study (2019) reveals that SES related performance differences have been increasing consistently over the past 50 years in many countries. Researchers also state

*CONTACT: Özlem Albayrakoğlu ✉ ozlemalbayrakoglu@gmail.com Bolu Abant Izzet Baysal University, Bolu, Turkiye

e-ISSN: 2148-7456 **/© IJATE** 2022

that the effect of school SES is more significant on students' mathematics achievement than the effect of student SES (Borman & Dowling, 2010; Sirin, 2005). Therefore, it seems crucial to eliminate the relationship between school SES and achievement for educational equity.

Student SES is defined as the social and economic background of students' parents. Accordingly, school SES is defined as the socioeconomic status of parents of students in a school. Student SES can be indicated by various ways such as parents' income, parents' education level, or home educational resources. Students with higher SES are likely to have an education in academically more advantaged schools (having qualified teaching and school resources). Likewise, students with lower SES probably will be educated in academically more disadvantaged schools (Berkowitz et al., 2017; Liu et al., 2015; Rumberger & Palardy, 2005; Sirin, 2005). The mentioned distribution of students to schools in a planned or an unplanned manner causes differences in achievement among schools as well as the individual performance of students. For example, a student attending a school with a higher SES is more likely to be successful compared to a student having a similar family structure but attending a lower SES school (Gustafsson et al., 2016; Liu et al., 2015; Mullis et al., 2016).

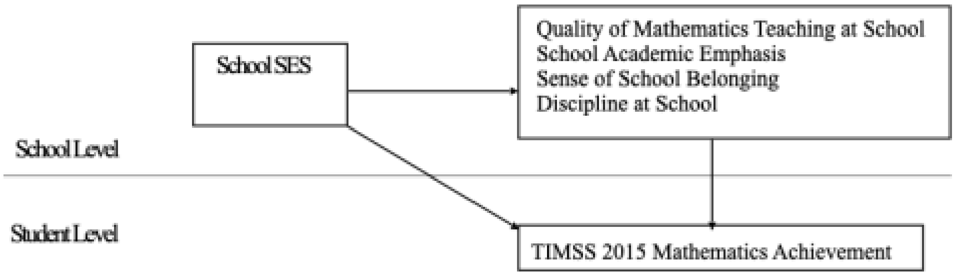
Besides, research shows that students who attend schools having qualified teaching, safe and supportive school climate, a high sense of school belonging, and giving high academic emphasis will have a higher probability to be successful in mathematics (Gustafsson et al., 2016; Nilsen et al., 2016; Olmez, 2020; Thrupp et al., 2002). There is little research and there are no comprehensive theories that address the relationships between school SES, school characteristics, and academic performance. However, about the theoretical nature of these relationships, some authors argue that a school's SES affects its characteristics (e.g., social climate), which in turn affects students' academic achievement (Berkowitz et al., 2017). That is, students' mathematics achievement is assumed to be influenced by school SES indirectly through school characteristics. This indirect relationship is named mediation and, in this mediation, the school characteristics are called mediators. Consistent with this argument the results of the previous studies outlined below suggest that school characteristics may have a potential mediational role between school SES and student mathematics achievement. For example, school SES may be related to mathematics achievement indirectly through the mediating role of school characteristics such as quality of teaching and school climate (Berkowitz et al., 2015; Gustafsson et al., 2016; Hansen & Strietholt, 2018; Liu et al., 2015; Nilsen et al., 2016; Schmidt et al., 2015). The educational practices that contribute to student achievement in a school are named as "quality of teaching at school" (Brophy & Good, 1986). The quality of mathematics teaching might be better in high SES schools (i.e., schools with high SES students) because these schools are more likely to have well-qualified mathematics teachers. Since the quality of mathematics teaching at school is associated with students' mathematics achievement, school SES may indirectly be related to mathematics achievement. On the other hand, low SES schools may not provide a safe and supportive learning environment because of disadvantages resulting from having low SES such as social problems, violence, and emotional and behavioral difficulties (Berkowitz et al., 2017; Liu et al., 2015). Without a safe and supportive learning environment in school, learning cannot become students' focus. The degree of physical and emotional safety and order of disciplinary situations provided for students by a school are conceptualized as "discipline at school" (Wang & Degol, 2016). Therefore, school SES has the potential to affect students' mathematics achievement via discipline at school (Liu et al., 2015). Similarly, research has shown that students' sense of belonging to school (Munk, 2007) may be another meaningful mediator between school SES and mathematics achievement. "A sense of school belonging" is conceptualized as that a student has the feeling of being an essential part of school/classroom life and activities and is accepted and valued by teachers and peers (Goodenow, 1993). Also, academic emphasis in school is another school characteristics that may be related to SES (Wu et al., 2013) and achievement

(Hoy, 2012; Olmez, 2020; Yavuz et al., 2017). "The school academic emphasis" is conceptualized as priority and importance given on learning and achievement by a school (Hoy et al., 2006). The work of Boonen et al. (2014) demonstrates that the school academic emphasis may have the potential to mediate the relationship between SES and mathematics achievement.

Despite numerous studies reported the relationship between school SES and student achievement, few studies investigated the underlying mechanism of this relationship (Berkowitz et al., 2017; Liu et al., 2015; Schmidt et al., 2015). In these studies, the joint effects of multiple mediators were not considered. In addition, examining these joint mediation effects in different cultures may shed light on understanding possible different school SES mechanism that is related to cultural difference. Viewing the gaps in previous studies, we investigated the multiple school characteristics mediating the relation between school SES and students' mathematics achievement. The existence of school-SES indirect effects in addition to its direct effect (the effect with no mediator) on student mathematics achievement creates a situation causing inequity in mathematics achievement (Berkowitz et al., 2017; Gustafsson et al., 2016; Nilsen et al., 2016). In such a situation, managing the educational environment at school in the way that this mediating role is eliminated can contribute to establishing equity in achievement. For example, if low SES schools were able to alter their climate to enhance student achievement, this might eliminate the negative effects of the low SES.

Thus, the purpose of this study is to investigate the mediational role of school characteristics between school SES and student mathematics achievement in order to understand the school SES mechanism better. Previous studies indicate weaker school SES effects in secondary schools than primary schools (e.g., Driessen, 2002). This finding may imply that the school SES effect diminishes over time. However, most of the recent studies have focused on the secondary level and found that school SES is a stronger predictor of student outcomes in secondary schools (e.g., Liu et al., 2015). To contribute to the generalizability of the recent findings at the secondary level, in this study, we investigated the school SES mechanism using Trends in International Mathematics and Science Study (TIMSS) 2015 eighth-grade data. Previus studies that use TIMSS data report the effect of school SES on student mathematics achievement (e.g., Akyuz, 2014). Similarly, TIMSS reports that if low SES students attend schools which are composed of students from affluent homes, they are more likely to achieve a higher level compared to low SES students who attend schools composed of students from less affluent or disadvantaged homes (Mullis et al., 2011; Mullis et al., 2016). Accordingly, we formulated the following hypotheses: 1) School SES is positively related to students' mathematics achievement and 2) School SES is positively and indirectly related to students' mathematics achievement through the mediation of school characteristics (quality of mathematics teaching at school, discipline at school, school academic emphasis, and sense of school belonging). The proposed model demonstrates these relations (Figure 1).

Figure 1. *The Proposed Model.*


School SES
Quality of Mathematics Teaching at School
School Academic Emphasis
Sense of School Belonging
Discipline at School
School Level
Student Level
TIMSS 2015 Mathematics Achievement

2. METHOD

2.1. Data and Participants

This study uses students' and school principals' data obtained through TIMSS 2015 administered by the International Association for the Evaluation of Educational Achievement (IEA). The sample includes 37 countries after removing Saudi Arabia due to single-sex schools in order to control student gender in the analysis. Therefore, data of 248.667 8th-grade students and 7135 schools in all remaining countries are used in the study (see Appendix 1). TIMSS uses a stratified two-cluster sampling design in each country with schools and classes randomly selected (Martin et al., 2016).

2.2. Variables

2.2.1. *TIMSS 2015 mathematics achievement*

This study used five plausible values of mathematics achievement scores in TIMSS 2015 data as the dependent variable. Plausible values are multiple imputed scores based on estimates regarding student ability distribution (Mullis et al., 2016; Wu, 2004). The scales created by TIMSS from student and school principal's responses were used in this study and explained below (Martin et al., 2016).

2.2.2. *School SES*

TIMSS surveys school principals' views to determine student percentage of economically advantaged and disadvantaged backgrounds in their schools. TIMSS uses these responses to describe schools as "affluent," "neither affluent nor disadvantaged," and "disadvantaged" based on students' economic backgrounds.

2.2.3. *School characteristics*

2.2.3.1. School Academic Emphasis. TIMSS asks school principals to indicate how they characterize their schools' emphasis on academic achievement by rating some items from very low to very high scale. Examples of these items are "parental expectations for student achievement," "teachers' degree of achievement in implementing the school's curriculum," "teachers' expectation for student achievement," "teachers' collaborative work to improve student achievement," "students' desire to do well in school", and "students' ability to reach school's academic goals." Then TIMSS combines these responses to describe school's academic emphasis as "medium", "high", and "very high".

2.2.3.2. Discipline at School. Similarly, an index score in the TIMSS 2015 database, based on principals' responses to the question "To what degree each of the following is a problem among 8th-grade students in your school?" was used in this study. Some of the issues scored by principals included "arriving late at school," "absenteeism," "intimidation and verbal abuse among students," "profanity," "cheating," thieving", and "vandalism." TIMSS uses these responses to describe schools' discipline problems as "hardly any", "minor", and "moderate to severe".

2.2.3.3. Sense of School Belonging. The extent of student sense of school belonging is categorized as "high sense of school belonging," "sense of school belonging", and "little sense of school belonging" by an index in the TIMSS 2015 database. It is derived from students' responses to the question "What do you think about your school?" Some of the item examples were "I like being in school," "I feel safe when I am at school," "I feel like I belong at this school," and "I like to see my classmates at school". This score was aggregated from student-level to school level data.

2.2.3.4. Quality of Mathematics Teaching at School. Relevant literature indicates that it is beneficial to consider the relationship between achievement and variables built up from students’ responses on topics of teaching practices at class, quality of teaching, and so on to obtain reasonable results (Eriksson et al., 2019). The study used an index score from the TIMSS database aggregated from the student level to the school level regarding the quality of mathematics teaching (Gustafsson et al., 2016). The index obtained from student responses to items such as: “I know what my teacher expects me to do,” “My teacher is easy to understand,” “My teacher gives me interesting things to do,” “My teacher has clear answers to my questions,” and “My teacher is good at explaining mathematics”. The index scored through “very engaging teaching,” “engaging teaching”, and “less than engaging teaching”.

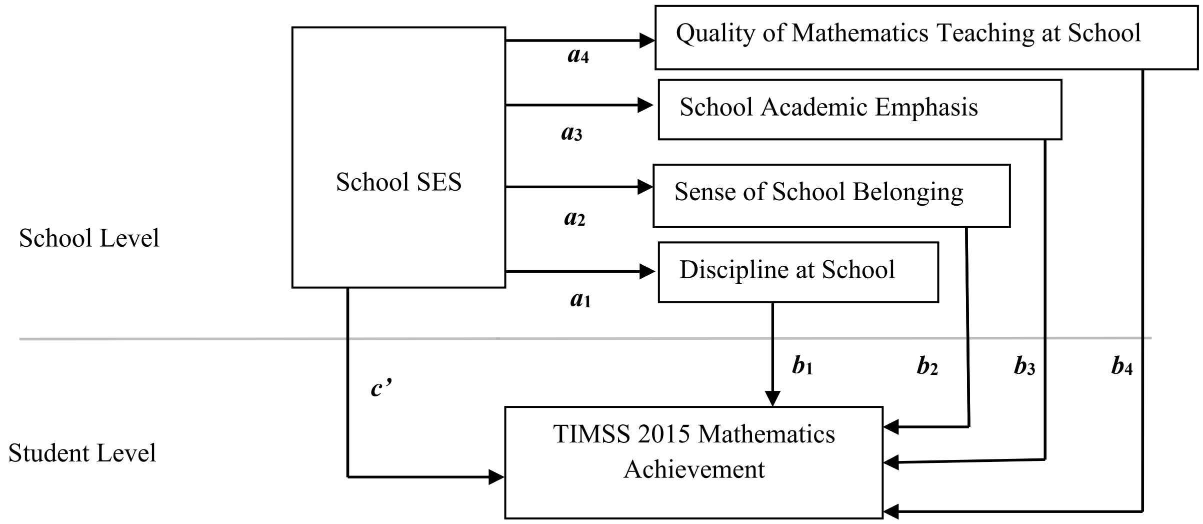
2.2.3.5. Controlling Variables. The variables “sense of school belonging” and “quality of mathematics teaching at school” are used as student-level control variables as suggested since they were aggregated from student-level data (Armor et al., 2017). In addition, previous studies reveal that student gender (Contini et al., 2017), self-confidence (Ker, 2016; Wang et al., 2012), and student SES (Sirin, 2005) may be related to mathematics achievement. Therefore, these variables were also used as control variables. Student gender variable was coded 1 for girls and 0 for boys. For student SES, this study used the “home educational resources” scale. At this scale, students with “many resources” have 100 or more books, private rooms, and the Internet at home and at least one parent holds a Bachelor’s degree. Students with “few resources,” on the contrary, are described as having 25 or fewer books, no Internet connection, and no private room at home, and none of the parents holds a higher degree than a secondary school degree. Other students were assigned to the “some resources” category. For self-confidence, the scale constructed from student responses to items such as, “I usually do well in mathematics” and “Mathematics is more difficult for me than for many of my classmates" was used. On this scale, students were categorized as “very confident", “confident”, and “not confident”.

2.3. Statistical Analyses

In mediation models, one or more variables transmit the effect of an independent variable (X) on an independent variable (Y). These variables are named mediator (M). In this study, there are four mediators (quality of mathematics teaching at school, school academic emphasis, sense of school belonging, and discipline at school) and these mediators transmit the effect of school SES (X) on student mathematics achievement (Y). Figure 2 visualizes the effects of these mediators.

As seen in Figure 2, the independent variable (school SES) and mediating variables (quality of mathematics teaching at school and school academic emphasis, sense of school belonging, and discipline at school) are at the school level (level 2), and the dependent variable (student mathematics achievement) is at the student level (level 1). In this multilevel structure, mediation is referred to as the 2-2-1 mediation model (Bauer et al., 2006; Krull & MacKinnon, 2001). That is, in this mediation model, the independent variable (X) and mediators (M) are at level 2, and the dependent variable is at level 1. This mediation was tested by a multilevel mediation approach as indicated by Zhang et al. (2009). The following steps explain the multilevel mediation used in this study.

Figure 2. *Multilevel multiple mediation (indirect) effect. Student level control variables are not shown in the figure for clarity of the model.*


Quality of Mathematics Teaching at School
School Academic Emphasis
Sense of School Belonging
Discipline at School
School SES
a4
a3
a2
a1
School Level
Student Level
c'
b1
b2
b3
b4
TIMSS 2015 Mathematics Achievement

Step 1. Multiple regression analysis was performed to calculate the effect of school SES (X) on student mathematics achievement (Y), and this effect was indicated by ***c***.

Step 2. In this step, both school SES (X) and mediators (M) were included as simultaneous predictors of student achievement (Y) in multilevel regression analysis. The effects of mediators were indicated by $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, $\boldsymbol{b}_3$, and $\boldsymbol{b}_4$. Similarly, the effect of school SES was indicated by ***c′***.

Step 3. Standard regression analyses were performed to calculate the effects of school SES (X) on mediators (M), and these effects were indicated by $\boldsymbol{a}_1$, $\boldsymbol{a}_2$, $\boldsymbol{a}_3$, and $\boldsymbol{a}_4$.

As seen, ***c*** coefficient denotes the effect of school SES on student achievement, whereas the ***c'*** signifies the effect of school SES on student achievement in the presence of mediators. In mediation analysis, ***c-c′*** estimates the total mediation effect and ***a.b*** estimates the mediation effect for a single mediator. In our analysis, $\boldsymbol{a}_1.\boldsymbol{b}_1$, $\boldsymbol{a}_2.\boldsymbol{b}_2$, $\boldsymbol{a}_3.\boldsymbol{b}_3$, and $\boldsymbol{a}_4.\boldsymbol{b}_4$ estimate the mediation effects for four mediators.

International Data Base Analysis Program (IDB analyzer) [IEA, 2017] was used to compute the effects indicated in Step 3. In Step 1 and Step 2, the effects were estimated using HLM6 (Raudenbush & Bryk, 2002). In HLM analyses, student-level variables were group-mean centered, while school-level variables were grand-mean centered, and the random intercept fixed slope method was used. In the analyses, gender, self-confidence, student SES, quality of mathematics teaching, and sense of school belonging were considered as control variables at the student level. Sampling weights in the TIMSS database were used both at school and student levels. Before the analyses, missing data were imputed with the SPSS expectation-maximization algorithm and the variables were standardized. In addition, before the mediation analysis, IDB Analyzer (IEA, 2017) was used to perform the descriptive statistics and correlations between the variables. Finally, to determine whether the mediation effects are statistically significant, the code calculating the Monte Carlo Confidence Intervals prepared by Preacher and Selig (2012) was used in R (R Development Core Team, 2017).

3. FINDINGS

Appendix 2 presents descriptive statistics (i.e., means, standard deviations, and associated standard errors) for the variables included in the analysis. We first examined the effects of school SES and school characteristics on student mathematics achievement. Then, we displayed the relationships between school SES and school characteristics. Finally, we explained the mediating effects arising from these relationships.

3.1. Predicting Student Mathematics Achievement

As seen in Table 1, school SES effect on student mathematics achievement is positive and statistically significant in practically most of the educational systems, except for a few (China, Taiwan, Italy, Kazakhstan, Lebanon, Egypt, Russia, Thailand, and Oman). This effect is signified as the ***c*** coefficient in Table 1. However, in most of the educational systems, the initially positive association of school SES with mathematics achievement is reduced once school characteristics are accounted for. Also, the statistically significant positive school SES effect on student mathematics achievement disappears in Georgia, Kuwait and Norway. In Table 1, coefficient ***c'*** signifies the school SES effect on student mathematics achievement in the presence of the school characteristics. Concerning the proposed model, the sense of school belonging and school academic emphasis appear to be strong predictors of student mathematics achievement across some of the educational systems because of their effects (i.e., ***b*** coefficients in Table 1) considerably high and significant. However, discipline at school or the quality of mathematics teaching at school appears to be important factors for student mathematics achievement only in a few educational systems (e.g., Hungary and Kazakhstan). The last column of Table 1 shows the explained variance at school level when school SES and school characteristics simultaneously predicted student mathematics achievement in the multilevel regression analyses. On average, this value is 31% across educational systems.

Table 1. *The effect of school SES and school characteristics on mathematics achievement.*

	coef. *c*		coef. *c′*	coef. *b*				
	Sc SES	R^2(%)	Sc SES	DS (b_1)	SSB (b_2)	SAE (b_3)	QMT (b_4)	R^2(%)
	β SE		β SE	β SE	β SE	β SE	β SE	
Australia	0.33 (0.03)***	32	0.17 (0.04)***	0.00 (0.03)	0.28 (0.04)***	0.05 (0.04)	-0.02 (0.03)	57
Bahrein	0.16 (0.07)*	6	0.13 (0.06)*	0.12 (0.04)**	0.21 (0.05)***	0.19 (0.05)***	-0.05 (0.06)	33
Botswana	0.30 (0.04)***	32	0.20 (0.04)***	0.03 (0.03)	0.05 (0.04)	0.26 (0.04)***	-0.02 (0.03)	54
Canada	0.16 (0.03)***	8	0.09 (0.03)**	0.01 (0.04)	0.17 (0.04)***	0.15 (0.04)***	0.00 (0.04)	25
Chile	0.43 (0.05)***	31	0.39 (0.05)***	0.09 (0.04)*	0.11(0.04)*	0.11 (0.07)	-0.06 (0.04)	40
China-Taiwan	0.20 (0.11)	-	0.14 (0.07)	0.11 (0.05)*	-0.04 (0.06)	0.24 (0.06)***	0.16 (0.06)*	42
Egypt	0.15 (0.08)	-	0.16 (0.07)	-0.01 (0.07)	0.02 (0.08)	0.01 (0.06)	0.05 (0.09)	-
England	0.40 (0.07)***	20	0.18 (0.08)*	-0.01 (0.06)	0.49 (0.07)***	0.07 (0.07)	-0.07 (0.07)	46
Georgia	0.12 (0.06)*	3	0.10 (0.07)	-0.02 (0.07)	0.13 (0.07)	-0.01 (0.07)	0.01 (0.07)	6
Hong Kong	0.34 (0.06)***	19	0.12 (0.05)*	0.02 (0.05)*	0.48 (0.05)***	0.08 (0.05)	-0.12 (0.05)*	52
Hungary	0.41 (0.06)***	32	0.31 (0.05)***	0.18 (0.06)**	0.10 (0.07)	0.08 (0.06)	-0.04 (0.08)	43
Iran	0.25 (0.05)***	16	0.19 (0.05)**	0.05 (0.04)	-0.14 (0.05)**	0.16 (0.06)*	0.03 (0.05)	25
Ireland	0.28 (0.06)***	21	0.15 (0.04)**	0.08 (0.05)	0.25 (0.07)***	0.03 (0.04)	-0.05 (0.04)	43
Israel	0.36 (0.05)***	28	0.32 (0.05)***	0.06 (0.05)	0.06 (0.05)	0.04 (0.06)	-0.06 (0.06)	23
Italy	0.08 (0.05)	-	0.06 (0.05)	0.03 (0.05)	0.02 (0.04)	0.02 (0.05)	-0.05 (0.05)	2
Japan	0.13 (0.04)***	11	0.11 (0.03)***	0.03 (0.04)	-0.05 (0.04)	0.13 (0.03)***	0.06 (0.03)*	24
Jordan	0.16 (0.05)**	10	0.12 (0.05)*	0.00 (0.04)	0.04 (0.04)	0.22 (0.04)***	0.01 (0.05)	27
Kazakhstan	0.13 (0.07)	-	0.11 (0.07)	0.07 (0.07)	-0.08 (0.09)	0.04 (0.05)	0.25 (0.09)**	11
Korea	0.25 (0.04)***	35	0.19 (0.03)***	-0.09 (0.03)**	0.11 (0.04)**	0.12 (0.03)***	-0.09 (0.03)**	51
Kuwait	0.020 (0.09)*	9	0.14 (0.08)	0.07 (0.08)	0.21 (0.08)*	0.11 (0.08)	-0.02 (0.09)	21
Lebanon	0.09 (0.07)	-	0.07 (0.06)	0.05 (0.05)	-0.10 (0.05)	0.22(0.05)***	0.10(0.06)	18

Table 1. *Continues*

Lithuania	0.21 (0.04)***	16	0.20 (0.04)***	0.09 (0.04)	-0.10 (0.05)*	0.00(0.04)	0.02(0.05)	20
Malaysia	0.29 (0.05)***	18	0.23 (0.05)***	0.07 (0.05)	-0.02 (0.07)	0.17 (0.05)**	0.14 (0.07)*	29
Malta	0.21 (0.06)**	17	0.13 (0.05)*	0.07 (0.06)	0.20 (0.04)***	0.14 (0.04)**	0.08 (0.04)*	62
Morocco	0.31 (0.03)***	37	0.22 (0.04)***	-0.01 (0.04)	-0.09 (0.04)*	0.10 (0.06)	0.08 (0.04)*	42
New Zealand	0.35 (0.04)***	44	0.28 (0.03)***	0.02 (0.04)	0.19 (0.07)**	0.07 (0.04)	0.00 (0.04)	57
Norway	0.11 (0.03)**	10	0.05 (0.03)	-0.02 (0.03)	0.11 (0.04)**	0.12 (0.03)***	-0.02 (0.04)	29
Oman	0.08 (0.05)	-	0.0 8(0.04)	0.06 (0.04)	0.07 (0.06)	0.00 (0.04)	0.09 (0.06)	-
Qatar	0.09 (0.04)*	1	0.16 (0.04)**	0.08 (0.04)	0.29 (0.07)***	0.05 (0.05)	0.10 (0.06)	28
Russia	0.02 (0.06)	-	0.02 (0.06)	0.10 (0.06)	0.06 (0.06)	0.17 (0.06)**	0.10 (0.07)	10
Singapore	0.42 (0.05)***	31	0.30 (0.06)***	0.04 (0.05)	0.24 (0.06)***	0.10 (0.05)*	0.05 (0.04)	47
Slovenia	0.08 (0.03)*	1	0.08 (0.03)*	0.02 (0.03)	0.02 (0.04)	0.12 (0.04)**	0.08 (0.05)	17
South Africa	0.35 (0.10)**	22	0.32 (0.10)**	0.19 (0.06)**	-0.01 (0.08)	0.02 (0.08)	-0.03 (0.07)	28
Sweden	0.28 (0.04)***	28	0.23 (0.06)***	0.01 (0.05)	0.10 (0.05)*	0.05 (0.06)	0.05 (0.04)	36
Thailand	0.14 (0.08)	-	0.13 (0.09)	0.04 (0.05)	0.16 (0.07)*	0.07 (0.06)	-0.13 (0.08)	14
Turkey	0.31 (0.08)***	23	0.21 (0.07)**	0.04 (0.05)	-0.03 (0.05)	0.21 (0.07)**	0.13 (0.05)*	40
UAE	0.20 (0.03)***	8	0.07 (0.03)*	0.05 (0.03)	0.21 (0.04)***	0.19 (0.03)***	0.07 (0.04)*	35
USA	0.32 (0.03)***	34	0.20 (0.03)***	-0.06 (0.05)	0.28 (0.04)***	0.09 (0.04)*	-0.02 (0.04)	46

$^{***}p<0.001$; $^{**}p<0.01$; $^{*}p<0.05$; Sc SES: School SES; SSB: Sense of School Belonging; DS: Discipline at School; QMT: Quality of Mathematics Teaching at School; SAE: School Academic Emphasis; R^2: The variance explained at school level.

3.2. Predicting School Characteristics

Table 2 displays the school SES effect on school characteristics (i.e., coefficient ***a***). Evaluating interactions of school SES with school characteristics revealed that the school SES affects the school academic emphasis almost in all educational systems, confirming that the higher the school SES is, the higher the school academic emphasis is. However, a statistically significant relationship cannot be observable for Bahrein, Israel, Qatar, and Slovenia. In ten educational systems, there is a statistically significant positive relationship between school SES and the sense of school belonging, but a negative relation between school SES and sense of school belonging appears in Botswana, Morocco, South Africa, Qatar and Turkey. In these educational systems the higher the school SES is, the lower the sense of school belonging is. In twenty educational systems, a statistically significant and positive relation between school SES and discipline at school exists. Although the school SES interaction with other school characteristics is likely in the majority of educational systems, school SES can affect the quality of mathematics teaching at school positively only in Israel and the UAE, and negatively in South Africa, Qatar, and Thailand.

Table 2. *School SES effect on school characteristics.*

	coef. *a*			
	DS (a_1)	SSB (a_2)	SAE (a_3)	QMT (a_4)
	β SE	β SE	β SE	β SE
Australia	0.40 (0.06) ***	0.40 (0.09) ***	0.51 (0.06) ***	0.16 (0.07)
Bahrein	0.09 (0.10)	0.00 (0.08)	0.13 (0.08)	0.14 (0.10)
Botswana	0.31 (0.07) ***	-0.16 (0.08) *	0.38 (0.07) ***	0.08 (0.08)
Canada	0.19 (0.09) *	0.00 (0.08)	0.45 (0.06) ***	-0.13 (0.06)
Chile	0.11 (0.08) ***	-0.08 (0.08)	0.26 (0.08) ***	-0.04 (0.08)
China-Taiwan	0.12 (0.15)	-0.05 (0.18)	0.40 (0.11) ***	-0.21 (0.13)
Egypt	-0.01 (0.10)	-0.20 (0.08)	0.46 (0.08) ***	-0.14 (0.09)
England	0.20 (0.09) *	0.39 (0.06) ***	0.54 (0.07) ***	0.21 (0.09)
Georgia	-0.13 (0.13)	0.05 (0.14)	0.31 (0.07) ***	0.03 (0.11)
Hong Kong	0.23 (0.07) ***	0.36 (0.09) ***	0.56 (0.06) ***	0.02 (0.10)
Hungary	0.25(0.09) ***	0.11 (0.10) *	0.50 (0.08) ***	-0.10 (0.10)
Iran	-0.10 (0.09)	-0.19 (0.09)	0.23 (0.09) ***	-0.07 (0.07)
Ireland	0.26 (0.09) *	0.41 (0.09) ***	0.43 (0.07) ***	0.17 (0.09)
Israel	0.34 (0.07) ***	0.02 (0.08)	0.42 (0.06) ***	-0.03 (0.08)
Italy	0.12 (0.14)	0.11 (0.09)	0.21 (0.07) **	-0.19 (0.09)
Japan	0.15 (0.12) *	0.24 (0.17) *	0.25 (0.13) **	0.02 (0.21)
Jordan	-0.02 (0.10)	-0.01 (0.10)	0.21 (0.09) ***	-0.03 (0.08)
Kazakhstan	-0.14 (0.07)	0.04 (0.10)	0.07 (0.08) *	0.15 (0.09)
Korea	-0.11 (0.09)	0.02 (0.17)	0.27 (0.12) **	-0.22 (0.17)
Kuwait	0.25 (0.10) *	0.03 (0.11)	0.38 (0.11) **	0.00 (0.12)
Lebanon	-0.05 (0.14)	0.01 (0.08)	0.14 (0.12) **	-0.02 (0.09)
Lithuania	0.06 (0.10)	-0.07 (0.09)	0.06 (0.10) *	-0.14 (0.10)
Malaysia	0.12 (0.09) **	-0.03 (0.07)	0.34 (0.08) ***	-0.01 (0.07)
Malta	0.18 (0.20)	0.11 (0.14)	0.22 (0.15)	0.25 (0.15)
Morocco	0.20 (0.08) *	-0.18 (0.12) *	0.64 (0.07) *	0.09 (0.07)
New Zealand	0.37 (0.06) ***	0.17 (0.11)	0.51 (0.07) ***	-0.04 (0.09)
Norway	0.10 (0.10)	0.37 (0.11) ***	0.27 (0.12) *	0.25 (0.14)
Oman	0.06 (0.07)	-0.01 (0.06)	0.09 (0.07) *	0.04 (0.07)
Qatar	0.08 (0.01)	-0.21 (0.09) *	0.13 (0.09)	-0.20 (0.09) *
Russia	-0.03 (0.10)	-0.14 (0.10)	0.20 (0.08) **	-0.15 (0.08)
Singapore	0.16 (0.08) *	0.30 (0.08) ***	0.42 (0.06) ***	0.02 (0.08)
Slovenia	0.07 (0.08)	-0.04 (0.10)	0.05 (0.11)	-0.11 (0.07)
South Africa	0.09 (0.14) ***	-0.20 (0.07) **	0.19 (0.12) ***	-0.08 (0.10) **
Sweden	0.39 (0.08) ***	0.23 (0.09)	0.45 (0.07) ***	-0.11 (0.08)
Thailand	0.06 (0.10) **	-0.14 (0.11)	0.08 (0.09) ***	-0.23 (0.08) ***
Turkey	0.11 (0.10)	-0.29 (0.09) ***	0.37 (0.11) ***	0.00 (0.08)
UAE	0.12 (0.05) **	0.34 (0.04) ***	0.23 (0.05) ***	0.11 (0.05) *
USA	0.45 (0.09) ***	0.49 (0.08) ***	0.51 (0.06) ***	0.23 (0.09)

$^{***}p<0.001$; $^{**}p<0.01$; $^{*}p<0.05$. SSB: Sense of School Belonging; DS: Discipline at School; QMT: Quality of Mathematics Teaching at School; SAE: School Academic Emphasis.

3.3. Test of Multiple Mediation

Table 3 depicts statistically significant values and confidence intervals regarding the mediating role of school characteristics between the school SES and student mathematics achievement. As seen in Table 3, the mediating effect of sense of school belonging between school SES and student mathematics achievement in seven educational systems (Australia, England, Hong Kong, Norway, Singapore, The UAE and the USA) is statistically significant and positive. However, in Qatar, the statistically significant mediating role of sense of school belonging is negative. In nine educational systems (Botswana, Canada, Iran, Jordan, Korea, Malaysia, Norway, Turkey, and the USA) the school academic emphasis between school SES and mathematics achievement has a statistically significant and positive mediating role. The mediating effect of discipline at school between the relation on school SES and mathematics achievement is, only in Hungary, found statistically significant and positive. In none of the educational systems, the quality of mathematics teaching at school has no mediating role between the school SES and student mathematics achievement.

Table 3. *The mediating role of school characteristics.*

	Mediating Effects					
	DS $(a_1.b_1)$	SSB $(a_2.b_2)$	SAE $(a_3.b_3)$	QMT $(a_4.b_4)$	Total $(c\text{-}c')$	Confidence Interval $(c\text{-}c')$
Australia		0.11^{**}			0.11^{**}	(0.044-0.200)
Botswana			0.10^{**}		0.10^{**}	(0.045-0.167)
Canada			0.07^{*}		0.07^{**}	(0.020-0.125)
England		0.20^{**}			0.20^{**}	(0.099-0.308)
Hong Kong	0.01	0.17^{**}			0.18^{**}	(0.059-0.309)
Hungary	0.05^{*}				0.05^{*}	(0.009-0.096)
Iran			0.03^{*}		0.03^{*}	(0.004-0.084)
Ireland		0.10^{**}			0.10^{**}	(0,022-0,213)
Jordan			0.05^{*}		0.05^{*}	(0.007-0.093)
Korea			0.03^{*}		0.03^{*}	(0.004-0.070)
Malaysia			0.06^{**}			(0.011-0.125)
Norway		0.04^{**}	0.03^{*}		0.07^{**}	(0.015-0.153)
Qatar		-0.06^{*}			-0.06^{*}	(-0.127; -0.009)
Singapore		0.07^{**}	0.04		0.11^{**}	(0.039-0.202)
Turkey			0.08^{**}		0.08^{**}	(0.007-0.193)
UAE		0.07^{**}	0.04	0.01	0.12^{**}	(0.076-0.178)
USA		0.10^{**}	0.04^{*}		0.14^{**}	(0.094-0.288)

$^{**}p$<0.01; * $^{**}p$<0.05. SSB: Sense of School Belonging; DS: Discipline at School; QMT: Quality of Mathematics Teaching at School; SAE: School Academic Emphasis.

4. DISCUSSION and CONCLUSION

In this study, we examined the mediating role of school characteristics between school SES and student mathematics achievement across educational systems. Results showed that effects of school SES vary across the participating countries in TIMSS 2015 and part of school SES effects on student mathematics achievement can be explained by school characteristics. Also, the test of mediation revealed that the school academic emphasis and sense of school belonging were the variables that might have the potential to transmit the effect of school SES on student mathematics achievement.

4.1. Predicting Mathematics Achievement

The finding on the effect of school SES on mathematics achievement is consistent with the work of Chmielewski (2019), which states that inequity resulting from school SES does not decrease and continues to exist in many countries. Previous studies suggested that the school SES might be less influencing on achievement within centralized educational systems, implementing a standard curriculum with central exams (Broer et al., 2019). In this study, results regarding some educational systems are in line with this view, such as Italy and Lebanon. However, although the educational system in Norway has a decentralized organization, the results regarding Norway does not support the relationship between school SES and mathematics achievement. Besides, results also show that the statistically significant positive relationship between school SES and achievement may exist in both centralized (e.g., Turkey) and decentralized (e.g., Korea, Sweden, and the USA) education systems. This situation signifies that the inequity in educational systems may not be explained only by a decentralized education system.

Concerning the effects of school characteristics on mathematics achievement, we found a positive significant effect of school academic emphasis in eighteen educational systems. This result confirms that students attending schools where they are encouraged to do their best with higher academic expectations might have a higher mathematics achievement in line with previous research (Akyuz 2014; Brault et al., 2014; Hoy et al., 2006; Nilsen & Gustafsson, 2014; Nilsen et al., 2016).

For the sense of school belonging, the effect was significant and positive in eighteen educational systems. This result is in line with the studies stating that students with a higher sense of belonging towards the school's social and academic structure might have higher mathematics achievement (Hoy et al., 2006; Nilsen & Gustafsson, 2014; Wang & Degol, 2016). Lei et al. (2016) claim that the relationship between the sense of school belonging and achievement might culturally be dependent and is higher in Western educational systems. The present study does not make a clear-cut distinction but confirms that this relationship is observed more in Western educational systems (England, Ireland, and the USA) and Western-dominated educational systems (Hong Kong, Singapore, Qatar and the UAE).

For the discipline at school, in most of the educational systems, the results are in line with the work of Ma and Wilkins (2002), who stated that discipline at school might not be a predictor of achievement when variables such as school academic emphasis and school SES were considered. However, there were some exceptions in which this effect was statistically significant. We found a statistically significant positive effect of discipline at school on student mathematics achievement in five educational systems (Bahreyn, Chile, Hong Kong, Hungary and South Africa), revealing that students perform better in mathematics in a safe and peaceful school climate with fewer discipline problems (McCoy et al., 2013; Nilsen & Gustafsson, 2014). Another exception was Korea, in which this effect was negative. In Korea, the disciplinary climate of schools might not have satisfied the expectations of secondary level students with higher mathematics achievement. This finding, however, was contrary to those of Shin et al. (2009), whose research findings indicated a positive relationship between school discipline and Korean students' PISA mathematics achievement.

Similarly, we did not observe a significant effect of the quality of mathematics teaching at school on student mathematics achievement in most of the educational systems. The exceptions in which this effect is positively significant are China-Taiwan, Japan, Kazakhstan, Malaysia, Malta, Morocco, Turkey, and the UAE. This finding supports the view that the quality of mathematics teaching is a key factor influencing student achievement at least in some of the countries (Baumert et al., 2010; Klieme et al., 2009). Other exceptions in which this effect is negatively significant are Hong Kong and Korea. The explanation of this negative effect might be the negative responses of students with a higher level of performance in mathematics. In these countries, students with higher mathematics performance may have higher teaching expectations from their schools and the schools may not satisfy these expectations.

4.2. Predicting School Characteristics

The effect of the school SES on the school characteristics differed across educational systems. With regard to the effect of school SES on discipline at school, we found a statistically significant positive effect in twenty-one educational systems. As stated by Brantlinger (2003), the reason of this positive effect might be attributable to the safe and ordered school climate demands of both students and their families in high SES schools. This result is also in line with the view that low SES schools have more disciplinary problems than those of the high SES schools (Bryk & Schneider, 2002; Thapa et al., 2013).

Similarly, we observed a positive effect of school SES on school academic emphasis in most of the educational systems, revealing that high SES schools have a high academic emphasis on achievement (Dumay & Dupriez, 2008; Nilsen & Gustafsson, 2014; Opdenakker & Van Damme, 2001). Bahrein, Israel, Malta, Slovenia and Qatar were the exceptions, where this effect was not statistically significant.

For the sense of school belonging, the effect of school SES was statistically significant in fifteen educational systems. The educational systems where the relation between school SES and sense of school belonging was positiveare mostly the developed countries such as Hong Kong, Japan, Singapore, and the USA. This result implies that schools in these countries might have more resources to connect students with high SES socially and emotionally to their schools. However, this effect was negative in Botswana, Morocco, South Africa, Qatar and Turkey. In this group of developing and low achieving countries, in contrast, schools might not have enough resources to fulfill expectations of students with high SES. Another explanation of this negative effect might be the negative rate of students with high SES due to their more critical perspectives towards their schools (Atlay et al., 2019).

In addition, we did not observe a statistically significant relationship between school SES and the quality of mathematics teaching at school across educational systems, with few exceptions. The effect of school SES on the quality of mathematics teaching at school was statistically significant and positive in Israel and the UAE, revealing that the quality of mathematics teaching in high SES schools is better than low SES schools. This effect was negative in Qatar, South Africa, and Thailand. Quality of teaching at school might be another school characteristics that students with high SES rate negatively due to their sense of entitlement (Atlay et al., 2019).

4.3. Mediating Role of School Characteristics

Previous research reported that school SES has an indirect effect on mathematics (Hoy et al., 2006) or science (Nilsen & Gustafsson, 2014) achievement through school academic emphasis. In this study, the finding of the mediating role of school academic emphasis between school SES and mathematics achievement in Botswana, Canada, Iran, Jordan, Korea, Malaysia, Norway, Turkey, and the USA is consistent with the previous research. Similarly, in this study, findings show that the sense of school belonging is another school characteristics that has the potential to mediate the relationship between school SES and student mathematics achievement in Australia, England, Hong Kong, Ireland, Norway, Singapore, and the USA. It seems that high SES schools in this group of educational systems might influence mathematics achievement by creating a healthy school environment and a sense of school belonging. However, the mediating role of the sense of school belonging in Qatar is unexpectedly negative, possibly due to negative ratings to items related to school belonging in high SES schools. These results show that the mechanism of school SES and achievement relationship differ across educational systems.

In their studies, Nielsen and Gustafsson (2014) stated that school academic emphasis has a mediating role in the relation between SES and science achievement in Norway. Combined with the current findings it appears that both school academic emphasis and sense of school belonging may have a mediating role on the relationship between school SES and mathematics achievement in Norway.

Previous research also stated that school disciplinary climate might have a mediating role between school SES and achievement (Berkowitz et al., 2015; Nilsen & Gustafsson, 2014; Liu et al., 2015). However, we observed only in Hungary, a small but statistically significant mediating role of discipline at school between school SES and mathematics achievement. Although school SES positively affects discipline at school in most of the countries, it seems that school SES does not influence student mathematics achievement through discipline at school. This finding shows the importance of examining the joint effects of school characteristics on student mathematics achievement.

In addition, in none of the educational systems, the results of this study did not support the view that mathematics teaching in high SES schools is more qualified, and thus students tend to have higher levels of achievement. This finding is similar to that of Hansen and Strietholt (2018), whose research findings suggested that the mediating role of quality of mathematics teaching, between SES (both in student and school levels) and mathematics achievement, may not be statistically significant in the presence of student self-confidence. Also, several studies showed that the quantity of instruction has a mediating role on the relation between school SES and achievement (Hansen & Strietholt, 2018; Rjosk et al., 2014; Schmidt et al., 2015). Combined with the current findings, it appears that the quality of instruction does not mediate the relationship of school SES and achievement as the quantity of instruction does. However, the mediating effects of quantity and quality of instruction should be detailed further by controlling the effect of student self-confidence.

The present study has some limitations. First, measurement of school characteristics is based on principal's or students' self-report measures. Studies considering observational data may strengthen the relationships found in this study. Second, the data used in this study are cross-sectional. Therefore, we cannot indicate definite causal conclusions. Third, the relations were examined in the mathematics domain, and the results might differ for other subjects.

In conclusion, we were able to find that the sense of school belonging and school academic emphasis might be two meaningful mediators accounting for the mechanism of school SES in some educational systems. Analyses suggest that the disadvantages of being a student in low SES schools might be alleviated if the sense of school belonging in low SES schools could be increased in Australia, England, Hong Kong, Norway, Singapore, the UAE, and the USA. This result emphasizes the importance of principals and teachers who are aware of the student needs such as the feeling of welcomed, respected, and supported by others in the school social environment. As students having a sense of school belonging are successful in mathematics, creating school environments to enhance students' feeling of a member of school/classroom life would be helpful to eliminate the negative effects of low SES. Similarly, analyses also suggest the need to give importance to academic emphasis in low SES schools to decrease the relationship between school SES and mathematics achievement in Botswana, Canada, Iran, Jordan, Korea, Malaysia, Norway, Turkey, and the USA. It seems that building a learning community that gives importance to academic achievement in low SES schools would improve the quality of student learning. Parents, teachers, and students are the members of the community who determine the school's emphasis on academic success. The notable involvement of these community members with a shared vision would contribute to reducing school SES effects. For example, parents, teachers, and students may benefit from parental engagement. When parents are engaged in education, they might have better contact with the teachers. In low SES schools, parents' and teachers' shared responsibility for encouraging learning may increase students' motivation to learn.

Declaration of Conflicting Interests and Ethics

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

Authorship Contribution Statement

Ozlem Albayrakoglu: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing - original draft. **Selda Yildirim**: Investigation, Resources, Methodology, Formal Analysis, Writing -original draft, Supervision, and Validation.

ORCID

Ozlem Albayrakoglu https://orcid.org/0000-0002-6234-5309
Selda Yildirim https://orcid.org/0000-0003-0535-4353

REFERENCES

Akyuz, G. (2014). The effects of student and school factors on mathematics achievement in TIMSS 2011. *Egitim ve Bilim*, *39*(172), 150-162.

Armor, D.J., Cotla, C.R., & Stratmann, T. (2017). Spurious relationships arising from aggregate variables in linear regression. *Quality and Quantity*, *51*(3), 1359-1379.

Atlay, C., Tieben, N., Fauth, B., & Hillmert, S. (2019). The role of socioeconomic background and prior achievement for students' perception of teacher support. *British Journal of Sociology of Education*, *40*(7), 970-991.

Bauer, D.J., Preacher, K.J., & Gil, K.M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: new procedures and recommendations. *Psychological Methods*, *11*(2), 142-163.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., & Tsai, Y.M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, *47*(1), 133–180.

Berkowitz, R., Glickman, H., Benbenishty, R., Ben-Artzi, E., Raz, T., Lipshtadt, N., & Astor, R.A. (2015). Compensating, mediating, and moderating effects of school climate on academic achievement gaps in Israel. *Teachers College Rec.*, *117*, article no: 070308, 1-34.

Berkowitz, R., Moore, H., Astor, R.A., & Benbenishty, R. (2017). A research synthesis of the associations between socioeconomic background, inequity, school climate, and academic achievement. *Review of Educational Research*, *87*(2), 425-469.

Boonen, T., Pinxten, M., Van Damme, J., & Onghena, P. (2014). Should schools be optimistic? An investigation of the association between academic optimism of schools and student achievement in primary education. *Educational Research and Evaluation, 20*, 3–24.

Borman, G., & Dowling, M. (2010). Schools and inequality: A multilevel analysis of Coleman's equality of educational opportunity data. *Teachers College Record*, *112*(5), 1201-1246.

Brantlinger, E.A. (2003). *Dividing classes: How the middle class negotiates and rationalizes school advantage.* Routledge Falmer.

Brault, M.C., Janosz, M., & Archambault, I. (2014). Effects of school composition and school climate on teacher expectations of students: A multilevel analysis. *Teaching and Teacher Education*, *44*, 148-159.

Broer, M., Bai, Y., & Fonseca, F. (2019). A Review of the Literature on Socioeconomic Status and Educational Achievement. In Socioeconomic *Inequality and Educational Outcomes* (pp. 7-17). Springer, Cham.

Brophy, J., & Good, T.L. (1986). *Teacher behavior and student achievement.* In M. C. Wittrock (Ed.), *Handbook of Research on Teaching* (3rd ed., pp. 328–375). Macmillan.

Bryk, A., & Schneider, B. (2002). *Trust in schools: A core resource for improvement*. Russell Sage Foundation.

Chmielewski, A.K. (2019). The global increase in the socioeconomic achievement gap, 1964 to 2015. *American Sociological Review*, *84*(3), 517-544. https://doi.org/10.1177/0003122419847165

Contini, D., DiTommaso, M.L., & Mendolia, S. (2017). The gender gap in mathematics achievement: Evidence from Italian data. *Economics of Education Review*, *58*, 32-42.

Driessen, G. (2002). School composition and achievement in primary education: A large scale multilevel approach. *Studies in Educational Evaluation, 28,* 347-368.

Dumay, X., & Dupriez, V. (2008). Does the school composition effect matter? Evidence from Belgian data. *British Journal of Educational Studies*, *56*, 440-477. http://dx.doi.org/10.1111/j.1467-8527.2008.00418.x

Eriksson, K., Helenius, O., & Ryve, A. (2019). Using TIMSS items to evaluate the effectiveness of different instructional practices. *Instructional Science*, *47*(1), 1-18.

Goodenow, C. (1993). The psychological sense of school membership among adolescents: Scale development and educational correlates. *Psychology in the Schools*, *30*(1), 79-90.

Gustafsson, J.E., Nilsen, T., & Hansen, K.Y. (2016). School characteristics moderating the relation between student socio-economic status and mathematics achievement in grade 8. Evidence from 50 countries in TIMSS 2011. *Studies in Educational Evaluation*, *57*, 16-30.

Hansen, K.Y., & Strietholt, R. (2018). Does schooling actually perpetuate educational inequality in mathematics performance? A validity question on the measures of opportunity to learn in PISA. *ZDM*, 1-16.

Hoy, W.K., Tarter, C.J., & Hoy, A.W. (2006). Academic optimism of schools: A force for student achievement. *American Educational Research Journal*, *43*(3), 425-446.

Hoy, W. (2012). School characteristics that make a difference for the achievement of all students: A 40-year odyssey. *Journal of Educational Administration*, *50*(1), 76-97.

International Association for the Evaluation of Educational Achievement. (2017). *IDB Analyzer (version 4.0)*. IEA Hamburg. http://www.iea.nl/data.html

Ker, H.W. (2016). The impacts of student- teacher and school-level factors on mathematics achievement: an exploratory comparative investigation of Singaporean students and the USA students. *Educational Psychology*, *36*(2), 254-276. https://doi.org/10.1080/01443410.2015.1026801

Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras study: Investigating effects of teaching and learning in Swiss and German Classrooms, in T. Janik, & T. Seidel (Eds.). *The power of video studies in investigating teaching and learning in the classroom*, pp. 137–160. Waxmann Verlag.

Krull, J.L., & MacKinnon, D.P. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate Behavioral Research,36*, 249-277.

Lei, H., Cui, Y. & M.M. Chiu. 2016. Affective teacher-student relationships and students 'externalizing behavior problems: A meta-analysis. *Frontiers in Psychology, 7*, 1-12. https://doi.org/10.3389/fpsyg.2016.01311

Liu, H., Van Damme, J., Gielen, S., & Van Den Noortgate, W. (2015). School processes mediate school compositional effects: model specification and estimation. *British Educational Research Journal*, *41*(3), 423-447.

Ma, X., & Wilkins, J.L. (2002). The development of science achievement in middle and high school: individual differences and school effects. *Evaluation Review*, *26*, 395-417.

Martin, M.O., Mullis, I.V.S., & Hooper, M. (Eds.). (2016). *Methods and Procedures in TIMSS 2015*.TIMSS and PIRLS International Study Center Boston College. http://timss.bc.edu/publications/timss/2015-methods.html

McCoy, D.C., Roy, A.L., & Sirkman, G.M. (2013). Neighborhood crime and school climate as predictors of elementary school academic quality: A cross-lagged panel analysis. *American Journal of Community Psychology*, *52*, 128–140. https://doi.org/10.1007/s10464-013-9583-5

Mullis, I.V.S., Martin, M.O., Foy, P., & Arora, A. (2011). *TIMSS 2011 International Results in Mathematics.* Boston College, TIMSS & PIRLS International Study Center, https://timssandpirls.bc.edu/timss2011/international-results-mathematics.html

Mullis, I.V.S., Martin, M.O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 International Results in Mathematics.* Boston College, TIMSS and PIRLS International Study Center. http://timssandpirls.bc.edu/timss2015/international-results/

National Council of Teachers of Mathematics [NCTM] (2000). *Principles and standards for school mathematics*. Author.Reston.VA.

Munk, T. (2007). *Full-school engagement as a mediator of ethnic and economic composition effects on grade 8 mathematics test scores: a two-level structural equation model.* [Unpublished doctoral dissertation]. https://doi.org/10.17615/8n2x-6s56

Nilsen, T., Blömeke, S., Hansen, K.Y., & Gustafsson, J.E. (2016). Are school characteristics related to equity? The answer may depend on a country's developmental level. *International Association for the Evaluation of Educational Achievement*. Policy Brief No. 10.

Nilsen, T., & Gustafsson, J.E. (2014). School emphasis on academic success: Exploring changes in science performance in Norway between 2007 and 2011 employing two-level SEM. *Educational Research and Evaluation*, *20*(4), 308-327.

Opdenakker, M.C., &Van Damme, J. (2001). Relationship between school composition and characteristics of school process and their effect on mathematics achievement. *British Educational Research Journal*, *27*(4), 407-432.

Olmez, I.B. (2020). Modeling mathematics achievement using hierarchical linear models. *Elementary Education Online, 19*(2), 944-957. https://doi:10.17051/ilkonline.2020.695837

Perry, L.B., & McConney, A. (2010). Does the SES of the school matter? An examination of socioeconomic status and student achievement using PISA 2003. *Teachers College Record*, *112*(4), 1137-1162.

Preacher, K.J., & Selig, J.P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures*, *6*(2), 77-98.

R Development Core Team (2017). *R: A language and environment for statistical computing*. The R foundation of statistical computing.

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage.

Rjosk, C., Richter, D., Hochweber, J., Lüdtke, O., Klieme, E., & Stanat, P. (2014). Socioeconomic and language minority classroom composition and individual reading achievement: The mediating role of instructional quality. *Learning and Instruction*, *32*, 63-72.

Rumberger, R. W., & Palardy, G. J. (2005). Does segregation still matter? The impact of student composition on academic achievement in high school. *Teachers'College Record*, *107*(9), 1999-2045.

Schmidt, W.H., Burroughs, N.A., Zoido, P., & Houang, R.T. (2015). The role of schooling in perpetuating educational inequality: An international perspective. *Educational Researcher*, *44*(7), 371-386.

Shin, J., Lee, H., & Kim, Y. (2009). Student and school factors affecting mathematics achievement: International comparisons between Korea, Japan and the USA. *School Psychology International*, *30*(5), 520-537.

Sirin, S.R. (2005). Socioeconomic status and academic achievement: A meta-analysis. *Review of Educational Research*, *75*(3), 417–453.

Thapa, A., Cohen, J., Guffey, S., & Higgins-D'Alessandro, A. (2013). A review of school climate research. *Review of Educational Research*, *83*(3), 357-385.

Thrupp, M., Lauder, H., & Robinson, T. (2002). School composition and peer effects. *International Journal of Educational Research*, *37*(5), 483-504.

Wang, Z., Osterlind, S.J., & Bergin, D.A. (2012). Building mathematics achievement models in four countries using TIMSS 2003. *International Journal of Science and Mathematics Education,10*(5), 1215-1242.

Wang, M.T., & Degol, J.L. (2016). School climate: A review of the construct, measurement, and impact on student outcomes. *Educational Psychology Review*, *28*(2), 315-352.

Wu, M. (2004). Plausible values. *Rasch Measurement Transactions*, *18*(2), 976-978.

Wu, J.H., Hoy, W.K., & Tarter, C.J. (2013). Enabling school structure, collective responsibility, and a culture of academic optimism. *Journal of Educational Administration*, *51*(2), 176-193.

Yavuz, H.C., Demirtasli, R.N., Yalcin, S., & Dibek, M.I. (2017). The effects of student and teacher level variables on TIMSS 2007 and 2011 mathematics achievement of Turkish students. *Education and Science, 42*(189), 27-47.

Zhang, Z., Zyphur, M.J., & Preacher, K.J. (2009). Testing multilevel mediation using hierarchical linear models: Problems and solutions. *Organizational Research Methods*, *12*(4), 695-719.

APPENDIX

Appendix 1. *Distribution of students and schools participating in TIMSS 2015 by country.*

	Student	School		Student	School
	N	N		N	N
Australia	10.338	285	Kuwait	4.503	168
Bahrein	4.918	105	Lebanon	3.873	138
Botswana	5.964	159	Lithuania	4.347	208
Canada	8.757	276	Malaysia	9.726	207
Chile	4.849	171	Malta	3.817	48
China-Taiwan	5.711	190	Morocco	13.035	100
Egypt	7.822	211	New Zealand	8.142	145
England	4.814	143	Norway	4.697	143
Georgia	4.035	153	Oman	8.883	301
Hong Kong	4.155	133	Qatar	5.403	131
Hungary	4.893	144	Russia	4.87	204
Iran	6.13	250	Singapore	6.116	167
Ireland	4.704	149	Slovenia	4.257	148
Israel	5.223	189	South Africa	12.514	292
Italy	4.481	161	Sweden	4.09	150
Japan	4.745	147	Thailand	6.482	204
Jordan	7.865	252	Turkey	6.079	218
Kazakhstan	4.887	172	USA	10.221	246
Korea	5.309	150	UAE	18.012	477

Appendix 2. *Descriptive Statistics about Countries Participating in TIMSS 2015 Assessment*.

	Student Level Descriptive Statistics											
	Student Gender		Confidence in Mathematics Learning		Student SES		Sense of School Belonging		Quality of Mathematics Teaching		TIMSS 2015 Mathematics Achievement	
	$\bar{X}$ SE	SD SE	$\bar{X}$ SE	SD SE	$\bar{X}$ SE	SD SE	$\bar{X}$ SE	SD SE	$\bar{X}$ SE	SD SE	$\bar{X}$ SE	SD SE
USA	1.50(0.01)	0.50(0.00)	1.81(0.01)	0.75(0.00)	2.15(0.01)	0.51(0.00)	2.23(0.01)	0.67(0.00)	2.22(0.02)	0.76(0.01)	518.30(3.08)	83.3(1.58)
Australia	1.49(0.02)	0.50(0.00)	1.72(0.01)	0.70(0.01)	2.19(0.01)	0.48(0.01)	2.29(0.01)	0.65(0.01)	2.09(0.02)	0.75(0.01)	504.96(3.1)	82.36(1.89)
UAE	1.50(0.03)	0.50(0.00)	1.83(0.01)	0.69(0.00)	2.01(0.01)	0.47(0.01)	2.02(0.01)	0.75(0.01)	2.31(0.01)	0.70(0.01)	464.78(2.0)	97.9(1.51)
Bahrein	1.52(0.01)	0.50(0.00)	1.72(0.01)	0.70(0.01)	1.95(0.01)	0.46(0.01)	2.28(0.02)	0.66(0.01)	2.22(0.03)	0.76(0.01)	453.95(1.44)	80.33(1.41)
Botswana	1.49(0.01)	0.54(0.00)	1.61(0.01)	0.62(0.01)	1.55(0.01)	0.54(0.00)	2.47(0.01)	0.59(0.01)	2.49(0.02)	0.66(0.01)	390.84(2.04)	83.40(1.13)
China-Taiwan	1.51(0.01)	0.50(0.00)	1.49(0.01)	0.66(0.01)	2.03(0.01)	0.52(0.01)	2.17(0.01)	0.58(0.01)	1.98(0.03)	0.69(0.01)	599.11(2.42)	97.18(1.69)
Morocco	1.54(0.01)	0.50(0.00)	1.68(0.01)	0.63(0.00)	1.47(0.01)	0.53(0.00)	2.70(0.01)	0.51(0.01)	2.52(0.02)	0.66(0.01)	384.39(2.25)	80.05(1.27)
South Africa	1.49(0.01)	0.50(0.00)	1.62(0.01)	0.65(0.01)	1.71(0.01)	0.51(0.01)	2.56(0.01)	0.57(0.01)	2.52(0.02)	0.64(0.01)	372.37(4.53)	87.07(3.02)
Georgia	1.53(0.01)	0.50(0.00)	1.67(0.02)	0.68(0.01)	2.16(0.01)	0.52(0.01)	2.38(0.01)	0.58(0.01)	2.44(0.02)	0.63(0.01)	453.20(3.44)	91.96(1.71)
Hong Kong	1.53(0.02)	0.50(0.00)	1.56(0.01)	0.67(0.01)	1.97(0.02)	0.51(0.01)	2.16(0.02)	0.65(0.01)	2.02(0.03)	0.71(0.01)	594.25(4.62)	78.41(2.80)
UK	1.49(0.02)	0.50(0.00)	1.80(0.02)	0.67(0.01)	2.14(0.01)	0.46(0.01)	2.24(0.02)	0.62(0.01)	2.18(0.03)	0.73(0.01)	518.26(4.17)	79.84(2.62)
Iran	1.52(0.01)	0.50(0.00)	1.74(0.02)	0.72(0.01)	1.73(0.02)	0.61(0.01)	2.38(0.02)	0.61(0.01)	2.43(0.02)	0.69(0.01)	436.35(4.64)	94.08(2.73)
Ireland	1.50(0.01)	0.50(0.00)	1.73(0.02)	0.71(0.01)	2.15(0.01)	0.48(0.01)	2.32(0.02)	0.64(0.01)	2.15(0.02)	0.75(0.01)	523.49(2.73)	73.95(2.31)
Israel	1.51(0.01)	0.50(0.00)	1.92(0.02)	0.74(0.01)	2.06(0.01)	0.39(0.01)	2.38(0.02)	0.66(0.01)	2.25(0.02)	0.76(0.01)	510.90(4.10)	102.01(2.32)
Sweden	1.52(0.01)	0.50(0.00)	1.76(0.02)	0.73(0.01)	2.19(0.01)	0.47(0.01)	2.25(0.02)	0.61(0.01)	2.10(0.03)	0.70(0.01)	500.72(2.76)	71.96(1.89)
Italy	1.51(0.00)	0.50(0.01)	1.75(0.01)	0.75(0.02)	1.97(0.01)	0.52(0.02)	2.16(0.01)	0.60(0.01)	2.13(0.01)	0.69(0.02)	494.39(1.75)	74.54(2.52)
Japan	1.49(0.01)	0.50(0.00)	1.41(0.01)	0.58(0.01)	2.16(0.01)	0.45(0.01)	2.14(0.02)	0.62(0.01)	1.70(0.02)	0.64(0.01)	586.47(2.27)	88.90(1.28)
Canada	1.49(0.01)	0.50(0.00)	1.92(0.01)	0.75(0.01)	2.19(0.01)	0.44(0.01)	2.37(0.01)	0.61(0.01)	2.32(0.02)	0.69(0.01)	527.28(2.15)	69.76(1.27)
Qatar	1.50(0.03)	0.50(0.00)	1.7680.01)	0.69(0.01)	2.05(0.01)	0.46(0.01)	2.23(0.02)	0.75(0.01)	2.23(0.02)	0.75(0.01)	437.11(2.99)	102.22(2.20)
Kazakhstan	1.51(0.01)	0.50(0.00)	1.86(0.02)	0.64(0.01)	2.00(0.02)	0.46(0.01)	2.64(0.01)	0.50(0.01)	2.45(0.02)	0.57(0.01)	527.81(5.28)	93.23(2.26)
Korea	1.53(0.01)	0.50(0.00)	1.53(0.01)	0.63(0.01)	2.35(0.01)	0.53(0.00)	2.17(0.01)	0.53(0.01)	1.67(0.02)	0.61(0.01)	605.74(2.60)	85.29(1.07)
Kuwait	1.50(0.03)	0.50(0.00)	1.78(0.02)	0.67(0.01)	1.92(0.01)	0.40(0.01)	2.45(0.02)	0.61(0.01)	2.35(0.02)	0.69(0.01)	392.47(4.65)	91.07(3.33)
Lithuania	1.50(0.01)	0.50(0.00)	1.75(0.02)	0.70(0.01)	2.09(0.01)	0.43(0.01)	2.30(0.02)	0.60(0.01)	2.22(0.03)	0.71(0.01)	511.31(2.77)	77.32(1.53)
Lebanon	1.47(0.02)	0.50(0.00)	1.79(0.02)	0.69(0.01)	1.87(0.01)	0.50(0.01)	2.45(0.01)	0.63(0.01)	2.52(0.02)	0.68(0.01)	442.43(3.63)	75.26(1.72)
Hungary	1.50(0.01)	0.50(0.00)	1.77(0.02)	0.74(0.01)	2.15(0.02)	0.52(0.01)	2.17(0.02)	0.63(0.01)	2.14(0.03)	0.72(0.01)	514.41(3.78)	93.39(2.24)
Malaysia	1.50(0.02)	0.50(0.00)	1.49(0.01)	0.57(0.00)	1.82(0.01)	0.49(0.01)	2.42(0.02)	0.57(0.01)	2.29(0.02)	0.65(0.01)	465.31(3.57)	86.64(2.05)
Malta	1.51(0.00)	0.50(0.00)	1.64(0.01)	0.70(0.00)	2.01(0.01)	0.50(0.01)	2.17(0.01)	0.67(0.01)	2.19(0.01)	0.76(0.01)	493.54(0.99)	88.44(0.88)
Egypt	1.47(0.02)	0.50(0.00)	1.80(0.02)	0.67(0.01)	1.76(0.01)	0.52(0.01)	2.56(0.02)	0.62(0.01)	2.57(0.02)	0.64(0.01)	392.23(4.12)	98.56(2.01)
Norway	1.50(0.01)	0.50(0.00)	1.87(0.02)	0.76(0.01)	2.28(0.01)	0.48(0.01)	2.45(0.02)	0.62(0.01)	2.10(0.03)	0.74(0.01)	511.54(2.25)	70.05(1.22)
Russia	1.51(0.01)	0.50(0.00)	1.66(0.02)	0.68(0.01)	2.08(0.01)	0.48(0.01)	2.28(0.01)	0.61(0.01)	2.33(0.02)	0.66(0.01)	538.00(4.66)	81.71(1.76)
Singapore	1.51(0.01)	0.50(0.00)	1.67(0.01)	0.69(0.01)	2.00(0.01)	0.48(0.01)	2.28(0.01)	0.61(0.01)	2.17(0.02)	0.67(0.01)	620.96(3.20)	82.13(2.15)
Slovenia	1.52(0.01)	0.50(0.00)	2.32(0.01)	0.67(0.01)	1.90(0.01)	0.39(0.01)	2.10(0.02)	0.57(0.01)	2.02(0.02)	0.64(0.01)	516.34(2.09)	69.35(1.35)
Chile	1.52(0.02)	0.50(0.00)	1.60(0.02)	0.69(0.01)	1.90(0.01)	0.45(0.01)	2.39(0.02)	0.67(0.01)	2.28(0.03)	0.76(0.01)	427.43(3.22)	79.96(1.92)
Thailand	1.46(0.02)	0.50(0.00)	1.34(0.01)	0.53(0.01)	1.66(0.01)	0.53(0.01)	2.56(0.01)	0.54(0.00)	2.34(0.02)	0.64(0.01)	431.42(4.76)	89.18(3.40)
Turkey	1.52(0.01)	0.50(0.00)	1.60(0.02)	0.72(0.01)	1.67(0.02)	0.59(0.01)	2.52(0.01)	0.61(0.01)	2.50(0.02)	0.66(0.01)	457.63(4.74)	105.41(2.78)
Oman	1.52(0.02)	0.50(0.00)	1.86(0.01)	0.68(0.00)	1.78(0.01)	0.53(0.00)	2.57(0.01)	0.58(0.01)	2.51(0.02)	0.63(0.01)	403.16(2.43)	96.13(1.29)
Jordan	1.50(0.03)	0.50(0.00)	1.82(0.01)	0.70(0.01)	1.83(0.01)	0.49(0.01)	2.60(0.02)	0.62(0.01)	2.60(0.02)	0.62(0.01)	385.55(3.23)	93.83(1.73)
New Zealand	1.49(0.02)	0.50(0.00)	1.68(0.01)	0.67(0.01)	2.12(0.01)	0.48(0.01)	2.35(0.01)	0.61(0.01)	2.09(0.03)	0.73(0.01)	492.72(3.36)	87.88(2.04)

School Level Descriptive Statistics

	School SES		Sense School Belonging		Discipline at School		Quality of Mathematics Teaching at School		School Academic Emphasis	
	$\bar{X}$ SE	SD SE	$\bar{X}$ SE	SD SE	$\bar{X}$ SE	SD SE	$\bar{X}$ SE	SD SE	$\bar{X}$ SE	SD SE
USA	2.53(0.09)	0.56(0.04)	2.33(0.04)	0.27(0.02)	2.53(0.01)	0.56(0.06)	2.28(0.04)	0.2980.02)	1.6(0.05)	0.53(0.04)
Australia	1.93(0.05)	0.71(0.04)	2.25(0.03)	0.29(0.03)	2.49(0.05)	0.51(0.01)	2.11(0.02)	0.31(0.02)	1.69(0.05)	0.66(0.03)
UAE	2.12(0.04)	0.84(0.02)	1.93(0.02)	0.49(0.01)	2.50(0.03)	0.56(0.02)	2.27(0.01)	0.29(0.01)	1.91(0.03)	0.59(0.02
Bahrein	2.21(0.06)	0.76(0.05)	2.32(0.02)	0.25(0.01)	2.47(0.07)	0.67(0.04)	2.28(0.03)	0.25(0.01)	1.73(0.05)	0.62(0.03)
Botswana	1.53(0.05)	0.72(0.03)	2.46(0.01)	0.15(0.01)	1.94(0.05)	0.57(0.03)	2.48(0.02)	0.24(0.02)	1.20(0.02)	0.45(0.03)
China-Taiwan	1.90(0.09)	0.68(0.04)	2.18(0.02)	0.17(0.01)	2.53(0.07)	0.52(0.01)	1.98(0.05)	0.31(0.02)	1.41(0.04)	0.55(0.02)
Morocco	1.57(0.05)	0.80(0.03)	2.71(0.02)	0.17(0.01)	1.71(0.07)	0.76(0.03)	2.53(0.02)	0.28(0.02)	1.25(0.03)	0.45(0.02)
South Africa	1.37(0.08)	0.69(0.08)	2.54(0.02)	0.18(0.01)	1.83(0.06)	0.67(0.04)	2.52(0.02)	0.25(0.02)	1.33(0.03)	0.50(0.02)
Georgia	1.65(0.08)	0.79(0.04)	2.41(0.03)	0.21(0.02)	2.66(0.04)	0.56(0.03)	2.46(0.02)	0.23(0.02)	1.67(0.05)	0.49(0.02)
Hong Kong	1.72(0.05)	0.73(0.03)	2.14(0.02)	0.24(0.02)	2.65(0.04)	0.48(0.02)	2.02(0.03)	0.31(0.02)	1.50(0.04)	0.60(0.03)
UK	2.08(0.05)	0.71(0.03)	2.23(0.02)	0.27(0.01)	2.76(0.03)	0.37(0.02)	2.18(0.04)	0.35(0.02)	2.05(0.04)	0.62(0.03)
Iran	1.58(0.05)	0.78(0.03)	2.43(0.03)	0.24(0.01)	2.61(0.05)	0.55(0.02)	2.50(0.03)	0.31(0.02)	1.49(0.05)	0.59(0.04)
Ireland	1.90(0.07)	0.84(0.05)	2.31(0.02)	0.22(0.02)	2.60(0.05)	0.52(0.03)	2.16(0.03)	0.27(0.02)	1.92(0.05)	0.61(0.04)
Israel	1.69(0.06)	0.75(0.03)	2.37(0.03)	0.27(0.02)	2.16(0.06)	0.64(0.03)	2.27(0.03)	0.29(0.02)	1.63(0.05)	0.54(0.02)
Sweden	2.40(0.07)	0.74(0.04)	2.25(0.02)	0.23(0.02)	2.24(0.05)	0.52(0.03)	2.09(0.03)	0.28(0.02)	1.49(0.06)	0.59(0.04)
Italy	2.03(0.09)	0.79(0.08)	2.16(0.02)	0.20(0.01)	2.18(0.07)	0.62(0.04)	2.13(0.03)	0.32(0.02)	1.31(0.04)	0.46(0.02)
Japan	2.38(0.05)	0.66(0.03)	2.22(0.05)	0.25(0.04)	2.55(0.07)	0.59(0.03)	1.78(0.07)	0.34(0.04)	1.58(0.07)	0.52(0.02)
Canada	2.05(0.06)	0.76(0.03)	2.38(0.02)	0.23(0.01)	2.52(0.04)	0.49(0.01)	2.37(0.04)	0.31(0.02)	1.60(0.05)	0.61(0.03)
Qatar	2.73(0.05)	0.62(0.05)	2.21(0.02)	0.28(0.02)	2.47(0.06)	0.67(0.04)	2.24(0.03)	0.30(0.02)	2.03(0.05)	0.63(0.04)
Kazakhstan	2.64(0.07)	0.59(0.06)	2.68(0.02)	0.19(0.01)	2.54(0.06)	0.75(0.04)	2.48(0.03)	0.21(0.01)	1.91(0.05)	0.48(0.05)
Korea	1.62(0.05)	0.64(0.03)	2.18(0.03)	0.16(0.01)	2.57(0.05)	0.59(0.03)	1.75(0.05)	0.27(0.05)	1.83(0.09)	0.59(0.05)
Kuwait	1.85(0.10)	0.82(0.04)	2.45(0.02)	0.24(0.02)	2.14(0.07)	0.70(0.03)	2.35(0.03)	0.32(0.02)	1.70(0.08)	0.62(0.07)
Lithuania	2.11(0.07)	0.77(0.03)	2.33(0.02)	0.23(0.01)	2.41(0.05)	0.53(0.02)	2.25(0.03)	0.31(0.03)	1.57(0.06)	0.53(0.02)
Lebanon	1.54(0.10)	0.84(0.06)	2.48(0.02)	0.24(0.01)	2.26(0.07)	0.80(0.03)	2.55(0.03)	0.28(0.02)	1.57(0.05)	0.54(0.02)
Hungary	1.73(0.07)	0.75(0.03)	2.18(0.02)	0.24(0.01)	2.18(0.05)	0.58(0.04)	2.16(0.04)	0.33(0.02)	1.58(0.04)	0.50(0.01)
Malaysia	1.40(0.05)	0.64(0.04)	2.46(0.02)	0.18(0.01)	2.52(0.06)	0.55(0.04)	2.34(0.02)	0.24(0.01)	1.83(0.06)	0.57(0.03)
Malta	2.30(0.07)	0.54(0.04)	2.22(0.03)	0.23(0.02)	2.51(0.08)	0.61(0.06)	2.20(0.03)	0.23(0.03)	1.83(0.06)	0.60(0.05)
Egypt	1.81(0.07)	0.77(0.04)	2.55(0.02)	0.23(0.01)	1.80(0.08)	0.79(0.03)	2.55(0.03)	0.25(0.02)	1.48(0.06)	0.58(0.03)
Norway	2.47(0.07)	0.58(0.04)	2.48(0.05)	0.25(0.03)	2.78(0.04)	0.40(0.03)	2.19(0.07)	0.38(0.05)	1.43(0.05)	0.48(0.01)
Russia	2.21(0.08)	0.85(0.04)	2.33(0.03)	0.31(0.03)	2.65(0.05)	0.50(0.03)	2.40(0.03)	0.30(0.02)	1.19(0.04)	0.39(0.03)
Singapore	2.14(0.05)	0.67(0.03)	2.27(0.01)	0.21(0.01)	2.73(0.03)	0.45(0.02)	2.18(0.02)	0.22(0.01)	1.80(0.04)	0.56(0.03)
Slovenia	1.82(0.05)	0.71(0.03)	1.92(0.02)	0.18(0.01)	1.56(0.04)	0.54(0.02)	2.02(0.03)	0.24(0.02)	2.64(0.05)	0.48(0.02)
Chile	1.36(0.04)	0.64(0.03)	2.40(0.03)	0.26(0.01)	2.20(0.06)	0.62(0.04)	2.35(0.03)	0.34(0.02)	1.27(0.04)	0.46(0.03)
Thailand	1.19(0.07)	0.61(0.07)	2.63(0.03)	0.19(0.01)	2.49(0.06)	0.58(0.04)	2.43(0.02)	0.22(0.01)	1.57(0.08)	0.57(0.03)
Turkey	1.55(0.09)	0.78(0.05)	2.60(0.03)	0.21(0.01)	1.94(0.08)	0.74(0.02)	2.59(0.03)	0.25(0.02)	1.34(0.06)	0.56(0.06)
Oman	2.11(0.06)	0.76(0.03)	2.56(0.02)	0.24(0.02)	2.32(0.06)	0.80(0.02)	2.52(0.02)	0.27(0.02)	1.62(0.03)	0.54(0.02)
Jordan	1.56(0.08)	0.73(0.04)	2.64(0.03)	0.23(0.02)	2.22(0.08)	0.78(0.03)	2.64(0.02)	0.22(0.02)	1.46(0.06)	0.55(0.01)
New Zealand	1.95(0.06)	0.82(0.04)	2.34(0.02)	0.23(0.02)	2.27(0.05)	0.56(0.03)	2.10(0.03)	0.28(0.02)	1.83(0.05)	0.64(0.03)

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 1, 118–137

https://doi.org/10.21449/ijate.909616

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate **_Research Article_**

Multidimensional Computerized Adaptive Testing Simulations in R

F. Gul Ince Araci[1,*], **Seref Tan**[1]

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

ARTICLE HISTORY

Received: Apr. 04, 2021
Revised: Nov. 25, 2021
Accepted: Dec. 18, 2021

Keywords:

Computerized adaptive testing,
Multidimensional,
R,
MCAT,
Simulation.

Abstract: Computerized Adaptive Testing (CAT) is a beneficial test technique that decreases the number of items that need to be administered by taking items in accordance with individuals' own ability levels. After the CAT applications were constructed based on the unidimensional Item Response Theory (IRT), Multidimensional CAT (MCAT) applications have gained momentum with the improvement of multidimensional IRT (MIRT) models in recent years. Researchers often benefit from simulation studies in order to design the final adaptive testing application and to test the effectiveness of adaptive testing applications they developed with different methods. Recently, R has become one of the most widely used programming languages in Monte Carlo Simulation studies since it is a free and open-source software. The aims of this study are to present the MCAT simulation process step by step in the R environment, to examine the effects of the conditions that researchers can handle during the simulation process according to two different dimensional models, and to examine the effect of treating multidimensional structures as unidimensional structures on simulation results. In this direction, datasets generated in accordance with within-item dimensionality and between-item dimensionality models, MCAT simulation studies were constructed with different customizations, and MCAT simulation results were compared with unidimensional CAT simulation results. All commands required for each simulation example were explained and results were shared for each condition.

1. INTRODUCTION

The integration of computers and the internet into education has gained tremendous momentum through the development of information technology. Although most of the exams are still applied as a paper-and-pencil method starting from the primary education level, internet-based distance education and test applications are rapidly increasing. In paper-and-pencil exam applications, the items that each examinee is expected to answer, the number of items are the same. The connection between abilities of an individual and the properties of the items are not taken into consideration. In other words, item difficulty cannot be matched with the examinee's ability. However, in CAT applications, the individual encounters the properties of items determined according to his/her ability level during the application process. In this way, test applications can be conducted with fewer items tailored to the individual, shorter time, and higher reliability. Moreover, test results and feedback can be presented to the individual as soon as the test ends (Weiner, 1993, Segall, 2005; Weiss

*CONTACT: F. Gul INCE ARACI ✉ gulincegazi@gmail.com Gazi University, Gazi Faculty of Education, Department of Measurement and Evaluation in Education, Ankara, Turkiye.

e-ISSN: 2148-7456 /© **IJATE** 2022

& Gibbons, 2007, Lin 2012). In addition, cheating during the testing time in cognitive tests is significantly reduced in online exams with CAT applications by using different questions to individuals. In addition to its advantages for individuals, CAT has also enabled researchers to form their own tests with different methods. Researchers can perform a wide variety of applications by differentiating the IRT model, starting the test, choosing the item to be administered, estimating the ability parameter, and changing the test stopping rules. For instance, when designing a test in order to decide which IRT model is suitable, simulation applications can be used to decide how many items will be included in the test and what the item exposure rate will be. Simulation applications allow the comparison of CAT applications constructed under different methods and constraints (Thompson & Wise, 2011; Meneghetti & Junior, 2018). Researchers can run CAT simulations based on various datasets: Monte Carlo simulations by generating data, post-hoc simulations on the basis of parameters derived from real-time applications, or hybrid simulation by imputing missing values to the real-time applications (Nydick & Weiss, 2009).

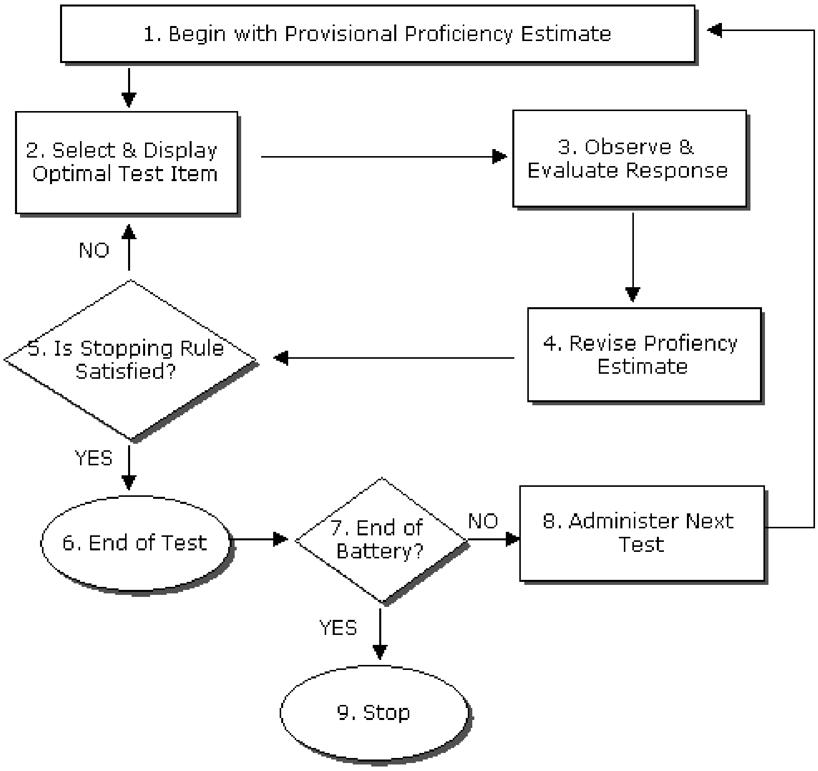
IRT models are frequently used in Monte Carlo Simulation studies in the field of psychometrics (Bulut & Sünbül, 2017). Most of the CAT studies performed in the literature are based on unidimensional IRT. Nevertheless, many psychological structures are multidimensional. Through MCAT applications developed using multidimensional IRT (MIRT) models in multidimensional structures, it is possible to decrease the number of items required to be administered to an individual to increase the precision of measurement and measure multiple traits at the same time (Seo & Weiss, 2015; Chalmers, 2016). In order to take these advantages of MCATs, there is a variety of software developed. One of the most popular software is R, which is a free and open-source platform. Real-time or simulation applications of MCATs, which researchers can customize by writing their own functions, can be performed on R (R Core Team, 2020). R allows researchers to customize their own applications by writing their own functions. The mirtCAT package (Chalmers, 2016) in R, which allows researchers to develop customized MCAT applications by writing their own code, consists of utile tools performing Monte Carlo simulations, and it is the only package that allows MCAT applications for now.

1.1. Computerized Adaptive Testing

Adaptive testing is an advanced test application where examinees encounter items according to their abilities, which is estimated based on the response pattern. Each individual completes a tailored test by preventing them from taking easy and difficult questions for them. Therefore, examinees encounter fewer items and save time (Embretson & Reise; 2000; Van der Linden, 2002).

In adaptive testing applications, the process starts with temporary θ estimation for an examinee. Frequently, the starting θ is considered to be 0. By presenting the first item in accordance with the starting rule to the examinee, the estimation is performed again, and the items are presented according to the item selection rule. This process continues until the stopping criterion is met. If the test does not consist of subscales, the application is stopped when the stopping criterion is met. If it consists of subscales, the other test starts, and the same procedures are repeated. The flowchart (Thissen & Mislevy, 2000), which represents the application process of adaptive testing, is presented in Figure 1.

Figure 1. *A flowchart showing the adaptive testing process**


1. Begin with Provisional Proficiency Estimate
2. Select & Display Optimal Test Item
3. Observe & Evaluate Response
4. Revise Profiency Estimate
5. Is Stopping Rule Satisfied?
NO
YES
6. End of Test
7. End of Battery?
NO
YES
8. Administer Next Test
9. Stop

*: From Thissen, D. and Mislevy, R. J., 2000. Testing Algorithms. In H. Wainer (Eds.). *Computerized Adaptive Testing: A Primer* (p.106). Lawrence Erlbaum Assc. Copyright 2000 by Lawrence Erlbaum Associates, Publishers."

Adaptive tests can be developed as a unidimensional CAT and MCAT according to the dimension of the measured construct. Six components are required to carry out an MCAT application (Weiss & Kingsbury, 1984; Thompson & Weiss, 2011; Chalmers, 2016):

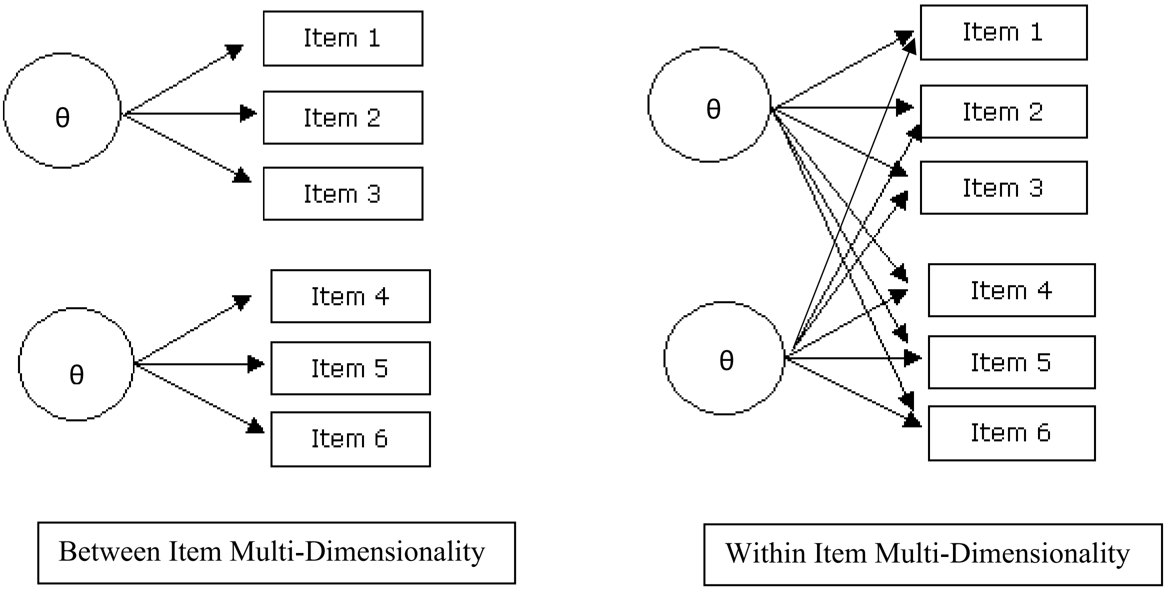
1. Multidimensional Item Response Theory (MIRT) model,
2. Calibrated item pool,
3. Starting rule,
4. Item selection method,
5. Estimation method,
6. Stopping rule.

These components are briefly explained below:

MIRT model**:** The interaction between examinees and test items may not always be unidimensional because, while answering the test item, the individual may need to use more than one ability or skill field, so MIRT models may be required in complex structures (Bock & Aitkin, 1981, Reckase, 2009, Chen, 2012). MIRT models are divided into two according to item level: within-item dimensionality and between-item dimensionality models. In the within-item dimensionality model, items load to all dimensions, and in the between-item dimensionality model, each item loads to a specific dimension as shown in Figure 2 (Wang & Chen, 2004).

One of the crucial steps in implementing adaptive testing applications is determining the model to be used in item bank calibration, ability estimation, and item selection methods (Magis et al., 2017). The methods to be used vary in accordance with the dimensionality of the model and whether the items have dichotomous or polytomous scored response categories (Weiss & Kingsbury, 1984, Ackerman, 1991; Wang & Chen, 2004, De Ayala, 2009). Therefore, it should be decided on the MIRT model to be used before MCATs are formed.

Figure 2. *Between and Within Item Multi-Dimensionality (Wang & Chen, 2004).*



Calibrated item pool: An item pool consisting of many quality items to be used during MCAT application should be created. There should be a sufficient number of items in the item pool suitable for individuals with different levels of the measured attribute. According to Stocking (1994), the item pool should have at least six times as many items as the test length. And, Reckase (2003) stated that a pool of approximately 200 items is appropriate for examinees sampled from a standard normal distribution. For CAT applications to work efficiently, it is essential to have sufficient quantity and quality of items. On the other hand, a high number of items will not be sufficient by itself. Many researchers have stated that the distribution of item parameters, content weighting, and item exposure rates should also be taken into consideration while developing the pool (He & Reckase, 2013).

Starting Rule: At this stage of MCAT application, it is usually required to define the initial estimation of the latent trait and the hyper-parameter distributions. Hyper-parameters are parameters obtained from the preliminary distribution without the real dataset observed. Suppose the first item administered to the examinee during the application is not determined specifically. In that case, the latent trait's initial estimation is used to determine the first item to be selected. Hyper-parameter distributions are used as a component in the item selection method and they provide prior distribution information while updating the latent trait estimation after the individual responds to each item during the MCAT application. MCATs can also be started with the methods of starting from the first item, starting according to the item selection method, and assigning the average ability level of the population as the initial theta. When the initial value of the examinees' latent trait estimates is unknown, it is common to assume it 0 (Thompson, 2007; Riggelsen, 2008; Chalmers, 2016).

Item selection method: During the MCAT application, after the examinee encounters the first item and estimates the ability parameters, the item selection method should be determined for determining which item will be presented next. Item selection methods are usually based on the idea of maximizing information about an examinee's location on the θ-coordinate or minimizing the error in the location estimation The basis of all item selection methods is maximizing or minimizing some criterion values in the final θ estimation. What makes these methods different from each other is the definition of the criterion (Reckase, 2009). While there are many item selection methods for CAT applications in the literature, there is a limited number of item selection methods available for MCAT applications. Some of these criteria are as follows: A-rule, E-rule, D-rule, T-rule, W-rule, Kullback-Leibler Information Criteria (KL), and Continuous Entropy method. While the basis of A-rule method is minimizing trace of the asymptotic co-variance matrix, the basis of E-rule is minimizing the information matrix and Wrule method based on maximizing the weighted information criteria. D-rule method is based on maximizing the determinant of the information matrix and T-rule based on maximizing the trace of the asymptotic covariance matrix. And other methods KL and Continious Entropy method, maximize posterior expected KL information and minimize the expected continuous entropy, respectively. And other methods KL and CL are based on maximizing the posterior expected KL information and minimizing the expected continuous entropy, respectively. (Veerkamp & Berger, 1997; Segall, 2001; Wang & Chang, 2004; Mulder & van der Linden, 2009; Wang & Chang 2011).

Estimation method: The estimation method for calculating the examinees' latent trait parameters should be selected. The Maximum Likelihood Method (MLE) (Lord and Novick, 1968), EAP and MAP (Segall, 1996) are the most frequently used methods. However, if all the answers are correct or incorrect, EAP and MAP methods are suggested to make estimations with low standard errors (Hambleton and Swaminathon, 1985). In addition to these methods, the weighted MLE (MWLE) method was revealed by Wang (2015) for multidimensional tests, which provides robust estimations

Stopping rule: At this stage of the CAT application, the stopping rule of testing should be determined. In CAT applications, stopping rules may be used in accordance with the fixed test length, standard error, change in the amount of the latent trait estimation, or the fixed application time. When the stopping rule is determined according to the fixed test length, erroneous results can be obtained in CAT applications for the examinees who are at the end of the skill distribution (Finkelman et al., 2009). For this reason, researchers can stop the CATs when a standard error level they previously specified is reached, and they obtain measurements that are more precisely. When specifying the standard error, if the test used is multidimensional, each dimension can be terminated with the same or different standard error values (Chalmers, 2016). If the standard error value is not determined for each dimension by customizing the codes to be used, the application stops in accordance with the standard error value specified for the first dimension and estimates according to the different standard error values for the other dimensions. Therefore, if the standard error-based stopping rule will be used in MCAT applications, it is essential to add the standard error value for each dimension to the code to be run.

1.2. Monte Carlo Simulations

Monte Carlo simulations have a crucial role in studies in the field of psychometrics. Within the scope of their studies, researchers may not be able to access empirical data or may not prefer to test applications for data collection purposes. One reason for the simulation requirement is that collecting empirical data can be time-consuming and costly when the number of items used in studies is long, and the number of examinees to be applied is high. In some studies, there are losses in the empirical data collected, and this loss of data affects the results of the analysis. Another, probably the most important, reason is that the working conditions to be examined cannot be obtained with real-time applications (Davey et al., 1997; Feinberg and Rubright, 2016). But this approach also has several limitations. Firstly, how realistic the conditions modeled in Monte Carlo simulation studies are affects the usefulness of the results. In this respect, the modeled conditions (e.g., assumed distribution of the parameters) should be defensible in terms of reality. Another limitation is that it is difficult to assess the quality of the random number generator in Monte Carlo simulations (Stone, 1993). Post-hoc and hybrid simulations can be done as a solution to the concern that the results obtained from Monte Carlo simulations cannot be generalized to real test applications. In post-hoc simulations, real item-response vectors obtained from paper-and-pencil test or adaptive test are used instead of generated item responses. In hybrid simulations, simulations are performed after missing value imputation based on empiricalreal data set (Thompson & Weiss, 2011). However, it is not easy to obtain real answers due to time and cost. Monte Carlo studies allow modeling realistic data conditions and can be used in competitor statistical comparisons that cannot be made with empirical data (Harvell et al., 1996). When there are a large number of conditions to manipulate, Monte Carlo simulations are preferred. Because, Monte Carlo simulations provide researchers with the opportunity to test a large number of models in a short time, which are hard to test in real life.

Monte Carlo simulation studies can be carried out by researchers in order to examine the applicability in CAT applications and to make an application plan. Post-hoc or hybrid simulation studies are preferred to determine the final application conditions (Thompson and Weiss, 2011). Monte Carlo and post-hoc simulations are frequently used in CAT applications performed in the literature. The two most crucial variables of Monte Carlo simulations are average test length and precision of the test scores. In traditional tests, the number of items in the test is constant, and the precision is variable. The number of items administered to examinees in adaptive tests is usually the variable, but it can be designed to provide equal precision to each examinee. In this regard, simulation studies are essential (Thompson & Weiss, 2011).

1.3. R Statistical Programming Environment

As a result of the development of computer technology, there are some commercial and open-source softwares that can carry out CAT simulations. Some of these softwares are CATSim, SimulCAT, SimuMCAT, Firestar software, and R software environment. CATSim is presented as a commercial product, and other software is presented as open-source access (Aybek, 2016). While unidimensional CAT simulations are possible with CATSim, SimulCAT, and Firestar software, MCATs simulations are possible with the SimuMCAT and the mirtCAT package in the R software environment.

The R programming language, which has been widely used in academic studies in recent years, is a programming language developed with the contributions of researchers from different parts of the world since 1997 (Hornik, 2020). The use of R has increased rapidly due to its open source code. R programming language offers the opportunity to be used in many fields such as statistics, data mining, machine learning and simulation applications. The R statistical programming environment (R Core Team, 2020) enables the opportunity to conduct simulation studies free of charge. Researchers who ask for generating data in R may generate data in accordance with a different probability distribution (normal, log-normal, uniform, etc.). There is a root name setting out each distribution, and usually, four functions are defined for each. Each distribution's commands begin with a letter to indicate functionality:

p: cumulative distribution function,

q: quantile function,

d: density function,

r: randomly generated numbers.

For instance, for log-normal distribution, rlnorm (the multivariate lognormal distribution), plnorm (the log normal cumulative distribution), dlnorm (the log normal probability density) and qlnorm (the log normal quantile) functions can be defined. Random data is generated for the rlnorm function according to the log-normal distribution. The qlnorm function sets the quantile of the log-normal distribution at a given cumulative density. Normal, log-normal, and uniform distributions are frequently used in studies where data generation is performed based upon IRT.

In this study, the steps of MCAT simulations according to within-item and between-item dimensionality models with the mirtCAT (version: 1.10) package in the RStudio (version: 1.3.1073) software environment will be demonstrated in terms of ease of use and prevalence.

After all required components are prepared, the function that starts MCAT simulations with the mirtCAT package is the mirtCAT() function. It is essential to introduce the item and individual parameters, IRT model, inter-dimensional correlations, starting rule of the test, item selection criteria and stopping rule to perform a multidimensional simulation with this function. The functions that are basically required to perform MCAT simulation with mirtCAT package are described in Table1, below (Chalmers, 2016).

Table 1. *Some functions to perform MCAT simulation with mirtCAT package*

mo: It is used in the model definition phase. The model defined in the mirt package is drawn to mirtCAT with this function. This object is required if test items are to be scored.
generate.mirt_object: It is the function used to form a mirt object from known population parameters and transfer it to mirtCAT.
method: It is used to determine the parameter estimation method. "EAP", "MAP", "ML", "WLE", "EAPsum", "fixed" are the methods that can be selected.
criteria: It is the function for determining the method of item selection. "seq", "random", "MI", "MEPV", "MLWI", "MPWI", "MEI", "IKL", "IKLP", "IKLn", "IKLPn", "Drule", "DPrule", "Erule", "EPrule", "Trule", "TPrule", "Arule", "APrule", "Wrule", "WPrule", "KL", "KLn".
start_item: It is the function by which the starting rule of MCAT application will be determined.

A MCAT design can be customized using different MIRT models, different item selection rules, different estimation methods etc. Since the methods to be used in a design will affect the measurement result, it is important to determine the most effective methods according to the application purpose. Besides these, interdimensional correlations and the dimensional structures are important issues for CATs (Su, 2016). Because interdimensional correlation can change the dimensionality of the structure and this in turn can change the MCAT implementation to be carried out.

1.4. Purposes

The purposes of this study are to present how MCAT designs can be generated and executed through Monte Carlo simulations in R environment; to show the effect of simulation conditions, which can be considered according to different dimensionality models, on the simulation results, and to investigate the effect of treating multidimensional structures as unidimensional structures. For the purposes, different Monte Carlo simulation studies were presented and the steps of simulations were demonstrated.

1.5. Research Questions

In line with the research purposes, answers to the following questions were sought:

1. How is an MCAT simulation designed according to the within-item dimensionality model affected by different item selection methods, ability parameter estimation methods and interdimensional correlations?
2. How is an MCAT simulation designed according to the between-item dimensionality model affected by different item selection methods, ability parameter estimation methods and interdimensional correlations?
3. How does treating each dimension of the multidimensional structure as a single dimension affect the simulation results?

2. METHOD

In this study, three different Monte Carlo simulation studies were carried out and all simulation steps were presented. In all three studies, data generation, Monte Carlo simulation steps and findings obtained as a result of analysis are presented. In order to answer the first research question, Simulation Study 1 is carried out. In this example, MCAT simulations were conducted according to the within-item dimensionality model with different conditions. Different conditions that could affect the precision of MCATs were considered: (1) interdimensional correlation, (2) item selection method, (3) parameter estimation method. In order to answer the second question, MCAT simulations conducted according to the between-item dimensionality model are carried out and the same conditions in the first study were examined in Simulation Study 2. Lastly, in order to seek an answer to the third questions, Simulation Study 3 is carried out. In this study, Unidimensional CAT (UCAT) simulations were conducted using the item and ability parameters of the multidimensional structure. MCAT and UCAT simulations performed with the data produced according to the between-item dimensionality model were compared.In all studies, RMSE, bias and $r(\theta i, \widehat{\theta} j)$ values obtained from all simulations were examined. R was used in order to complete simulation steps. The presented steps are completed in R for Windows 4.0.2. Simulation study examples of MCAT applications were performed on the mirtCAT (Chalmers, 2016) package.

2.1. Simulation Study 1: The Within-item Dimensionality Model

In the first simulation study, the within-item dimensionality model was handled. A simple non-customized MCAT simulation example is presented. Item selection methods, ability parameter estimation methods and the interdimensional correlations were examined as changing simulation conditions. In the first step of the MCAT simulation, packages to be used on the R platform should be downloaded.

```r
# Install required packages
install.packages("mirt")
install.packages("mirtCAT")
install.packages("mvtnorm")
install.packages("plyr")
install.packages("SimDesign")
```

After the downloading process is completed, the required packages should be activated. Before the analyses are carried out, the `set.seed()` command ensures that the outputs of the application are reproducible. Any number can be written in parentheses, and the same results are obtained when `set.seed()` is run with the same number.

```r
# Load packages into the current session
library(mirt)
library(mirtCAT)
library(mvtnorm)
library(plyr)
library(SimDesign)
# Set the seed for reproducible results
set.seed(1111)
```

After the packages were drawn and activated, item and ability parameters were generated. In accordance with the within-item dimensionality model, 2-dimensional MCAT simulations were carried out for 1000 examinees. For this example, parameters for a multidimensional test consisting of 300 dichotomous items and 2 dimensions were generated. The *a* parameters were drawn from the log normal distribution (*a* ~ *lnN* (.0, .3)), and item intercept parameters were drawn from the uniform distribution (*d~ U* (-2, 2)). After slopes and intercepts were generated, they were combined in a single object (parameters) with the `data.frame` function.

```
#Generate Multidimensional IRT parameters
testlength <- 300 # Bank size
N <- 1000 # Sample size
a <- matrix(rlnorm(testlength*2,.0,.3),testlength) # Generate item
discrimination parameters
d <- runif(n = testlength, -2, 2) # Generate intercepts
parameters <- data.frame(a, d) # Combine parameters in a single dataset
colnames(parameters) <- c('a1', 'a2', 'd') # Name the columns
```

In the next step, the variance-covariance matrix of the ability parameters for two-dimensional structure is demonstrated on a matrix. Ability parameters were drawn from the multivariate normal distribution ((θ~$(0,\boldsymbol{\Sigma})$) depending on the defined correlations. For the two-dimensional structure, the inter-dimensional correlation was determined as 0.3, 0.6 and 0.9, and parameters generated with `rmvnorm()` function.

```
#Set intercorrelations between latent traits
latent_cov <- matrix(c(1, r, r, 1), 2, 2)

#Generate multidimensional theta parameters
thetas <- rmvnorm(N, sigma = latent_cov)
```

Then the `mod` object required for MCAT Simulation was formed. This object is used while generating the response pattern and creating the MCAT design. The `generate_pattern ()` function was used to generate the response pattern.

```
# Create mirt_object
mod <- generate.mirt_object(parameters, itemtype = '2PL', latent_covariance
= latent_cov)

#Generate response data
responsepattern <- generate_pattern(mo = mod, Theta = thetas)
```

In the next step, required components were specified for the MCAT simulation to be conducted. These components were defined by the function `design()` and `mirtCAT()`. In these definitions, SE ≤0.4 was specified as stopping rule. Five item selection methods were examined for two dimensional complex model using two estimation methods. By using Arule, Drule, Trule, Wrule and KLn (Kullback-Leibler item selection method with root-N adjustment) item selection methods, estimations were made according to both EAP and MAP methods. The item starting rule for each condition is the same as for the item selection method. 5x2x3 (30) simulation application including stopping criterion, estimation method and correlation was carried out.

```
# Run the MCAT simulations with mirtCAT function and store results
design <- list(min_SEM = 0.4)
mcat1 <- mirtCAT(mo = mod, local_pattern = responsepattern, method = ' ',
start_item = " ", criteria = " ", design = design)
```

When the application is completed, the results can be reached by running the `mcat1` object where the results are saved. The mirtCAT package presents the avarage number of items administered, the ability parameter and true ability parameter for each dimension, and the standard errors of these parameters in the output of the simulation. If researchers ask for examining the test efficiency or the effect of different MCAT simulation designs on test efficiency; bias and

RMSE values can be computed. In this situation, firstly, an object should be formed in which theta estimates of both dimensions are listed. Theta estimates are collected in a single object with the `laply()` function that can be computed using the `bias()` and `RMSE()` commands of the SimDesign (Chalmers et al., 2020) package, with the following code:

```
#Show average number of items answered, theta estimations, bias and RMSE
itemsanswered <- laply(mcat1, function(x) length(x$items_answered))
mean(itemsanswered)
estimation1 <- laply(mcat1, function(x) x$thetas[1])
estimation2 <- laply(mcat1, function(x) x$thetas[2])
bias(thetas[,1], estimation1) # Compute bias
bias(thetas[,2], estimation2)
RMSE(thetas[,1], estimation1) # Compute root mean square error
RMSE(thetas[,2], estimation2)
cor(thetas[,1], estimation1)
cor(thetas[,2], estimation2)
```

According to the simulation outputs, the average number of items answered, bias, RMSE and the correlation between estimated θ and true $\hat{\theta}$ (r(θ i, $\hat{\theta}$ j)) obtained for both dimensions are presented. The conditions with interdimensional correlations of 0.3, 0.6, and 0.9 are presented in Table 2, Table 3 and Table 4, respectively.

Table 2. *Statistics from MCAT when interdimensional correlation is 0.3*

Item Selection	Estimation	MTL	Bias1	Bias2	RMSE1	RMSE2	r_1(θi, $\hat{\theta}$ j)	r_2(θi, $\hat{\theta}$ j)
Arule	EAP	67.057	-0.004	-0.013	0.374	0.398	0.926	0.917
	MAP	63.848	-0.008	-0.013	0.402	0.373	0.926	0.915
Drule	EAP	72.517	-0.02	0.002	0.373	0.398	0.927	0.917
	MAP	69.56	-0.05	-0.001	0.376	0.403	0.925	0.915
Trule	EAP	94.364	-0.017	0.000	0.355	0.396	0.934	0.917
	MAP	91.238	-0.018	0.001	0.359	0.399	0.932	0.916
Wrule	EAP	100.424	-0.016	0.004	0.357	0.392	0.933	0.919
	MAP	97.532	-0.016	0.003	0.359	0.395	0.932	0.918
KLn	EAP	103.496	-0.017	0.009	0.359	0.393	0.932	0.919
	MAP	100.6	-0.02	0.005	0.366	0.397	0.929	0.917

Note. MTL (Mean test length) represents the average number of items administered.

When the correlation between dimensions is 0.3, the MCAT application resulting in the least average number of items was performed with the MAP estimation method and the ARule stopping rule. However, when the Arule stopping rule was used, the RMSE value obtained for the first dimension was not below 0.40. The simulation application that resulted in the highest number of items was carried out with the KLn method. All MCAT applications ended with fewer items with the MAP estimation method. Correlation between estimated θ and true $\hat{\theta}$ calculations was high and similar in all conditions. All calculated bias values were negligible.

As seen in Table 3, the increase in interdimensional correlation decreased the number of items required to terminate the MCAT application. The simulation that resulted in the least number of items was carried out with the Arule stopping rule and the MAP estimation method. All applications performed with the MAP estimation method ended with fewer items than the applications performed with EAP. The RMSE value for the second dimension was not below 0.4

for Arule, Drule, Trule and Wrule methods. Only, the RMSE value obtained for both dimensions with the KLn method fell below 0.4. It should be noted that the KLn is the only method that can be used in common in unidimensional and multidimensional CAT applications. Correlation between estimated θ and true $\widehat{\theta}$ calculations and bias values were similar. The results obtained under the condition that the interdimensional correlation is 0.9 are presented in Table 4.

Table 3. *Statistics from MCAT when interdimensional correlation is 0.6*

Item Selection	Estimation	MTL	Bias1	Bias2	RMSE1	RMSE2	$r_1(\theta i, \widehat{\theta} j)$	$r_2(\theta i, \widehat{\theta} j)$
Arule	EAP	45.006	0.025	-0.024	0.375	0.41	0.928	0.918
	MAP	42.485	0.024	-0.023	0.385	0.414	0.924	0.916
Drule	EAP	46.05	0.024	-0.022	0.368	0.412	0.931	0.917
	MAP	43.839	0.017	-0.025	0.374	0.420	0.928	0.914
Trule	EAP	66.181	0.027	0.027	0.354	0.401	0.936	0.921
	MAP	58.440	0.026	-0.011	0.357	0.402	0.935	0.921
Wrule	EAP	64.706	0.023	-0.010	0.354	0.398	0.936	0.922
	MAP	63.014	0.023	-0.012	0.355	0.400	0.936	0.922
KLn	EAP	66.333	0.028	-0.016	0.352	0.393	0.937	0.925
	MAP	64.840	0.024	-0.019	0.352	0.396	0.937	0.924

Table 4. *Statistics from MCAT when interdimensional correlation is 0.9*

Item Selection	Estimation	MTL	Bias1	Bias2	RMSE1	RMSE2	$r_1(\theta i, \widehat{\theta} j)$	$r_2(\theta i, \widehat{\theta} j)$
Arule	EAP	12.832	0.019	0.009	0.375	0.386	0.932	0.926
	MAP	11.119	0.028	0.013	0.402	0.406	0.922	0.918
Drule	EAP	11.216	0.015	0.004	0.370	0.386	0.934	0.926
	MAP	9.820	0.022	0.01	0.390	0.401	0.927	0.92
Trule	EAP	10.849	0.022	0.008	0.356	0.380	0.940	0.928
	MAP	9.219	-0.018	0.035	0.383	0.398	0.931	0.922
Wrule	EAP	10.401	0.019	0.010	0.371	0.386	0.934	0.926
	MAP	8.975	0.031	0.022	0.388	0.388	0.929	0.926
KLn	EAP	10.84	0.026	0.018	0.367	0.376	0.935	0.930
	MAP	9.645	0.022	0.014	0.377	0.397	0.933	0.922

When the interdimensional correlation was 0.9, the number of items required to complete the MCAT simulation was greatly reduced. The result of the application that ends with the least number of items was obtained with the Trule stopping criterion and the MAP estimation method. When Arule and Drule stopping methods are used with MAP parameter estimation method, the RMSE value for the second dimension was not below 0.4. In line with the simulation results, it was observed that the Arule and Drule methods gave similar results in all conditions. However, since they finished the application with fewer items, it was observed that although they provided the desired RMSE value in the first dimension, they could not provide in the second dimension. Simulations performed using the MAP estimation method in all conditions resulted in fewer items than EAP. As the interdimensional correlation increased and the structure approached unidimensionality, the methods gave results closer to each other and in all conditions KLn provided the desired stopping rule for both dimensions. In all conditions, $r(\theta i, \widehat{\theta} j)$ obtained for both dimensions was high and similar. Bias for all dimensions was negligible.

2.2. Simulation Study 2: The Between-item Dimensionality Model

In simulation study 1, MCAT simulations were performed with the stopping rule not customized. In simulation study 2, MCAT simulations were performed according to the between item dimensionality model by customizing the stopping rule for each dimension. Interdimensional correlation values and stopping methods were used the same as Study 1. Due to the fact that the necessary packages are loaded in the first example, the packages will be activated directly in this example. The packages required for this study are called via commands written to the console. In order to MCAT Simulation results to be reproducible, the `set.seed()` command is used.

```
library(mirt)
library(mirtCAT)
library(mvtnorm)
library(plyr)
library(SimDesign)
set.seed(2222)
```

According to the between-item dimensionality model for the MCAT simulation, parameters for a multidimensional test consisting of 600 polytomous items and 2 dimensions were generated. Item parameters were generated for polytomous items with four categories and Multidimensional Graded Response Model (MGRM) was chosen as the MIRT model. The item parameters are distributed in the same way as in the study of Jiang, Wang, and Weiss (2016). The *a* parameters were drawn from the uniform normal distribution ($a \sim U\,(1.1,\, 2.8)$). First category boundary parameter (d_1) were drawn randomly from the uniform distribution (d_1~ (0.67, 2)), second category boundary parameter from (d_1~ (-0.67, 2-0.67)) and third category boundary parameter from (d_1~ (-0.67, -2)). Thus, all item bounce parameters ranged from [-2,2]. After that, the generated parameters were combined in a single dataset.

```
Generate Multidimensional IRT parameters
testlength <- 600 # Bank size
N <- 1000 # Sample size
# Generate  parameters
itemnames <- paste0("Item.", 1: testlength)
a <- matrix(runif(testlength *2, 1.1, 2.8), testlength)
a[1:300, 2] <- a[301:600, 1] <- 0
d1 <- runif(n = 600, min = 0.67, max = 2) # Generate first category boundary
parameter
d2 <- d1 - runif(n = 600, min = 0.67, max = 1.34)
d3 <- d2 - runif(n = 600, min = 0.67, max = 1.34)
d <- as.matrix(cbind(d1, d2, d3), ncol = 3)
parameters <- data.frame(a, d) # Combine parameters in a single dataset
colnames(parameters) <- c('a1', 'a2', paste0('d', 1:3))
```

In the next step, the variance-covariance matrix of the ability parameters for the two-dimensional structure is demonstrated on a matrix (cov). For the two-dimensional structure, interdimensional correlations were determined as 0.3, 0.6 and 0.9 between all dimensions. Ability parameters were drawn from the multivariate normal distribution ($\theta \sim (0, \boldsymbol{\Sigma})$) depending on the defined correlations. Then the mod object was created and the response pattern was generated using this object.

```
#Set intercorrelations between latent traits
latent_cov <- matrix(c(1, r, r, 1), 2, 2)
#Generate theta parameters for 2 dimensions
thetas <- rmvnorm(N, sigma = latent_cov)
#Create mirt_object
mod <- generate.mirt_object(parameters, itemtype = 'graded', latent_covari-
ance = cov)
#Generate response pattern
responsepattern <- generate_pattern(mo = mod, Theta = thetas)
```

In the next stage, unlike the first example, the minimum SE values were determined for each dimension by customizing the commands. The simulation was stopped on the condition that each dimension had a minimum SE value below 0.4 using `customNextItem()`, `extract.mirtCAT()` and `findNextItem ()` functions. In this regard, each dimension is considered as a block.

As item selection criteria, Arule, Drule, Trule, Wrule and KLn methods were used. EAP and MAP estimation methods were used as in the first example. The stopping rules of the application was determined by the `customNextItem()` function, the item selection method is defined by the `findNextItem()` function. A total of 30 simulations including the stopping rule (5), the estimation method (2) and the correlation value (3) were carried out.

```
customNextItem <- function(design, person, test){
browser()
      }
customNextItem <- function(design, person, test){
block1 <- 1:300
block2 <- 301:600
#Stop when the SE value falls below 0.4.
total <- sum(!is.na(extract.mirtCAT(person, 'items_answered')))
if(total< 300 && extract.mirtCAT(person, 'thetas_SE')[1] >= 0.4){
block <- block1
} else if(total < 600 && extract.mirtCAT(person, 'thetas_SE')[2] >= 0.4){
block <- block2
} else return(NA)
ret <- findNextItem(person=person, design=design, test=test, subset=block,
criteria = '')
ret
}
```

In the last step, simulation design was constructed with `mirtCAT()` function. Average number of items answered, bias, RMSE and the correlation between estimated θ and true $\widehat{\theta}$ values calculated with the commands presented below.

```
mcat2 <- mirtCAT(mo = mod, local_pattern = responsepattern, method = ' ',
start_item = " ", criteria = " ",
design = list(customNextItem=customNextItem))

#Show average number of items answered, bias, RMSE and r(θ i, ,θ .j).
itemsanswered <- laply(mcat2, function(x) length(x$items_answered))
mean(itemsanswered)
estimation1 <- laply(mcat2, function(x) x$thetas[1])
```

```
estimation2 <- laply(mcat2, function(x) x$thetas[2])
bias(thetas[,1], estimation1) # Compute bias
bias(thetas[,2], estimation2)
RMSE(thetas[,1], estimation1) # Compute root mean square error
RMSE(thetas[,2], estimation2)
cor(thetas[,1], estimation1)
cor(thetas[,1], estimation1)
```

The average number of items administered, bias, RMSE and the correlation between estimated θ and true $\hat{\theta}$ ($r(\theta$ i, $\hat{\theta}$ j)) obtained for both dimensions are as follows. The values calculated when the correlation is 0.3 are presented in Table 5.

Table 5. *Statistics from MCAT when interdimensional correlation is 0.3*

Item Selection	Estimation	MTL	Bias1	Bias2	RMSE1	RMSE2	$r_1(\theta i, \hat{\theta} j)$	$r_2(\theta i, \hat{\theta} j)$
Arule	EAP	12.035	-0.02	0.011	0.359	0.356	0.938	0.934
	MAP	9.986	-0.013	0.014	0.388	0.397	0.928	0.918
Drule	EAP	16.27	0.001	0.002	0.386	0.353	0.927	0.935
	MAP	13.887	0.022	0.014	0.403	0.398	0.922	0.918
Trule	EAP	12.035	-0.02	0.011	0.359	0.356	0.938	0.934
	MAP	9.986	-0.013	0.013	0.388	0.397	0.928	0.918
Wrule	EAP	12.035	-0.020	0.011	0.359	0.356	0.938	0.934
	MAP	9.986	-0.013	0.014	0.388	0.397	0.928	0.918
KLn	EAP	12.101	0.006	0.010	0.370	0.354	0.933	0.935
	MAP	10.240	0.006	0.002	0.390	0.398	0.927	0.918

When the interdimensional correlation for the between-item dimensionality model is 0.3, the number of items required to finish the simulation is similar for the conditions. However, when the Drule method was used as stopping rule, average number of items administered were higher compared to other methods. The condition with the highest number of items administered is the condition in which the Drule stopping rule and EAP estimation method are used. Under the condition that the Drule stopping rule and EAP estimation method are used, the RMSE value obtained for the first dimension is more than 0.4. The results obtained from simulations using the Trule and Wrule stopping rules are the same. All simulations ended with fewer items with the MAP estimation method. The calculated bias and $r_1(\theta$ i, $\hat{\theta}$ j) values are similar for all conditions. The values calculated for the condition that the interdimensional correlation is 0.6 are presented in Table 6.

Table 6. *Statistics from MCAT when interdimensional correlation is 0.6*

Item Selection	Estimation	MTL	Bias1	Bias2	RMSE1	RMSE2	$r_1(\theta i, \hat{\theta} j)$	$r_2(\theta i, \hat{\theta} j)$
Arule	EAP	11.52	-0.012	0.002	0.35	0.355	0.94	0.934
	MAP	9.661	-0.007	0.005	0.375	0.396	0.931	0.919
Drule	EAP	15.686	-0.003	-0.009	0.374	0.356	0.931	0.934
	MAP	13.643	0.018	0.009	0.389	0.394	0.926	0.919
Trule	EAP	11.52	-0.012	0.002	0.350	0.355	0.940	0.934
	MAP	9.661	-0.007	0.005	0.375	0.396	0.931	0.919
Wrule	EAP	11.52	-0.012	0.002	0.350	0.355	0.94	0.934
	MAP	9.661	-0.007	0.005	0.375	0.396	0.931	0.919
KLn	EAP	11.648	-0.01	-0.008	0.349	0.367	0.940	0.929
	MAP	10.015	0.003	-0.002	0.373	0.399	0.932	0.917

Under the condition that the interdimensional correlation is 0.6, the average number of items administered is fewer than the correlation is 0.3. The average number of items obtained from the simulation application performed with the Drule stopping rule is higher than other methods. The same results were obtained with Arule, Trule and Wrule methods. Bias and $r_1(\theta\ i, \hat{\theta}\ j)$ values are very close to each other for all conditions.

Finally, for the between-item dimensionality model, the values calculated according to the condition that the interdimensional correlation is 0.9 are presented in the Table 7.

Table 7. *Statistics from MCAT when interdimensional correlation is 0.9*

Item Selection	Estimation	MTL	Bias1	Bias2	RMSE1	RMSE2	$r_1(\theta i, \hat{\theta} j)$	$r_2(\theta i, \hat{\theta} j)$
Arule	EAP	9.302	0.002	-0.004	0.316	0.363	0.950	0.931
	MAP	7.793	0.010	0.006	0.354	0.390	0.938	0.922
Drule	EAP	13.283	-0.002	-0.001	0.342	0.367	0.941	0.930
	MAP	11.585	0.018	0.013	0.361	0.398	0.935	0.918
Trule	EAP	9.302	0.002	-0.004	0.316	0.363	0.950	0.931
	MAP	7.793	0.010	0.006	0.354	0.390	0.938	0.921
Wrule	EAP	9.302	0.002	-0.004	0.316	0.363	0.950	0.931
	MAP	7.793	0.010	0.006	0.354	0.390	0.938	0.921
KLn	EAP	9.346	0.002	-0.003	0.329	0.36	0.946	0.933
	MAP	7.885	0.015	0.013	0.358	0.394	0.937	0.920

When the interdimensional correlation is 0.9, that is, if the structure is similar to unidimensional structure, the average number of items administered is the fewest. As in other conditions, simulations performed with the MAP estimation method resulted fewer items than simulations performed with EAP method. The calculations obtained using the stopping rules Arule, Wrule and Trule are the same. Simulation with Drule method ended with more items and higher RMSE values than others. The bias values calculated for both dimensions are negligible.

2.3. Simulation Study 3: Comparison of MCAT and CAT Results

In the third simulation study presented , we investigate the effect of treating multi-unidimensional structures as unidimensional structures on adaptive testing results. In line with the purpose, using the item and ability parameters used in the second example, a unidimensional CAT simulation was performed and the outputs were compared with the MCAT simulation. Since it is an item selection method that can be used in both CATs and MCATs, the "KLn" method was used. MAP was used as the estimation method. The data generated for the two-dimensional structure is exported in csv format with the Haven package (Wickham & Miller, 2020). After obtaining the item and ability parameters with the commands example 2, the parameters were exported through the following commands, the ".csv" files were divided and saved for each dimension.

```
#Export parameters
library(haven)
df <- data.frame(a1 = a[,1], a2= a[,2], d1 = d[,1], d2 = d[,2], d3 = d[,3])
write.csv(df, "parameters.csv")
write.csv(thetas, "thetas.csv")
```

After the data sets were saved separately for each dimension, simulation studies continued with ".csv" files. In the UCAT simulation phase, SE($\hat{\theta}$) <0.4 stopping criteria is determined as in the MCAT examples.

```
#Design and start simulation
design = list(min_SEM = 0.4, max_items=300)
mcat3 <- mirtCAT(mo=mod, local_pattern=response, start_item = 'KLn', criteria
= 'KLn', design = design)
```

Average number of items administered, bias, RMSE and r (θ i, $\hat{\theta}$ j) values obtained from CAT simulation performed with MCAT parameters are presented in the Table 8.

Table 8. *Statistics from UCAT simulation.*

Dimension	Interdimensional Correlation	MTL	Bias	RMSE	r (θ i, $\hat{\theta}$ j)
1	0.3	5.335	0.003	0.398	0.924
2		5.130	-0.008	0.407	0.914
1	0.6	5.142	-0.004	0.412	0.917
2		5.120	0.004	0.412	0.911
1	0.9	5.326	-0.006	0.393	0.923
2		5.120	-0.005	0.401	0.917

Compared to MCAT and unidimensional CAT in terms of the average number of items administered, MCAT has a lower average number of items in all conditions. As the interdimensional correlation increases, the average number of items decreases. Unidimensional CAT simulation, on the other hand, resulted in a similar number of items in all conditions. In addition, in MCATs, SE ≤ 0.4 criterion was provided for both dimensions, whereas in CATs, this criterion was only provided for the first dimension when the correlation was 0.3 and 0.9. As in MCAT simulations, bias values are negligible in UCAT.

3. DISCUSSION and CONCLUSION

Since CATs are used for selection, classification and diagnosing purposes, it has important functions for society (Chang, 2015). Technological developments have increased the popularity of CAT applications. With CATs, test length and test session duration are reduced compared to the paper-pencil applications of both achievement tests and psychological scales. While this decrease, the increase in measurement precision makes adaptive testing applications more important. Through the widespread use of MIRT models, MCAT applications are becoming widespread. Researchers frequently apply simulations before CAT and MCAT applications to design the appropriate design for their studies. In this study, data were generated using Monte Carlo simulations by using within-item and between-item dimensionality models. With the generated data, MCAT simulation application codes customized according to different conditions were presented. R programming language was used in this study as it is an open-source and free software. The simulation findings obtained under different conditions are shared. The average number of items administered, RMSE, BIAS and r (θ i, $\hat{\theta}$ j) values obtained using different interdimensional correlation values, different item selection criteria and different parameter estimation methods were examined.

3.1. Main Findings

In this study, the steps of MCAT simulations according to within-item and between-item dimensionality models with the mirtCAT (version: 1.10) package in the RStudio (version: 1.3.1073) software environment were demonstrated. In more detail, multidimensional models applied at the item level to MCAT under within-item and between-item dimensional models using three interdimensional correlation levels, five item selection methods and two parameter estimation methods. MCAT and CAT results performed with data generated according to the between-item dimensionality model were compared. Results showed that MCAT simulations performed with data produced according to the multidimensional models, as the interdimensional correlation increased, the average number of items required to terminate the test decreased. In the MCAT simulations performed according to the within-item dimensionality model, the number of items required to complete the test was higher than the between-item dimensionality model. While increasing the correlation in the within-item dimensionality model greatly changes the average number of items, the average number of items is quite similar in the between-item model.

Wang and Chen (2004) concluded in their study that the higher the correlation between the traits, the less number of items required to reach the same test reliability degree and MCATs will be more efficient than CATs. Similarly, in this study, as the correlation between features increased, the number of items required to complete the MCAT according to the standard error rule decreased. In other words, as the correlation value between traits increases, the number of items required to achieve similar accuracy decreases. And, MCAT was more effective than CAT at meeting the required termination criteria.

When comparing UCATs and MCATs with data generated according to MIRT, the average number of items used in UCAT simulation is higher than MCAT. According to MCAT results, SE <0.4 rule was provided for each dimension, but according to UCAT results, this rule was not provided for all dimensions. A similar result was obtained in Paap, Born, and Braeken's (2018) study. They conducted simulations with the standard error-based termination rule for different design cells and concluded that while meeting the MCAT's termination criteria, CAT failed 80% to meet the termination criterion.

According to the findings obtained, ability parameter estimation method, interdimensional correlations and item selection methods did not much affect measurement fidelity. However, as in the Yao's (2013) study, it can be said that MAP performs similar or better than EAP. The bias values obtained for the different conditions indicate that MCATs give unbiased estimates of ability. The size of interdimensional correlation, item selection criterion and parameter estimation method did not have a considerable effect on the calculated BIAS values for three examples. An interesting finding obtained as a result of the simulations was that KLn was the only method that provided the standard error stopping criteria, regardless of the methods used.

3.2. Future Directions

On the basis of results, for MCAT simulations that researchers will design according to the standard error-based stop rule, it is suggested to add MAP as the estimation method to the simulation conditions. If the standard error rule cannot be defined separately for each dimension, it is recommended to add the KLn rule as the item selection criterion to the simulation conditions. It should be noted that if the stopping rule is not defined for each dimension in MCAT applications, the standard error-based stop rule may not be provided. If the stopping rule is defined by customizing for each dimension, the items continues to be applied until the termination rule is met in all dimensions. Therefore, it is required to specify the standard error rule by customizing it at the desired level for each dimension. If the structure is multidimensional, it is recommended to use MCAT instead of applying separate CAT to each dimension.

Lastly, although the number of MCAT studies has increased in the last decade, more research is needed to investigate scenarios beyond the factors included in the study. For example, different stopping rules, content balancing and other MIRT models can also be investigated. It is important for MCAT practitioners to know with which criteria they can perform MCAT applications more efficiently and effectively.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number:** Since the data in this study were generated by Monte Carlo simulations, there is no need for an ethics committee document.

Authorship Contribution Statement

F. Gul Ince Araci: Investigation, Methodology, Resources, Visualization, Software, Formal Analysis, and Writing the original draft. **Seref Tan:** Methodology, Supervision, and Validation.

Orcid

F. Gul Ince Araci https://orcid.org/0000-0001-5620-6911
Seref Tan https://orcid.org/0000-0002-9892-3369

REFERENCES

Ackerman, T.A. (1991). The use of unidimensional parameter estimates of multidimensional items in adaptive testing. *Applied Psychological Measurement*, *15*(1), 13-24.

Aybek, E.C. (2016*). Kendini Değerlendirme Envanteri'nin bilgisayar ortamında bireye uyarlanmış test (BOBUT) olarak uygulanabilirliğinin araştırılması [An investigation of applicability of the self assessment inventory as a computerized adaptive test (CAT)]* [Doctoral Dissertation, Ankara University]. https://dspace.ankara.edu.tr/xmlui/bitstream/handle/20.500.12575/37233/eren_can_aybek.pdf?sequence=1&isAllowed=y

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algortihm. *Pschometrika, 46*(4), 443-459.

Bulut, O., & Sünbül, Ö. (2017). Monte carlo simulation studies in item response theory with the R programming language. *Journal of Measurement and Evaluation in Education and Psychology*, *8*(3), 266-287. https://doi.org/10.21031/epod.305821

Boyd, A.M., Dodd, B.G., & Choi, S.W. (2010). *Polytomous models in computerized adaptive testing.* In M. L. Nering & R. Ostini (Eds.), Handbook of polytomous item response theory models (pp. 229–255). Routledge.

Chalmers, R.P. (2015). mirtCAT: Computerized adaptive testing with multidimensional item response theory. *R package version 0.6*, *1*. https://CRAN.Rproject.org/package=mirtCAT

Chalmers, R.P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software*, *71*(5), 139. https://doi.org/10.18637/jss.v071.i05

Chalmers, P., Sigal, M., Oguzhan, O., & Chalmers, M. P. (2020). SimDesign: Structure fororganizing monte carlo simulation designs. *R package version 2.2.* https://CRAN.R-project.org/package=SimDesign

Chen, J. (2012*). Applying Item Response Theory methods to design a learning progression based science assessment* [Unpublished Doctoral Dissertation]. Michigan State University.

Davey, T., Nering, M. L., & Thompson, T. (1997). *Realistic simulation of item response data* (Vol. 97, No. 4). ACT, Incorporated.

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

Embretson, S.E., & Reise, S.P. (2000*). Item response theory for psychologists*. Erlbaum.

Feinberg, R.A., & Rubright, J.D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, *35*(2), 36-49. https://doi.org/10.1111/emip.12111

Finkelman, M., Nering, M.L., & Roussos, L.A. (2009). A conditional exposure control method for multidimensional adaptive testing. *Journal of Educational Measurement*, *46*(1), 84103. http://doi.org/10.1111/j.1745-3984.2009.01070.x

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Academic Publishers.

He, W., & Reckase, M.D. (2014). Item pool design for an operational variable-length computerized adaptive test. *Educational and Psychological Measurement*, *74*(3), 473-494. https://doi.org/10.1177/0013164413509629

Hornik, K., & FAQ, R. (2010). Frequently asked questions on R. *The R project for Statistical*. https://CRAN.R-project.org/doc/FAQ/RFAQ.html

Lin, H. (2012). *Item selection methods in multidimensional computerized adaptive testing adopting polytomously-scored items under multidımensional generalized partial credit model* [Unpublished Doctoral Dissertation, University of Illinois]. https://hdl.handle.net/2142/34534

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of zihinsel test scores*. Oxford.

Magis, D., Yan, D., & von-Davier, A. (Eds.). (2017). *Computerized adaptive and multistage testing with R: Using packages catr and mstr*. Springer.

Meneghetti, D.D.R., & Junior, P.T.A. (2017). *Application and simulation of computerized adaptive tests through the package catsim*. https:// arxiv.org/pdf/1707.03012.pdf

Mulder, J., & van der Linden, W.J. (2009). Muldimensional adaptive testing with optimal design criteria for item selection. *Psychometrika, 74*(2), 273-296. https://doi.org/10.1007/s11336-008-9097-5

Nydick, S., & Weiss, D.J. (2009). A hybrid simulation procedure for developments of CATs. *In Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* https://www.iacat.org/sites/default/files/biblio/cat09nydick.pdf

Paap, M.C., Born, S., & Braeken, J. (2019). Measurement efficiency for fixed-precision multidimensional computerized adaptive tests: Comparing health measurement and educational testing using example banks. *Applied Psychological Measurement*, *43*(1), 68-83. https://doi.org/10.1177/0146621618765719

R Core Team (2020). R: *A language and environment for statistical computing* [Computer software manual]. *http://www.R-project.org/*

Reckase, M.D. (2009*). Multidimensional item response theory: Statistics for social and behavioral sciences*. Springer.

Riggelsen, C. (2008). Learning Bayesian networks: a MAP criterion for joint selection of model structure and parameter. In *2008 Eighth IEEE International Conference on Data Mining* (pp. 522-529). IEEE.

Segall, D.O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*(2), 331-354.

Segall, D.O. (2001). General ability measurement: An application of multidimensional itemresponse theory. *Psychometrika*, *66*, 79-97.

Segall, D.O. (2005). *Computerized adaptive testing.* In K. Kempf-Leonard (Ed.),Encyclopedia of Social Measurement. Academic Press.

Seo, D.G., & Weiss, D.J. (2015). Best Design for Multidimensional Adaptive Testing With the Bifactor Model. *Educational and Psychological Measurement, 75*(6), *954-978.*

Su, Y.H. (2016). A comparison of constrained item selection methods in multidimensional computerized adaptive testing. *Applied psychological measurement*, *40*(5), 346-360. https://doi.org/10.1177/0146621616639305

Team, R. (2020). RStudio: Integrated Development for R (1.3.1073) [Computer software]. RStudio. https://rstudio.com/products/rstudio/

Thissen, D., & Mislevy, R.J., 2000. Testing algorithms. In H. Wainer (Eds.). *Computerized Adaptive Testing*. Lawrence Erlbaum Assc.

Thompson, N.A. (2007). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment, Research & Evaluation, 12*(1), 1-13.

Thompson, N.A., & Weiss, D.J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation. 16*(1*). 1-9.*

Van der Linden, W., & Glas, G. A. W. (2002). *Computerized adaptive testing: theory and practice.* Kluwer Academic Publishers.

Veerkamp, W.J., & Berger, M.P. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics*, *22*(2), 203-226. https://doi.org/10.3102/10769986022002203

Wainer, H., Dorans, N.J., Flaugher, R., Green, B.F., Mislevy, R.J. Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: a primer*. Lawrence Erlbaum.

Wang, C., & Chen, H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement,* 28*, 295-316.* https://doi.org/10.1177/0146621604265938

Wang, C. (2015). On latent trait estimation in multidimensional compensatory item response models. *Psychometrika*, *80*(2), 428-449. https://doi.org/10.1007/s11336-013-9399-0

Weiss, D.J., & Kingsbury, G.G. (1984). Application of computer adaptive testing to educational problems. *Journal of Educational Measurement, 21(4), 361–375.*

Weiss, D.J., & Gibbons, R.D. (2007). Computerized adaptive testing with the bifactor model. *Paper presented at the New CAT Models session at the 2007 GMAC Conference on Computerized Adaptive Testing.* https://mail.iacat.org/sites/default/files/biblio/cat07weiss%26gibbons.pdf

Yao, L. (2013). Comparing the performance of five multidimensional CAT selection procedure with different stopping rules. *Applied Psychological Measurement*, *37*(1), 3-23. https://doi.org/10.1177/0146621612455687

International Journal of Assessment Tools in Education

2022, Vol. 9, No. 1, 138–164

https://doi.org/10.21449/ijate.946530

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

Academic Intellectual Capital Scale: A Validity and Reliability Study

Ugur Ozalp[1,*], **Munevver Cetin**[2]

[1]Marmara University, Faculty of Education, Department of Educational Sciences, İstanbul / Turkiye

ARTICLE HISTORY

Received: June 01, 2021

Revised: Dec. 24, 2021

Accepted: Jan. 04, 2022

Keywords:

Academic intellectual capital,
Intellectual capital,
Higher education.

Abstract: The aim of this study was to develop a scale instrument for measuring academic intellectual capital in the Turkish higher education context depending on student perceptions. The sample consisted of students of higher education institutions in the 2020-2021 academic year. Data were gathered in two stages. Exploratory Factor Analysis (EFA) was conducted in the first stage and Confirmatory Factor Analysis (CFA) was conducted in the second stage. The EFA sample consisted of 538 students studying in 96 higher education institutions while the CFA sample consisted of 492 students studying in 112 higher education institutions. Principal Axis Factoring (PAF) extraction and Promax rotation methods were used in EFA. Results of EFA showed that the scale had a three-factor structure with 20 items. The three-factor structure was confirmed with CFA. Cronbach's alpha, stratified alpha, Composite Reliability and McDonald's omega were calculated in order to determine the reliability of the scores obtained from the scale. Item discrimination was verified by calculating item-total correlation and item-remainder correlation. Also, t-test was carried out between upper and lower 27% to check item discrimination. Analyses were conducted making use of R (ver. 4.1.2) and RStudio (ver. 2021.09.1 build 372). Overall, results showed that the structure of Academic Intellectual Capital Scale was valid. The measurement tool was concluded to have three factors and 20 items, all in affirmative form.

1. INTRODUCTION

The ever-changing nature of competition makes it obligatory for organizations to step ahead of their competitors in the context of meeting the expectations of the stakeholders. Representing the future-facing side of the societies and serving as a bridge between theory and practice, higher education institutions are also affected by this competitive environment. Academic intellectual capital of higher education institutions is among the variables that are effective in making difference in the competition.

Together with technological, economic, social and political innovations, intellectual capital is among the concepts that the fourth industrial revolution has brought along, seeking the ways to overcome encountered problems in management, planning, practice, strategy, analysis,

*CONTACT: Ugur OZALP ✉ ozalp.ugur@gmail.com 🖃 Marmara University, Faculty of Education, Department of Educational Sciences, İstanbul / Turkiye

cooperation, human resources, change and leadership (el Hamdi et al., 2019; Mohamed, 2018; Schneider, 2018; Suciu & Năsulea, 2019). Like many other concepts in social sciences, there is no consensus on the definition of intellectual capital (de Castro et al., 2010). Some scholars focused on its being knowledge-based (Bontis et al., 2002; Cabrita & Bontis, 2008; Chang et al., 2008; Dzinkowski, 2000; Holland & Holland, 2010; Nahapiet & Ghoshal, 1998), some others focused on its providing competitive advantage (de Castro et al., 2010; Delgado-Verde & Cruz-González, 2010; Hsu & Fang, 2009) and some others focused on its having potential to turn into inter-organizational value (Martínez-Torres, 2006; Sohrabi et al., 2010).

Intellectual capital is a combination of all the intangible assets and skills of the members of the organization. Managing this combination serves as a useful tool in the value creation process for the administrators (Brătianu & Pînzaru, 2015). Besides, it affects the decision-making processes of stakeholders by presenting valid and transparent data (Ramírez & Gordillo, 2013; Todericiu & Stanit, 2016). It also contributes to strengthening the long-term vision of the organization, increases the satisfaction that the stakeholders experienced as a result of sense of confidence towards the organization, and helps positive corporate image and corporate reputation building (Ramirez et al., 2016).

Intellectual capital represents the total knowledge of the members of an organization. In other words, it is the collective ability of members which involves perception of knowledge and learning. Organizations are able to gain competitive advantage by making decisions involving production thanks to their intellectual capital which represents intangible assets they possess. Intellectual capital stems from interactions of organizations with their environments and its value increases as long as it is used. Apart from constituting a competitive advantage providing factor, intellectual capital is also an essential tool for creating internal value (Kelly, 2004b; Ren, 2009; Roos et al., 1997; Semenov, 2016).

As it depends mostly on knowledge, it is impossible to completely eradicate intellectual capital. In addition, being knowledge-dependent prevents it from value loss and its value constantly increases. In addition to its being at no cost for organizations, it also constitutes both input and output of the value creation process in the organization. Because it is in the minds of the members and placed in the processes of the organization, intellectual capital is also an inimitable source (Dean & Kretschmer, 2007; Sohrabi et al., 2010).

Although there are various classifications regarding the dimensions of intellectual capital, it was observed that a considerable number of studies classified it as human capital, structural capital and relational capital (Bontis, 1998; Carson et al., 2004; Chan, 2009; de Castro et al., 2010; Delgado-Verde & Cruz-González, 2010; Huang et al., 2007; O'Donnell & O'Regan, 2000; Pedrini, 2007; Saint-Ogne, 1996).

Human capital represents know-how, experiences and skills of the members of the organizations (de Castro et al., 2010). It is used for expressing the importance of the abilities and problem-solving skills of the individuals for the organization (Suciu & Năsulea, 2019). It is the collective knowledge and experience that provides sustainable competitive advantage to the organization (Kelly, 2004a). Fitz-enz (2019) puts forward that it is the combination of the elements that an individual brings to the organization such as intellect, commitment, imagination and creativity. It refers to the knowledge, skills and abilities in the minds of the members of the organization that they use for achieving organizational goals. As it does not belong to the organization, losing members is a threat for the organization in terms of human capital. One of the most important skills for the organizations is to preserve the human capital they have and thus, become the center of attraction for the human capital their competitors possess (Bontis, 1998; Bontis et al., 2000; Chen et al., 2004; Demir, 2018; Görmüş, 2009; Kaya & Kesen, 2014; Kutlu, 2009).

Structural capital refers to the processes, procedures, strategies and policies that shape and develop the organization. Structural capital includes the organizational structure and technological infrastructure of the organization (de Castro et al., 2010; Suciu & Năsulea, 2019). Members of the organization provide intellectual input that shapes structural capital. This aspect makes it specific to the organization (Sohrabi et al., 2010). Structural capital is implicit knowledge acquired through language and narratives embedded in the social interactions between members of the organization and includes organizational capabilities developed to meet market requirements. In this context, it can be stated that all management tools, infrastructures, R&D studies, patents or trademarks used for increasing the efficiency and productivity of the organization are part of the structural capital. Organizations with strong structural capital have a supportive culture that allows organization members to make innovative attempts, fail and learn from the experience of failure (Bontis, 1998, 2002; Bontis et al., 2000; Dzinkowski, 2000; Mura & Longo, 2013).

Relational capital expresses the sum of assets of the organization regarding its relations with its environment. Relational capital refers to the relations of an organization with the stakeholders, beneficiaries of its products or services, its external environment, suppliers, government agencies, the society and its competitors (Bontis, 2002; Bozbura & Toraman, 2004; de Castro et al., 2010; Fitz-enz, 2019; Sohrabi et al., 2010; Suciu & Năsulea, 2019). It is the basic indicator of turning intellectual capital into production and added value. Without relational capital, it is not possible to create marketing value or obtain corporate performance. Relational capital blooms on human capital and structural capital. As it depends on customer loyalty and relations with suppliers which are out of the boundaries of influence of the organization, it is the most difficult dimension of intellectual capital to build. Just like human capital, it is not owned by the organization. It is important to turn relational capital into a part of structural capital (Baş et al., 2014; Bontis, 1998; Bontis et al., 2000; Chen et al., 2004; Dzinkowski, 2000).

Intellectual capital is an important power source for an organization in competition. Expressing the innovative power and innovative potential of organizations, intellectual capital is also important for higher education institutions that adopted long-term sustainable development as principle. Measurement of intellectual capital and sharing the results with the stakeholders provide higher education institutions with the opportunity to strengthen the perception regarding their reputation (Kelly, 2004b; Matos et al., 2019; Suciu & Năsulea, 2019).

In the era of a knowledge-based economy, increasing intellectual capital potential depends on education. In addition to providing other benefits, education has an indisputably important role in the future of countries with the economic incomes it brings (Chatterji & Kiran, 2017; Jakubowska & Rosa, 2014). Intellectual capital affects the efficiency of instruction and research which are among the duties of higher education institutions and constitutes input for education simultaneously (Lu, 2012; Sánchez et al., 2009). Higher education institutions produce and market certificates presenting evidence for the degree earned as product, and instruction, learning and socialization opportunities as services. Perception regarding the products and services directly affects the value attributed to them (Brenca & Gravite, 2013).

In an academic context, intellectual capital refers to intangible assets such as innovation capacity, patents owned, skills of the members or social level of acceptance (Ramírez & Gordillo, 2014). Kelly (2004b) puts forward that academic intellectual capital is the knowledge of the faculty members and its reflection on turning the knowledge into values. In this respect, the added value that academic intellectual capital provides both for the society and the higher education institutions which are expected to contribute to economic growth, to lead up social developments and to promote entrepreneurship should be investigated (Brătianu & Pînzaru, 2015; Mariani et al., 2018). Academic intellectual capital comprises the input of the knowledge creation process in higher education institutions. It refers to all intangible sources that provide basis for knowledge and have the potential to provide a competitive advantage. Consequently, in an academic context, intellectual capital indicates elements beyond accounting (Leitner, 2004).

Academic intellectual capital is directly related to the qualifications of the members of the organization and it refers to the intellectual value of human potential in education, research and socialization processes. Elements such as qualifications of faculty staff, use of physical and technological resources for improving instruction and research, student or faculty mobility, and ownership of intellectual properties are within the scope of academic intellectual capital (Brenca & Garleja, 2013; Silva & Ferreira, 2019). In addition to these, academic intellectual capital has a positive influence on the life quality of societies by affecting the sustainable development of the countries (Pedro et al., 2020).

In educational contexts in which both the input and the output are people and knowledge, it is of great importance to effectively and efficiently manage intellectual capital – the intangible assets (Basile, 2009; Karakuş, 2008; Kelly, 2004b; Ramírez Córcoles & Tejada Ponce, 2013). Measurement of academic intellectual capital is expected to lead to managerial, cultural and organizational changes and it is important as it will set the future route of the higher education institution (Kelly, 2004b; Todericiu & Şerban, 2015). In addition to its providing an indicator for the quality of instruction, measuring academic intellectual capital is also expected to provide insight about the competitive advantage of the institution in an international context (Lu, 2012).

A number of scales were developed to measure intellectual capital. However, a significant number of them focus on business organizations and most of them depend on the opinions of senior executives of the firms. For example, Bontis (1998) developed a tool for measuring the intellectual capital of the firms and carried out the study with MBA students who represented the organizations they worked in. Another example is a study by Chen et al. (2004) which was carried out by participation of entrepreneurs, general managers or the top executives of high-tech enterprises. Youndt and Snell (2004) also developed an intellectual capital scale targeting top-level executives of firms. Another scale developed by Subramaniam and Youndt (2005) involved the executives and vice presidents of human resources of enterprises. Huang et al. (2010) developed a scale with the participation of managers of companies. Another scale is of Han and Li (2015) that was developed with the participation of middle or senior managers of firms. Another intellectual capital scale by Asiaei and Jusoh (2017) used chief financial executives as the data source. Another example is by Urban and Joubert (2017) in which their data source was CEOs or owners of enterprises. Apart from making use of scales developed for business environment in the academic context, it was observed that scales for measuring academic intellectual capital were not common in the literature. For example, de Frutos-Belizón et al. (2019) developed an academic intellectual capital scale for measuring the perceptions of academics and researchers.

In the above-mentioned scales, it was observed that decision-makers are used as a data source in general. Cabrita and Vaz (2005) propose that evaluation of intellectual capital requires awareness in terms of organizational strategy and these strategically aware individuals are mainly chief executive officers, directors or top-level administrators. However, in this study, student perceptions regarding intellectual capital are in focus. We believe that, for educational institutions, students constitute both the input and the output of the process. From this point of view, it is thought that the scale developed in this study will contribute to the literature in terms of reflecting perceptions of different stakeholders of educational processes in a higher education context.

2. METHOD

In this section, information regarding the sample is presented and steps of scale development are explained in detail.

2.1. Sample

The snowball sampling method which allows data collection in case of population listing is not possible or it is impossible to compile the entire list was used in this study (Fink, 2010). Snowballing started with 40 students studying at various higher education institutions in the 2020-2021 academic year. A total of 1117 students from 112 institutions were reached for data collection.

OECD/Eurostat (2018) puts forward that it is appropriate to make use of online data collection techniques in academic studies. In addition to its being a low-cost way in terms of both time and money, online data collection also enables researchers to gather data in electronic format. Thus, it becomes easier to analyze the data (Harris et al., 2007; Tajvidi & Karami, 2015). Online data collection also provides the researcher with the comfort of suppressing the missing data by not allowing the participant to continue without answering certain questions (OECD/Eurostat, 2018). However, on the other hand, online data gathering also holds the probability of low participation level or poor data (Sultan & Wong, 2019). Data in this study were gathered online by making use of Google Forms. For securing the data quality, a control item requesting participants to choose a certain answer (*For this item, please choose 'partially true' option*) was also included in the form.

Data were collected in two stages: 574 students participated in the first stage in which EFA was conducted and 543 students participated in the second stage in which CFA was conducted. However, 36 participants from EFA and 51 participants from CFA were excluded from the analyses as they were confirmed to give the same answer for all the items and/or did not follow the control item. In the first stage, data were gathered from 538 students studying in 96 universities included in the study. Following EFA, in the second stage, data were collected from 492 students in 112 universities who didn't get involved in the first stage of the study. Data regarding the participants are presented in Table 1.

Table 1. *Participants.*

	1st Stage		2nd Stage	
	N	%	N	%
Female	352	65.43	307	62.40
Male	186	34.57	185	37.60
Associate	52	9.67	60	12.19
Bachelor's	313	58.18	208	42.28
Master's	138	25.65	162	32.93
Doctoral	35	6.50	62	12.60
State University	471	87.55	413	83.94
Foundation University	67	12.45	79	16.06
Research University	45	8.36	74	15.04
Candidate Research University	38	7.07	25	5.08
Other State University	388	72.12	314	63.82
Foundation University	67	12.45	79	16.06
Total	538	100	492	100

Table 1 shows that there were 352 female and 186 male participants in the 1st stage while there were 307 female and 185 male participants in the 2nd stage. Besides, 52 associate degree students, 313 bachelor's degree students, 138 master's degree students and 35 doctoral students participated in the 1st stage; 60 associate degree students, 208 bachelor's degree students, 162 master's degree students and 62 doctoral students participated in the 2nd stage. Out of 471 state university students who participated in the 1st stage, 45 were studying at research universities, 38 were studying at candidate research universities, and 388 were studying at other state universities. Out of 413 state university students who participated in the 2nd stage, 74 were studying at research universities, 25 were studying at candidate research universities and 314 were studying at other state universities.

DeVellis (2017) emphasizes that sample size in EFA is a controversial issue. Similarly, Johnson and Morgan (2016) state that there is no universal rule of thumb for sample size in EFA. However, they put forward that the more the number of participants the better EFA will result. Field (2018) claims that it is essential to have more than 300 participants in order for the results of EFA to be reliable. On the other hand, Irwing and Hughes (2018) assert that the number of participants in EFA is expected to exceed 500 if it is aimed to generalize the results. Similarly, Worthington and Whittaker (2006) put forward that there need to be over 300 participants for CFA. In this perspective, it is possible to state that 538 participants for EFA and 492 participants for CFA are sufficient.

2.2. Development of the Scale

In scale development, primarily, answers for the following questions are sought (Lane et al., 2016): What is the measured structure? Who will be the participants? How will the results be used? What will the scale format be? Büyüköztürk et al. (2020) propose that a scale can be developed in seven steps: (1) defining the purpose of the scale, (2) determining the feature to be measured, (3) preparing the draft item pool, (4) technical supervising and inspecting in terms of language, (5) gathering expert opinions, (6) collecting data, (7) evaluating psychometric aspects of the scale. In this study, the abovementioned steps were followed for scale development.

2.2.1. *Purpose of the scale*

At this stage, the target group of the scale, how the results will be interpreted and how the results will be used is decided (American Educational Research Association, 2014; Büyüköztürk et al., 2020). In this context, the target group of the Academic Intellectual Capital Scale was decided to be students who are studying in higher education institutions. Also, it was decided that the results of the scale to be used for evaluating the level of perceived academic intellectual capital level of the higher education institutions.

2.2.2. *Feature to be measured*

According to Johnson and Morgan (2016), researchers develop scales to measure the knowledge level, behavior or perceptions of the participants. At this phase, it is decided whether the scale should focus on apprehension, attitude, self-efficacy or academic success (Büyüköztürk et al., 2020). In this study, it was decided to measure the level of perception of the participants with the scale.

2.2.3. *Draft item pool*

Different techniques such as literature review, interview or consulting expert opinions for item development are widely used. It is important to consider that the number of items in the pool should both be manageable for the researcher and not be time-consuming for the participants (Büyüköztürk et al., 2020; DeVellis, 2017; Johnson & Morgan, 2016). Carpenter (2018) emphasizes that reviewing literature holds importance in determining the factor structure of a phenomenon. At this stage, it was decided to review the literature for preparing the draft item pool. Following the literature review, it was inferred that academic intellectual capital might have three underlying factors: academic human capital, academic structural capital, academic relational capital. It was also decided that the scale would be in five-point Likert format. Finally, a draft item pool consisting of 90 items was prepared and the options for the items were decided: *(1) not true at all*, *(2) partially true*, *(3) fifty-fifty*, *(4) true to a great extent*, *(5) completely true*.

2.2.4. *Technical supervision and inspection in terms of language*

At this stage, together with language clarity, the convenience of the items for the structure intended to be measured are inspected (Büyüköztürk et al., 2020; Lane et al., 2016). For this reason, draft item pool was sent to a panel of 4 language experts who hold a bachelor's degree in the Turkish Language. Depending on the panel's feedback on punctuation and grammar, items in the draft pool were revised.

2.2.5. *Opinions of panel of experts*

Content validity refers to the level of the items' representing the structure intended to be measured (Markus & Smith, 2010; Martinez, 2017). Evaluation of content validity allows researchers to eliminate the items which do not serve the purpose of the scale (Litwin, 2002). Content validity also serves as an indicator of construct validity (Markus & Lin, 2010).

Wilson et al. (2012) state that the most widely used technique for evaluating the content validity in most of the fields such as education, health, organizational development, marketing, psychology is the Content Validity Ratio (CVR) proposed by Lawshe (1975). CVR, calculated depending on the opinions of a panel of experts, provides a quantitative basis for evaluating the items before deciding on the inclusion of them in the scale (Gilbert & Prion, 2016). As this approach is built upon gathering the opinions of field experts, it holds great importance to decide on the members of the panel of experts for ensuring the content validity (American Educational Research Association, 2014).

In order to determine content validity, draft item pool was sent to a panel of experts. Experts were decided depending on the criterion sampling method. Criteria for the experts were stated as follows: having a Ph.D. degree in the educational administration field, having research on higher education management and working in a higher education institution. 1 scholar holding professor title, 6 scholars holding associate professor title and 6 scholars holding assistant professor title, totally 13 academics from 9 higher education institutions were reached for expert opinion.

Depending on expert opinions, CVR for each item was calculated using Lawshe's formula (1975) and evaluated using Content Validity Criterion (CVC) proposed by Ayre and Scally (2014). Ayre and Scally (2014) inform that CVC for a panel of 13 experts is .538. Following the opinions of experts, it was determined that 31 items out of 90 were found to be suitable for the scale. 31 items in the pool are presented in Table 2.

Please note that items written in English in Table 2 are provided only to give insight about items in Turkish, thus the readers should handle the item pool accordingly.

Table 2. *Item pool for Academic Intellectual Capital Scale (31 items).*

#	Item in Turkish	English Translation
1	Üniversitemizde bilimsel araştırmaya odaklanmış güçlü bir akademik kültür vardır.	There is a strong academic culture focused on scientific research in our university.
2	Üniversitemizdeki öğretim elemanları, öğrencileri girişimciliğe teşvik eder.	Faculty staff in our university leads students in entrepreneurship.
3	Üniversitemizdeki öğretim elemanları, yüksek akademik niteliklere sahiptir.	Faculty staff in our university has high academic qualifications.
4	Üniversitemiz, verilen eğitim içeriğini destekleyecek nitelikte dijital donanıma sahiptir.	Our university has the digital equipment to support the content of the education it provides.
5	Üniversitemizdeki öğrenciler, birbirlerinin fikirlerine değer verir.	Students in our university value each other's ideas.
6	Üniversitemiz, alanlarının en başarılı öğretim elemanlarına sahiptir.	Our university has the most successful faculty staff in their fields.
7	Üniversitemiz, verilen eğitim içeriğini destekleyecek nitelikte bina, donatı, vb. fiziki olanaklara sahiptir.	Our university has physical facilities such as buildings and hardware to support the content of the education it provides.
8	Üniversitemizde karar verilirken dış paydaşların (çevre, yerel yönetimler, iş dünyası vb.) fikirleri dikkate alınır.	Opinions of external stakeholders (environment, local authorities, business world, etc.) are taken into account in decision-making in our university.
9	Üniversitemizde yeterli sayıda öğretim elemanı görev yapar.	There is a sufficient number of faculty staff in our university.
10	Üniversitemizde ihtiyaçlara cevap verecek nitelikte bir bilgi yönetim sistemi (ders seçimi, not takibi vb.) kullanılır.	An information management system (for course selection, academic record tracking, etc.) that satisfies the needs is used in our university.
11	Üniversitemizde karar verilirken mezun öğrencilerin fikirleri dikkate alınır.	Opinions of alumni are taken into account in decision-making in our university.
12	Üniversitemizdeki öğretim elemanları, çalışmalarını iş birliği içerisinde yürütür.	The faculty staff carries out their studies in cooperation in our university.
13	Üniversitemizdeki kütüphane olanakları yeterlidir.	Library facilities are sufficient in our university.
14	Üniversitemizin, iş dünyasında faaliyet gösteren kurumlarla iş birliği protokolleri vardır.	Our university has cooperation protocols with institutions in the business world.
15	Üniversitemiz, ihtiyaca cevap verecek nitelikte bir e-öğrenme platformuna sahiptir.	Our university has an e-learning platform that satisfies the needs.
16	Üniversitemizin, sektördeki kuruluşlarla imzalanmış mezun işe alım protokolleri vardır.	Our university has recruitment protocols with institutions in the sector for the graduates.
17	Üniversitemizin başka üniversitelerle iş birliği protokolleri vardır.	Our university has cooperation protocols with other universities.
18	Üniversitemiz bünyesinde işlevsel bir teknoloji transfer birimi vardır.	Our university has a functional technology transfer unit.
19	Üniversitemizde, bilimsel anlayışı topluma yaymaya yönelik etkinlikler düzenlenir.	Activities for disseminating scientific perspective to society are organized in our university.
20	Üniversitemizde farklı kültürel birikimleri olan kişiler uyum içinde çalışır.	People with diverse cultural backgrounds work in harmony in our university.

Table 2. *Continued.*

#	Item in Turkish	English Translation
21	Üniversite yönetimi, bilgiye kolay ulaşım olanakları sunar.	The university administration offers easy access to information.
22	Üniversitemizde çevre sorumluluğuna ilişkin etkinlikler düzenlenir.	Activities related to environmental responsibility are organized in our university.
23	Üniversitemizdeki bilgi yönetim sistemi (ders seçimi, not takibi vb.), öğretim elemanları tarafından etkin bir şekilde kullanılır.	The information management system (for course selection, academic record tracking, etc.) in our university is effectively used by the faculty members.
24	Üniversitemiz, yeni iş girişimi (start-up) firmalarını destekler.	Our university supports start-up companies.
25	Üniversitemizdeki öğretim elemanları, üniversitemizin kurumsal hedeflerini gerçekleştirmek için çaba sarf eder.	The faculty staff strives for achieving the corporate objectives of our university.
26	Üniversitemizdeki bilgi yönetim sistemi (ders seçimi, not takibi vb.), öğrenciler tarafından etkin bir şekilde kullanılır.	The information management system (course selection, academic record tracking, etc.) in our university is effectively used by the students.
27	Üniversitemizdeki öğretim elemanları, öğrencilerden gelen geri bildirimlere önem verir.	The faculty staff at our university value the feedback from the students.
28	Üniversitemizdeki öğrenciler, yaşadıkları sorunları yöneticilere açık bir biçimde dile getirebilir.	The students at our university can overtly utter the problems they face to the administrators.
29	Üniversitemiz, mezun öğrencileriyle irtibat halindedir.	Our university keeps in contact with the alumni.
30	Üniversitemiz, özgün fikirleriyle bilinen öğretim elemanlarına sahiptir.	Our university has faculty staff known for their peculiar ideas.
31	Üniversitemizdeki öğretim elemanları, öğrencileri ekip çalışması yapmaya teşvik eder.	The faculty staff at our university encourage students for teamwork.

2.2.6. *Data collection*

At this stage, data are collected using a draft scale. Once the construct and the content of the scale are evaluated as satisfactory, it is inferred that the draft scale is ready for data collection (Büyüköztürk et al., 2020; DeVellis, 2017; Johnson & Morgan, 2016). Psychometric aspects of the scale are determined depending on the data collected at this stage (Irwing & Hughes, 2018; Netemeyer et al., 2003).

Data were collected in two steps. First, the draft scale was used and 538 participants were reached. Using the data from the first step, EFA was conducted and the number of the items reduced. Second, using the final version of the scale depending on the EFA results, another 492 participants were reached and data collected for conducting CFA.

2.2.7. *Evaluation of psychometric aspects of the scale*

Once data are gathered, scale is shaped using statistical techniques at this stage (Büyüköztürk et al., 2020). As it covers validity and reliability analyses, it is possible to call this stage the heart of the scale development process (DeVellis, 2017). Mainly, two types of analyses were followed at this stage: EFA and CFA.

2.3. Data Analysis

R (version 4.1.2) (R Core Team, 2021) and RStudio (version 2021.09.1 build 372) (RStudio Team, 2021) were used to analyze the data. *data.table* (Dowle & Srinivasan, 2020), *dplyr* (Wickham et al., 2020), *EFAtools* (Steiner & Grieder, 2020), *EFA.dimensions* (O'Connor, 2020), *lavaan* (Rosseel, 2012), *psych* (Revelle, 2020), *rela* (Chajewski, 2009), *semTools* (Jorgensen et al., 2021), *ShinyItemAnalysis* (Martinková & Drabinová, 2018), *QuantPsych* (Fletcher, 2015) and *sirt* (Robitzsch, 2021) packages were used in the analyses.

3. FINDINGS

Findings of EFA, CFA and reliability analyses are presented in this section.

3.1. Exploratory Factor Analysis

With the help of EFA, it is possible to reduce the number of items in a scale, thus variance explained by the scale can be maximized (Netemeyer et al., 2003). EFA is used for determining underlying non-observable factor structures through observable variables (Hayashi & Yuan, 2010). In order to determine the factorial structure of the Academic Intellectual Capital Scale, EFA was conducted following the five-step model proposed by Williams et al. (2010).

3.1.1. *Checking the data for suitability*

In order to determine if the data are suitable for EFA, its factorability should be checked first. Field (2018) draws attention that the correlation coefficient between the variables shouldn't be lower than .30 and should not exceed .80. The correlation matrix was investigated and it was observed that there was no correlation coefficient above .80 or below .30 between the variables. Another way of determining suitability for factoring of data is checking the anti-image correlation matrix. Şencan (2005) puts forward that elements that are off-diagonal in the anti-image correlation matrix should be below .30 and diagonal elements of the matrix should be above .50. The anti-image correlation matrix was obtained using *rela* package (Chajewski, 2009). Examining the anti-image correlation matrix showed that all diagonal elements were above .50 and off-diagonal elements were below .30 indicating that the data were suitable for factorability.

EFA is a method that depends on Pearson product-moment correlation and it assumes that the data are normally distributed. For this reason, violation of this assumption holds potential to affect the EFA results in an unintended way (Watkins, 2021). Also, Field (2018) and Şencan (2005) emphasize that in order to obtain generalizable results from EFA, normal distribution of data is essential. Tabachnick and Fidell (2014) propose that skewness and kurtosis are indicators of univariate normal distribution of data. Şencan (2005) puts forwards that skewness and kurtosis of each item should be evaluated individually to check univariate normality. Leech et al. (2015) state that skewness and kurtosis should be between +1 and -1 to define data as normally distributed in terms of univariate normality. To check univariate normality, skewness and kurtosis values of each item were calculated using *psych* package (Revelle, 2020). It was found out that skewness ranged between -.82 and .10 while kurtosis ranged between -.70 and .61. Depending on these results it is possible to say that the univariate normality assumption was met. Multivariate normality was checked through multivariate skewness and multivariate kurtosis tests by making use of *QuantPsych* package (Fletcher, 2015). Multivariate normality tests resulted in significant *p* value meaning that multivariate normality was violated.

Bartlett Sphericity Test and Kaiser-Meyer-Olkin (KMO) Test are other ways of checking the suitability of data for EFA. With the help of these two tests, the factorability of the data is determined (Carpenter, 2018). Bartlett Sphericity Test is expected to be statistically significant, and KMO is expected to be above .50 (Field, 2018; Field et al., 2012). Using *EFAtools* package Bartlett Sphericity Test and KMO Test were conducted (Steiner & Grieder, 2020). Bartlett

Sphericity Test was found to be statistically significant ($\chi^2_{(465)}$ = 10715, p = .000) and KMO was .966. These results showed that the data are suitable for factorability, thus for EFA.

3.1.2. *Selection of factor extraction method*

There are various factor extraction methods in EFA such as image analysis, principal component analysis, principal axis factoring, maximum likelihood and so on (Watkins, 2021). According to Fabrigar et al. (1999) principal axis factoring (PAF) has advantage of requiring no distributional assumptions. Since multivariate normality assumptions wasn't met, it was decided to use PAF as the factor extraction method.

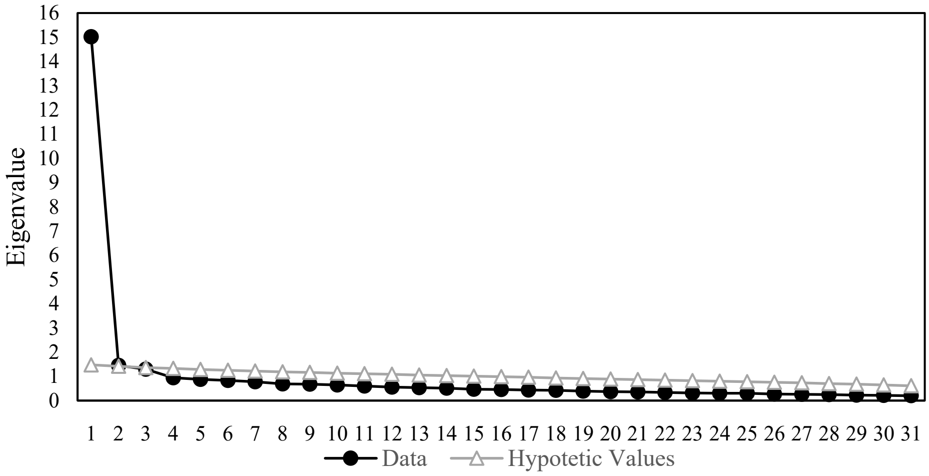
3.1.3. *Determining the number of factors*

EFA aims to reduce multiple items to fewer common structures. In scale development, it is important to determine the number of underlying factors as it is expected to reach similar results with the same scale on different samples. There are several techniques in determining the number of factors such as Kaiser rule (eigenvalue over 1), visually interpreting scree plot, parallel analysis and Velicer's Minimum Average Partial (MAP) Test (Costello & Osborne, 2005; Netemeyer et al., 2003; Preacher et al., 2013; Williams et al., 2010).

Williams et al. (2010) emphasize that using multiple techniques for determining the number of factors should be preferred to get better results. In this study, it was decided to use the Kaiser rule, parallel analysis and MAP Test together to decide on the number of factors to extract.

Eigenvalues of variance explained were calculated and it was observed that there were three factors with eigenvalue over 1. Parallel analysis was also interpreted using *psych* package (Revelle, 2020). Parallel analysis plot is shown in Figure 1. Interpretation of plot in Figure 1 shows that the scale might have 2 factors.

Figure 1. *Scree plot for parallel analysis.*



Another empirical way of determining the number of factors is the MAP Test. In this test, the partial correlation matrix is calculated after the extraction of each of the factors. The average of partial correlations is calculated for each matrix. When the appropriate number of factors is reached, it is expected to have the minimum average (Watkins, 2021). MAP Test was conducted using *EFA.dimensions* package (O'Connor, 2020). The results of the MAP Test are shown in Table 3.

Table 3 presents evidence that average squared partial correlation and 4th power partial correlation decreased until reaching the third factor and started increasing after it. Depending on the results of the MAP Test, it is possible to state that the scale has three factors.

Table 3. *Map Test results.*

	PC^{2*}	PC^{4**}		PC^{2*}	PC^{4**}		PC^{2*}	PC^{4**}
0	.22147	.05393	11	.02434	.00233	22	.09020	.02355
1	.01177	.00042	12	.02713	.00261	23	.10457	.03051
2	.01101	.00035	13	.03015	.00310	24	.12499	.03851
3	**.00992**	**.00034**	14	.03324	.00395	25	.14970	.05581
4	.01089	.00036	15	.03716	.00440	26	.18220	.07630
5	.01194	.00041	16	.04208	.00549	27	.23538	.11878
6	.01320	.00057	17	.04860	.00674	28	.33232	.20348
7	.01474	.00074	18	.05527	.00878	29	.48390	.35518
8	.01653	.00095	19	.06202	.01087	30	1.00000	1.00000
9	.01879	.00124	20	.06918	.01416			
10	.02099	.00184	21	.07845	.01847			

* Squared partial correlation

** 4th power partial correlation

Kaiser rule, parallel analysis results and the result of MAP test were evaluated together. While parallel analysis pointed out 2 factors, Kaiser rule and MAP test indicated 3 factors. Depending on the empirical results and literature review, it is inferred that the scale had three factors.

3.1.4. *Selection of rotational method*

In order to simplify the data structure and to interpret the data structure easily, rotational methods are used (Costello & Osborne, 2005; Motta, 2017). Basically, there are two types of rotational methods: oblique and orthogonal. When correlation is expected to be between factors, oblique methods are used and oblique rotational methods allow reaching statistically accurate factor structures (Field, 2018; Motta, 2017; Schmitt, 2011; Williams et al., 2010). Although there are various oblique rotational methods, Direct Oblimin and Promax are the prominent ones (Brody, 2017). Besides, Costello and Osborne (2005) put forward that it is not always possible to draw a strict line between the issues in fields such as education and psychology. Depending on the literature review, the structure of academic intellectual capital was inferred to arise from correlated elements. As a result, it was decided to use Promax oblique rotational method.

3.1.5. *Interpretation*

At this stage, items of the factors are determined and the factors are named (Williams et al., 2010). Factors are identified depending on factor loadings of the items (Johnson & Morgan, 2016). Tabachnick and Fidell (2014) propose that the lower bound for an item loading to be accepted is .32 whereas Johnson and Morgan (2016) put forward that the lower bound should be .40. In this study, it was decided that the lower bound for item loading would be .40 and it was decided to remove any items lower than that value from the scale. Also, in some cases, some of the items might have loading on more than one factor (Welch, 2010). Overlapping items with less than .20 difference in factor loadings were also decided to remove from the scale (Child, 2006). EFA is conducted using *psych* package (Revelle, 2020).

EFA was conducted and items with lower than .40 item loading were removed from the scale (respectively items 13, 9, 5, 28, 8, 20, 12 and 25). Subsequently, overlapping items were also removed from the scale (respectively items 27, 29 and 21). After each item removal, the analysis was repeated. In the end, there were 20 items left for EFA. Bartlett Sphericity Test and KMO test were also carried out with 20 items. The results of Bartlett Sphericity Test were statistically significant ($\chi^2_{(190)} = 6523.40$, $p = .000$) and KMO was .949 indicating that data of 20 items were suitable for EFA. The results of EFA are presented in Table 4.

Table 4. *EFA results.*

Items	Factors			Communalities
	1	2	3	
Item 03	.89	.04	-.14	.67
Item 06	.79	-.09	.07	.60
Item 30	.74	0	.09	.66
Item 02	.68	.11	-.04	.54
Item 31	.60	.02	.14	.53
Item 01	.50	.19	.09	.51
Item 10	.02	.89	-.15	.64
Item 26	.02	.72	0	.55
Item 04	.02	.69	.06	.56
Item 07	.03	.65	.02	.46
Item 23	.11	.65	.07	.60
Item 15	-.02	.55	.25	.54
Item 24	.08	-.18	.87	.64
Item 16	-.12	0	.87	.62
Item 18	-.11	.23	.66	.59
Item 17	-.01	.05	.65	.46
Item 22	.20	-.05	.63	.57
Item 19	.18	.05	.59	.58
Item 14	-.04	.24	.54	.50
Item 11	.17	.02	.48	.39
Variance Explained (%)	17.74	17.40	21.01	

Table 4 shows that items 3, 6, 30, 2, 31 and 1 were under factor 1 (factor loadings varied between .89 and .50), items 10, 26, 4, 7, 23 and 15 were under factor 2 (factor loadings varied between .89 and .55), and items 24, 16, 18, 17, 22, 19, 14 and 11 were under factor 3 (factor loadings varied between .87 and .48). Factor number 1 explained 17.74%, factor number 2 explained 17.40% and factor number 3 explained 21.01% of the total variance. The scale explained 56.15% of total variance.

Items in the factors were evaluated and, factor number 1 with 6 items was named *Academic Human Capital*, factor number 2 with 6 items was named *Academic Structural Capital*, and factor number 3 with 8 items was named *Academic Relational Capital*. Holistically, it is possible to state that Academic Intellectual Capital Scale is composed of three factors and 20 items, all of which are in affirmative form.

Interfactor correlations provided by *fa* function from *psych* package (Revelle, 2020) are presented in Table 5.

Table 5. *Interfactor correlations.*

	Academic Human Capital	Academic Structural Capital	Academic Relational Capital
Academic Human Capital	1	.690	.728
Academic Structural Capital	.690	1	.735
Academic Relational Capital	.728	.735	1

Table 5 shows that Academic Human Capital had positive interfactor correlation with Academic Structural Capital (.690) and positive interfactor correlation with Academic Relational Capital (.728). Academic Structural Capital had positive interfactor correlation with Academic Relational Capital (.735). Following that, a Pearson product-moment correlation test was conducted. The summary table of the correlation test was obtained by using *data.table* package (Dowle & Srinivasan, 2020). The results are presented in Table 6.

Table 6. *Correlation test results.*

	Academic Human Capital	Academic Structural Capital	Academic Relational Capital
Academic Human Capital	1	.675	.718
Academic Structural Capital	.675	1	.715
Academic Relational Capital	.718	.715	1
Academic Intellectual Capital	.880	.881	.923

Table 6 shows that Academic Human Capital had statistically significant positive correlation with Academic Structural Capital (r = .675, *p* = .000); statistically significant positive correlation with Academic Relational Capital (r = .718, *p* = .000); and statistically significant positive correlation with scale total score (r = .880, *p* = .000). Academic Structural Capital had statistically significant positive correlation with Academic Relational Capital (r = .715, *p* = .000); and statistically significant positive correlation with scale total score (r = .881, *p* = .000). Academic Relational Capital had statistically significant positive correlation with scale total score (r = .923, *p* = .000). It was found out that all the factors and scale total scores had statistically significant positive correlation. This result revealed that all the factors measure a similar structure.

3.2. Confirmatory Factor Analysis

Ullman (2014) points out that in addition to having an adequate number of participants, normal distribution of data is also important in CFA. To check univariate normality (Tabachnick & Fidell, 2014), skewness and kurtosis values of each item were calculated using *psych* package (Revelle, 2020). It was found out that skewness ranged between -.89 and .63 while kurtosis ranged between -.86 and .04. Depending on these results it is possible to infer that the data had univariate normality. Multivariate normality tests of multivariate skewness and multivariate kurtosis (Fletcher, 2015) resulted in significant *p* value meaning that multivariate normality was violated.

In order to confirm a model in CFA, there are various parameters to check. Kline (2015) suggests that χ^2, degree of freedom, the significance of χ^2, RMSEA, CFI and SRMR are the minimum parameters to look for in CFA. Schermelleh-Engel et al. (2003) draw attention that there is no consensus on which parameters to control in CFA. On the other hand, Kline (2015) emphasizes that each parameter represents a different aspect of the scale under investigation and there is no single parameter to confirm the proposed model.

Brown (2015) puts forward if the data is categorical and normality assumption is violated, Maximum Likelihood estimation method should not be used in CFA. Instead, it is possible to use one of several estimators such as ULS, WLS and WLSMV. Irwing et al. (2018) and Schmitt (2011) propose that WLSMV estimator should be used with categorical data. In addition, Li (2016) and Bagheri and Saadati (2021) assert that WLSMV has no assumptions regarding distribution of the data. In this respect, CFA with WLSMV estimator was conducted using *lavaan* package (Rosseel, 2012). Fit indices and the results of CFA are presented in Table 7.

Table 7. *Fit indices and CFA results.*

Parameter	Result	Perfect Fit	Acceptable Fit
χ^2/df	2.354	$0 \leq \chi^2/df \leq 2$	$2 \leq \chi^2/df \leq 5$
RMSEA	.053	0 ≤ RMSEA ≤ .05	.05 < RMSEA ≤ .08
SRMR	.031	0 ≤ SRMR ≤ .05	.05 < SRMR ≤ .10
NFI	.844	.95 ≤ NFI ≤ 1.00	.90 ≤ NFI < .95
NNFI	.890	.97 ≤ NNFI ≤ 1.00	.90 ≤ NNFI < .97
CFI	.903	.97 ≤ CFI ≤ 1.00	.95 ≤ CFI < .97
GFI	.998	.95 ≤ GFI ≤ 1.00	.80 ≤ GFI < .95
AGFI	.997	.90 ≤ AGFI ≤ 1.00	.80 ≤ AGFI <.90

Source: Awang (2012), Byrne (2016), Doll et al. (1994), Forza and Filippini (1998), Greenspoon and Saklofske (1998), Hooper et al. (2008), Hu and Bentler (1999), Schermelleh-Engel et al. (2003), Schumacker and Lomax (2016), Segars and Grover (1993), Steiger (2007)

The CFA results showed that χ^2 = 393.223, degree of freedom was 167 and significance was *p* = .000. Hair et al. (2018) put forward that scales having 12 to 30 items with over 250 participants are expected to have a statistically significant *p* value for χ^2. CFA results also revealed that χ^2/df was 2.354 < 5 (Doll et al., 1994; Hooper et al., 2008), RMSEA was .053 < .08 (Hooper et al., 2008; Schermelleh-Engel et al., 2003; Schumacker & Lomax, 2016), SRMR was .031 < .05 (Byrne, 2016; Doll et al., 1994; Schumacker & Lomax, 2016), GFI was .998 > .80 (Forza & Filippini, 1998; Greenspoon & Saklofske, 1998) and AGFI was .997 > .80 (Forza & Filippini, 1998; Segars & Grover, 1993). It also showed that NFI was .844 which was not far from the cut value (.90) proposed by Schermelleh-Engel et al. (2003), NNFI was .890 which was not far from the cut value (.90) proposed by Awang (2012) and Forza and Filippini (1998) and CFI was .903 which was not far from the cut value (.95) proposed by Hu and Bentler (1999). Evaluated together, CFA confirmed the proposed model for Academic Intellectual Capital Scale. Measurement model for the scale is presented in Figure 2.

Figure 2. *Measurement model for the Academic Intellectual Capital Scale.*

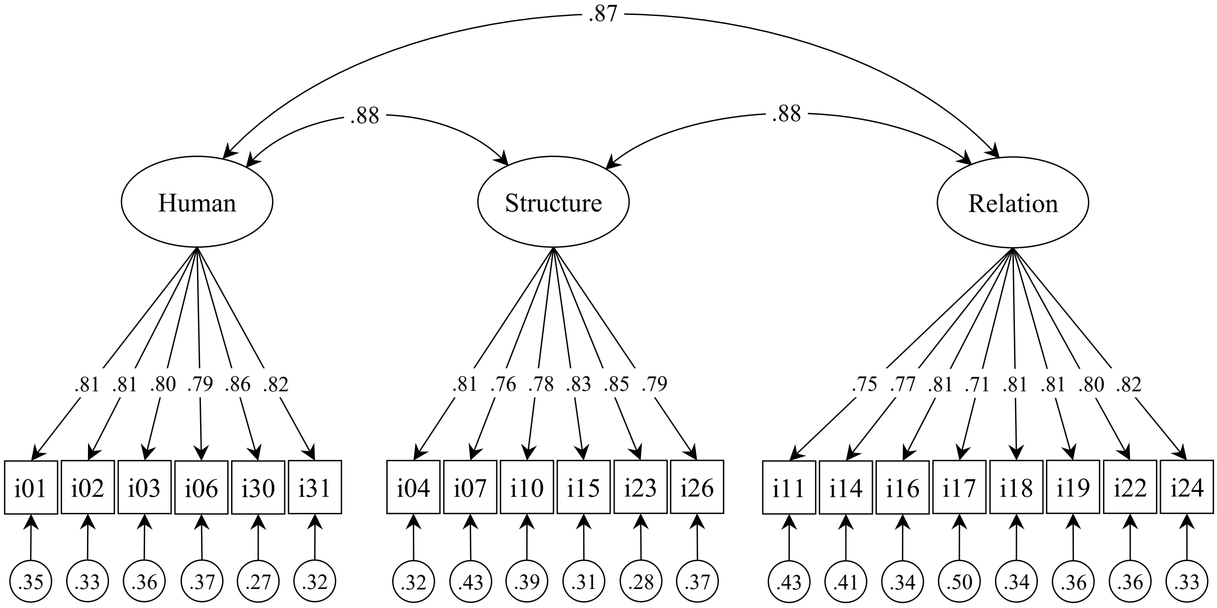


Figure 2 shows the factor loadings of the items. None of the error variances were linked. The figure also demonstrates the confirmed model of Academic Intellectual Capital Scale consisting of Academic Human Capital, Academic Structural Capital and Academic Relational Capital subdimensions.

3.3. Reliability

In order to test the reliability of the scores obtained from Academic Intellectual Capital Scale, reliability coefficient was calculated, independent samples t-test was conducted between upper 27% scores and lower 27% scores, and item-total, item-remainder correlation was calculated. Reliability analyses were carried out on a sample of 1030 participants by combining EFA and CFA data sets (538 + 492).

Field (2018) states that Cronbach's Alpha (α) is the most widely used internal consistency coefficient for scales. However, Osburn (2000) suggests that stratified alpha (α_s) provides more accurate results in terms of reliability. On the other hand, Rae (2007) draws attention that both alpha and stratified alpha should be calculated. In this perspective, it was decided to calculate stratified alpha for whole scale in addition to Cronbach's Alpha. Additionally, Composite Reliability (cR) and McDonald's Omega (ω) are other measures for internal consistency (Irwing & Hughes, 2018; Netemeyer et al., 2003). While Hair et al. (2018) put forward that cR is a more robust way of calculating internal consistency, Irwing and Hughes (2018) claim that ω is a more exact solution. McDonald's Omega was calculated using *pych* package (Revelle, 2020), Cronbach's Alpha and Composite Reliability were calculated using *semTools* package (Jorgensen et al., 2021) and stratified alpha was calculated using *sirt* package (Robitzsch, 2021). Calculated α, α_s, cR and ω coefficients are presented in Table 8.

Table 8. *Internal reliability test results.*

	Cronbach's Alpha (α)	Stratified Alpha (α_s)	Composite Reliability (cR)	MacDonald's Omega (ω)
Academic Human Capital	.906		.905	.904
Academic Structural Capital	.898		.897	.896
Academic Relational Capital	.913		.914	.909
Total Score	.957	.963	.962	.963

Alpha over .80 presents evidence for a very good internal consistency whereas alpha over .90 is an indicator of perfect consistency (DeVellis, 2017; Kline, 2015). On the other hand, Composite Reliability over .70 (Hair et al., 2018) and McDonald's Omega over .80 (Feißt et al., 2019) demonstrate that internal consistency of the scale is ensured. Scores on Table 8 indicate that the Academic Intellectual Capital Scale had a high internal consistency.

In order to determine item discrimination, together with item-total and item-remainder correlation, independent samples t-test between top 27% scores and bottom 27% scores were carried out. Netemeyer et al. (2003) and Dorans (2018) put forward that a low or negative item-remainder correlation coefficient is proof that the item does not serve the purpose of the scale. While Johnson and Morgan (2016) claim that items with item-remainder correlation coefficient should be above .20, Field (2018) defends that item with an item-remainder correlation coefficient below .30 should be removed from the scale. Item-total and item-remainder correlation coefficients were calculated using *ShinyItemAnalysis* package (Martinková & Drabinová, 2018) and independent samples t-test between upper and lower scores were calculated using *dplyr* package (Wickham et al., 2020). The results are presented in Table 9.

Table 9. *Item analysis results.*

Dimension	Item	r_{it}	*p*	r_{ir}	*p*	Top 27% $\bar{X}$	Bottom 27% $\bar{X}$	*t*	*df*	*p*
Academic Human Capital	i01	.744	.000	.712	.000	4.46	2.25	42.115	558	.000
	i02	.735	.000	.704	.000	4.45	2.31	41.686	558	.000
	i03	.735	.000	.703	.000	4.72	2.49	43.559	558	.000
	i06	.728	.000	.693	.000	4.50	2.10	46.153	558	.000
	i30	.783	.000	.756	.000	4.41	2.24	40.372	558	.000
	i31	.745	.000	.714	.000	4.47	2.31	40.887	558	.000
Academic Structural Capital	i04	.749	.000	.717	.000	4.50	2.19	42.934	558	.000
	i07	.689	.000	.651	.000	4.47	2.20	44.092	558	.000
	i10	.722	.000	.687	.000	4.55	2.23	43.955	558	.000
	i15	.759	.000	.726	.000	4.48	2.03	47.120	558	.000
	i23	.779	.000	.750	.000	4.53	2.30	42.549	558	.000
	i26	.724	.000	.690	.000	4.59	2.40	41.250	558	.000
Academic Relational Capital	i11	.695	.000	.653	.000	4.24	1.49	69.739	558	.000
	i14	.729	.000	.695	.000	4.42	2.13	45.764	558	.000
	i16	.752	.000	.719	.000	4.26	1.66	67.259	558	.000
	i17	.678	.000	.643	.000	4.36	2.32	38.895	558	.000
	i18	.763	.000	.733	.000	4.30	1.90	54.482	558	.000
	i19	.775	.000	.746	.000	4.36	1.90	55.220	558	.000
	i22	.758	.000	.726	.000	4.47	2.10	44.905	558	.000
	i24	.759	.000	.728	.000	4.25	1.75	68.127	558	.000

Table 9 shows that item-total correlation coefficients (r_{it}) varied between .678 and .783, item-remainder correlation coefficients (r_{ir}) varied between .643 and .756, and all the values obtained were statistically significant. In addition, Table 9 demonstrates that there is a statistically significant difference between the upper 27% scores and the lower 27% scores in favor of the upper segment for all the items. Reviewed together, results of item analyses presented evidence that the Academic Intellectual Capital Scale consisted of discriminating items.

4. DISCUSSION and CONCLUSION

In this study, perceptions of students regarding the academic intellectual capital of higher education institutions were in focus. Determining student perceptions in terms of academic intellectual capital holds importance as the students constitute both the input and the output of the educational process the higher education institutions provide. From this perspective, it is expected that this study will contribute to the literature. In this study, a scale with a valid structure for measuring academic intellectual capital levels of higher education institutions depending on student perceptions was developed. Item pool consisting of 90 items for the scale was formed after an extensive literature review. 4 language experts evaluated the initial item pool for language suitability and a panel of 13 scholars evaluated the items to ensure content validity. 59 items were eliminated depending on the opinions of field experts. The draft item pool had 31 items.

Data were collected in two stages. In the first stage, 538 students from 96 universities participated in the study and the draft item pool consisting of 31 items was used. In the second stage, 492 students who didn’t take part in the first stage from 112 universities participated in the study.

In the first stage, the main aim was to reveal scale structure through EFA. Results of EFA revealed that Academic Intellectual Capital Scale had three factors. These were Academic Human Capital, Academic Structural Capital and Academic Relational Capital. There were 6 items in the Academic Human Capital factor, 6 items in the Academic Structural Capital factor and 8 items in the Academic Relational Capital factor. The Academic Human Capital factor explained 17.74% of the total variance, the Academic Structural Capital explained 17.40% of the total variance and the Academic Relational Capital factor explained 21.01% of the total variance. The total variance explained by the scale was 56.15%. In the second stage of the study, the theoretical model proposed by the results of EFA was validated by CFA. Results of CFA confirmed that Academic Intellectual Capital Scale consisted of three factors and 20 items, all in affirmative form. The scale is structured in 5-point Likert-type with options ranging from *(1) not true at all* to *(5) completely true*. A score between 20 and 100 can be obtained from the scale. The higher the obtained score, the better the perception of students regarding the academic human capital, academic structural capital and academic relational capital, and vice versa.

The reliability of the scores obtained from the scale was tested by Cronbach's Alpha, stratified alpha, Composite Reliability and McDonald's Omega. Cronbach's Alpha for Academic Human Capital score was .906, for Academic Structural Capital score was .898, for Academic Relational Capital score was .913, and for the total score was .957. Stratified alpha was .963. Composite Reliability for Academic Human Capital score was .905, for Academic Structural Capital score was .897, for Academic Relational Capital score was .914 and for the total score was .962. McDonald's Omega for Academic Human Capital score was .904, for Academic Structural Capital score was .896, for Academic Relational Capital score was .909 and for the total score was .963. Results of the reliability tests proved that the scale had internal consistency.

Item discrimination was inspected by calculating item-total and item-remainder correlation coefficients. In addition, a t-test was conducted between upper 27% scores and lower %27 scores for all the items. Item-total and item-remainder correlations revealed that all the items in the scale served the purpose of the scale. Results of the t-test showed that there was a statistically significant difference between upper 27% scores and lower 27% scores in favor of upper scores. Items in the scale were proved to be discriminating. The final form of the Academic Intellectual Scale is provided in Appendix.

Managing academic intellectual capital is among the inevitable outcomes of the knowledge era. It was observed that the studies in the literature on measurement tools regarding academic intellectual capital were limited. The Academic Intellectual Capital Scale developed in this study was statistically proven to be a measurement tool with a valid structure. In this context, it is expected the Academic Intellectual Capital Scale contributes to the literature. With the help of this scale, administrators of higher education institutions may have the opportunity to get a clearer picture of the student perceptions regarding academic intellectual capital.

It should be kept in mind that this study only covers the perceptions of the students of higher education institutions. Similar studies on perceptions of faculty staff, non-academic staff and/or administrators are suggested to be carried out to provide a more explicit view of the academic intellectual capital of higher education institutions. Also, it should be noted that the data was gathered using snowball sampling method which is one of non-probability sampling techniques. A probability sampling technique may be used in the future studies. In addition, the fact that items with loading below .40 were removed from the scale during EFA might have led to a reduced content validity.

Acknowledgments

This study was produced from the first author's doctoral dissertation under the supervision of the second author.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number:** Marmara University, 15.02.2021, 2100042416.

Authorship Contribution Statement

Ugur Ozalp: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing the original draft. **Munevver Cetin**: Supervision and Validation.

Orcid

Ugur OZALP https://orcid.org/0000-0001-6790-5304
Munevver CETIN https://orcid.org/0000-0002-1203-9098

REFERENCES

American Educational Research Association. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Asiaei, K., & Jusoh, R. (2017). Using a robust performance measurement system to illuminate intellectual capital. *International Journal of Accounting Information Systems*, *26*, 1–19. https://doi.org/10.1016/j.accinf.2017.06.003

Awang, Z. (2012). *Structural equation modeling using AMOS graphic*. Universiti Teknologi MARA Publication Centre (UPENA).

Ayre, C., & Scally, A.J. (2014). Critical values for Lawshe's content validity ratio: Revisiting the original methods of calculation. *Measurement and Evaluation in Counseling and Development*, *47*(1), 79–86. https://doi.org/10.1177/0748175613513808

Bagheri, A., & Saadati, M. (2021). Generalized structural equations approach in the of elderly self-rated health. *Journal of Physics: Conference Series*, *1863*(1), 1-10. https://doi.org/10.1088/1742-6596/1863/1/012041

Basile, C.G. (2009). *Intellectual capital: The tangible assets of professional development schools*. State University of New York Press.

Baş, M., Mısırdalı Yangil, F., & Aygün, S. (2014). Entelektüel sermaye alanında yapılan lisansüstü tez çalışmalarına yönelik bir içerik analizi: 2002 – 2012 dönemi [A content analysis on the dissertations on intellectual capital: 2002 – 2021 period]. *International Journal of Management Economics and Business*, *10*(23), 207–226. https://doi.org/10.17130/ijmeb.2014.10.23.618

Bontis, N. (1998). Intellectual capital: an exploratory study that develops measures and models. *Management Decision*, *36*(2), 63–76. https://doi.org/10.1108/00251749810204142

Bontis, N. (2002). Managing organizational knowledge by diagnosing intellectual capital: Framing and advancing the state of the field. In N. Bontis (Ed.), *World Congress on Intellectual Capital Readings* (pp. 13–56). Elsevier.

Bontis, N., Crossan, M. M., & Hulland, J. (2002). Managing an organizational learning system by aligning stocks and flows. *Journal of Management Studies*, *39*(4), 437–469. https://doi.org/10.1111/1467-6486.t01-1-00299

Bontis, N., Keow, W.C.C., & Richardson, S. (2000). Intellectual capital and business performance in Malaysian industries. *Journal of Intellectual Capital*, *1*(1), 85–100. https://doi.org/10.1108/14691930010324188

Bozbura, F.T., & Toraman, A. (2004). Türkiye'de entelektüel sermayenin ölçülmesi ile ilgili model çalışması ve bir uygulama [A model study of measurement intellectual capital in Turkey and an application]. *İTÜ Dergisi*, *3*(1), 55-66. http://itudergi.itu.edu.tr/index.php/itudergisi_d/article/viewFile/709/643

Brătianu, C., & Pînzaru, F. (2015). Challenges for the university intellectual capital in the knowledge economy. *Management Dynamics in the Knowledge Economy*, *3*(4), 609–627. https://www.managementdynamics.ro/index.php/journal/article/view/153/102

Brenca, A., & Garleja, R. (2013). Intellectual capital in the higher education institutions of Latvia in the context of international trade. *European Conference on Intellectual Capital*, 495.

Brenca, A., & Gravite, A. (2013). Intellectual capital import for the benefit of higher education. *Bulgarian Comparative Education Society*, 1–6.

Brody, N. (2017). Factor analysis: Rotated matrix. In M. Allen (Ed.), *The SAGE encyclopedia of communication research methods* (pp. 526–531). Sage Reference.

Brown, T.A. (2015). *Confirmatory factor analysis for applied research*. The Guilford Press.

Büyüköztürk, Ş., Kılıç Çakmak, E., Erkan Akgün, Ö., Karadeniz, Ş., & Demirel, F. (2020). *Bilimsel araştırma yöntemleri [Scientific research methods in education]*. Pegem Akademi. https://doi.org/10.14527/9789944919289

Byrne, B.M. (2016). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Routledge.

Cabrita, M. do R., & Bontis, N. (2008). Intellectual capital and business performance in the Portuguese banking industry. *International Journal of Technology Management*, *43*(1–3), 212–237. http://dx.doi.org/10.1504/IJTM.2008.019416

Cabrita, M. do R., & Vaz, J.L. (2005). Intellectual capital and value creation: Evidencing in Portuguese banking industry. In D. Remenyi (Ed.), *Proceedings of the European Conference on Knowledge Management, ECKM* (pp. 98–106). University of Limerick.

Carpenter, S. (2018). Ten steps in scale development and reporting: A guide for researchers. *Communication Methods and Measures*, *12*(1), 25-44. https://doi.org/10.1080/19312458.2017.1396583

Carson, E., Ranzijn, R., Winefield, A., & Marsden, H. (2004). Intellectual capital: Mapping employee and work group attributes. *Journal of Intellectual Capital*, *5*(3), 443–463. https://doi.org/10.1108/14691930410550390

Chajewski, M. (2009). *rela: Scale item analysis*. R package version 4.1.

Chan, K.H. (2009). Impact of intellectual capital on organisational performance: An empirical study of companies in the Hang Seng Index (Part 1). *The Learning Organization*, *16*(1), 4–21. https://doi.org/10.1108/09696470910927641

Chang, S.C., Chen, S.S., & Lai, J.H. (2008). The effect of alliance experience and intellectual capital on the value creation of international strategic alliances. *Omega*, *36*(2), 298–316. https://doi.org/10.1016/j.omega.2006.06.010

Chatterji, N., & Kiran, R. (2017). Relationship between university performance and dimensions of intellectual capital: An empirical investigation. *Eurasian Journal of Educational Research*, *71*, 215–232. https://doi.org/10.14689/ejer.2017.71.12

Chen, J., Zhu, Z., & Yuan Xie, H. (2004). Measuring intellectual capital: A new model and empirical study. *Journal of Intellectual Capital*, *5*(1), 195-212. https://doi.org/10.1108/14691930410513003

Child, D. (2006). *The essentials of factor analysis*. Continuum.

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 and Evaluation*, *10*(7).

de Castro, G.M., Verde, M.D., Sáez, P.L., & López, J.E.N. (2010). *Technological innovation: An intellectual capital based view*. Palgrave Macmillan.

de Frutos-Belizón, J., Martín-Alcázar, F., & Sánchez-Gardey, G. (2019). Conceptualizing academic intellectual capital: Definition and proposal of a measurement scale. *Journal of Intellectual Capital*, *20*(3), 306–334. https://doi.org/10.1108/JIC-09-2018-0152

Dean, A., & Kretschmer, M. (2007). Can ideas be capital? Factors of production in the postindustrial economy: A review and critique. *Academy of Management Review*, *32*(2), 573–594. https://doi.org/10.5465/AMR.2007.24351866

Delgado-Verde, M., & Cruz-González, J. (2010). An intellectual capital-based view of technological innovation. In P. L. Sáez, G. M. de Castro, J. E. N. López, & M. Delgado-Verde (Eds.), *Intellectual capital and technological innovation: Knowledge-based theory and practice* (pp. 166–193). Information Science Reference.

Demir, S. (2018). Entelektüel sermaye ile öğretmenlerin iş doyumları arasındaki ilişki üzerine bir çalışma [A study on the relationship between intellectual capital and teachers' job satisfaction]. *İnönü University Journal of the Faculty of Education*, *19*(3), 205–215. https://doi.org/10.17679/inuefd.385908

DeVellis, R.F. (2017). *Scale development: Theory and applications*. Sage.

Doll, W.J., Xia, W., & Torkzadeh, G. (1994). A confirmatory factor analysis of the end-user computing satisfaction instrument. *MS Quarterly*, *18*(4), 453–461.

Dorans, N.J. (2018). Scores, scales, and score linking. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development2* (pp. 573–605). Wiley Blackwell.

Dowle, M., & Srinivasan, A. (2020). *data.table: Extension of "data.frame."*

Dzinkowski, R. (2000). The measurement and management of intellectual capital: An introduction. *International Management Accounting Study*, 32–36.

el Hamdi, S., Abouabdellah, A., & Oudani, M. (2019). Industry 4.0: Fundamentals and main challenges. *12th International Colloquium on Logistics and Supply Chain Management, LOGISTIQUA 2019*, 1–5. https://doi.org/10.1109/LOGISTIQUA.2019.8907280

Fabrigar, L.R., Wegener, D.T., MacCallum, R.C., & Strahan, E. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Pyschological Methods*, *4*(3), 272–299. https://doi.org/10.1037/1082-989x.4.3.272

Feißt, M., Hennigs, A., Heil, J., Moosbrugger, H., Kelava, A., Stolpner, I., Kieser, M., & Rauch, G. (2019). Refining scores based on patient reported outcomes - statistical and medical perspectives. *BMC Medical Research Methodology*, *19*(1), 167. https://doi.org/10.1186/s12874-019-0806-9

Field, A. (2018). *Discovering statistics using IBM SPSS Statistics* (5th Editio). Sage.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage.

Fink, A. (2010). Survey Research Methods. In P. Peterson, E. Baker, & B. McGraw (Eds.), *International encyclopedia of education* (pp. 152–160). Elsevier.

Fitz-enz, J. (2019). *The ROI of human capital: Measuring the economic value of employee performance.* Amacom.

Fletcher, T.D. (2015). *Package 'QuantPsyc.'* https://cran.r-project.org/web/packages/QuantPsyc

Forza, C., & Filippini, R. (1998). TQM impact on quality conformance and customer satisfaction: A causal model. *International Journal of Production Economics*, *55*(1), 1–20. https://doi.org/10.1016/S0925-5273(98)00007-3

Gilbert, G.E., & Prion, S. (2016). Making sense of methods and measurement : Lawshe's content validity index. *Clinical Simulation in Nursing*, *12*(12), 530–531. https://doi.org/10.1016/j.ecns.2016.08.002

Görmüş, A.Ş. (2009). Entelektüel sermaye ve insan kaynaklari yönetiminin artan önemi [The growing importance of intellectual capital and human resource management]. *Afyon Kocatepe University Journal of Economics and Administrative Sciences*, *10*(1), 57–75.

https://acikerisim.aku.edu.tr/xmlui/handle/AKU/1310

Greenspoon, P.J., & Saklofske, D.H. (1998). Confirmatory factor analysis of the multidimensional Students' Life Satisfaction Scale. *Personality and Individual Differences*, *25*(5), 965–971. https://doi.org/10.1016/S0191-8869(98)00115-9

Hair, J.F., Black, W.C., Babin, B.J., & Anderson, R.E. (2018). *Multivariate data analysis*. Cengage.

Han, Y., & Li, D. (2015). Effects of intellectual capital on innovative performance: The role of knowledge-based dynamic capability. *Management Decision*, *53*(1), 40–56. https://doi.org/10.1108/MD-08-2013-0411

Harris, V., Brett, J., Hirst, S., McClelland, Z., Phizackerley-Sugden, E., & Brown, S. (2007). Using a quality enhancement audit approach to review provision for international students: A case study. In E. Jones & S. Brown (Eds.), *Internationalising higher education* (pp. 95–106). Routledge.

Hayashi, K., & Yuan, K.-H. (2010). Exploratory factor analysis. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 458–465). Sage.

Holland, J., & Holland, J. (2010). Globalization of instruction: Developing intellectual capital. In P. L. Sáez, G. M. de Castro, J. E. N. López, & M. Delgado-Verde (Eds.), *Intellectual capital and technological innovation: Knowledge-based theory and practice* (pp. 39–54). Information Science Reference.

Hooper, D., Coughlan, J., & Mullen, M.R. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, *6*(1), 53–60. https://doi.org/10.21427/D79B73

Hsu, Y.H., & Fang, W. (2009). Intellectual capital and new product development performance: The mediating role of organizational learning capability. *Technological Forecasting and Social Change*, *76*(5), 664–677. https://doi.org/10.1016/j.techfore.2008.03.012

Hu, L.T., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Huang, C.C., Luther, R., & Tayles, M. (2007). An evidence-based taxonomy of intellectual capital. *Journal of Intellectual Capital*, *8*(3), 386-408. https://doi.org/10.1108/14691930710774830

Huang, C.C., Tayles, M., & Luther, R. (2010). Contingency factors influencing the availability of internal intellectual capital information. *Journal of Financial Reporting and Accounting*, *8*(1), 4–21. https://doi.org/10.1108/19852511011055916

Irwing, P., Booth, T., & Hughes, D. J. (2018). *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development*. Wiley Blackwell.

Irwing, P., & Hughes, D.J. (2018). Test development. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 3–48). Wiley Blackwell.

Jakubowska, A., & Rosa, A. (2014). Significance of higher education in creating of intellectual capital. *Human Resources Management & Ergonomics*, *8*, 48–60.

Johnson, R.L., & Morgan, G.B. (2016). *Survey scales: A guide to development, analysis, and reporting*. The Guilford Press.

Jorgensen, T.D., Pornprasertmanit, S., Schoemann, A.M., & Rosseel, Y. (2021). *semTools: Useful tools for structural equation modeling*.

Karakuş, M. (2008). Eğitim örgütlerinde entelektüel sermayenin yönetimi [Management of intellectual capital in educational organizations]. *Milli Egitim*, *178*, 334–349.

Kaya, N., & Kesen, M. (2014). İnsan kaynaklarının insan sermayesine dönüşümü: Bir literatür taraması [Transforming human resources into human capital: A scan of the literature]. *Journal of Academic Researches and Studies*, *6*(10), 23-38. https://dergipark.org.tr/tr/do wnload/article-file/180494

Kelly, A. (2004a). The intellectual capital of schools: Analysing government policy statements on school improvement in light of a new theorization. *Journal of Education Policy*, *19*(5), 609–629. https://doi.org/10.1080/0268093042000269180

Kelly, A. (2004b). *The intellectual capital of schools - Measuring and managing knowledge responsibility and reward: Lessons from the commercial sector*. Springer Science+Business Media, Inc.

Kline, R.B. (2015). *Principles and practices of structural equation modelling*. The Guilford Press.

Kutlu, H.A. (2009). Entelektüel sermaye: Türkiye muhasebe sisteminde raporlanabilir mi? [Intellectual capital: Can it be reported in the accounting system of Turkey?] *Hacettepe University Journal of Economics and Administrative Sciences, 27*(1), 235–257. https://dergipark.org.tr/en/download/article-file/302681

Lane, S., Raymond, M.R., Haladyna, T.M., & Downing, S.M. (2016). Test development process. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3–18). Routledge.

Lawshe, C.H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*, 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Leech, N.L., Barrett, K.C., & Morgan, G.A. (2015). *IBM SPSS intermediate statistics*. Routledge.

Leitner, K.H. (2004). Intellectual capital reporting for universities: Conceptual background and application for Austrian universities. *Research Evaluation*, *13*(2), 129–140. https://doi.org/10.3152/147154404781776464

Li, C.H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936–949. https://doi.org/10.3758/s13428-015-0619-7

Litwin, M.S. (2002). *How to assess and interpret survey psychometrics*. Sage Publications.

Lu, W.M. (2012). Intellectual capital and university performance in Taiwan. *Economic Modelling*, *29*(4), 1081–1089. https://doi.org/10.1016/j.econmod.2012.03.021

Mariani, G., Carlesi, A., & Scarfò, A.A. (2018). Academic spinoffs as a value driver for intellectual capital: The case of the University of Pisa. *Journal of Intellectual Capital*, *19*(1), 202–226. https://doi.org/10.1108/JIC-03-2017-0050

Markus, K.A., & Lin, C. (2010). Construct validity. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 229–233). Sage.

Markus, K.A., & Smith, K.M. (2010). Content validity. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 238–243). Sage.

Martínez-Torres, M. R. (2006). A procedure to design a structural and measurement model of intellectual capital: An exploratory study. *Information and Management*, *43*(5), 617–626. https://doi.org/10.1016/j.im.2006.03.002

Martinez, L.S. (2017). Validity, face and content. In M. Allen (Ed.), *The SAGE encyclopedia of communication research methods* (pp. 1822–1824). Sage Reference.

Martinková, P., & Drabinová, A. (2018). ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. *The R Journal*, *10*(2), 503–515. https://doi.org/10.32614/RJ-2018-074

Matos, F., Vairinhos, V., Selig, P. M., & Edvinsson, L. (2019). *Intellectual capital management as a driver of sustainability: Perspectives for organizations and society*. Springer.

Mohamed, M. (2018). Challenges and benefits of industry 4.0: An overview. *International Journal of Supply Operating Management*, *5*(3), 256-265. https://doi.org/10.22034/2018.3.7

Motta, G. (2017). Factor analysis. In M. Allen (Ed.), *The SAGE encyclopedia of communication* *research methods* (pp. 502–505). Sage Reference.

Mura, M., & Longo, M. (2013). Developing a tool for intellectual capital assessment: An individual-level perspective. *Expert Systems*, *30*(5), 436-450. https://doi.org/10.1111/j.1468-0394.2012.00650.x

Nahapiet, J., & Ghoshal, S. (1998). Social capital, intellectual capital, and the organizational change. *Academy Ol Managemeni Review*, *23*(2), 242–266.

Netemeyer, R.G., Bearden, W.O., & Sharma, S. (2003). *Scaling procedures: Issues and applications*. Sage Publications.

O'Connor, B.P. (2020). *EFA.dimensions: Exploratory factor analysis functions for assessing dimensionality*.

O'Donnell, D., & O'Regan, P. (2000). The structural dimensions of intellectual capital : Emerging challenges for management and accounting. *Southern African Business Review*, *4*(2), 14–20.

OECD/Eurostat. (2018). *Oslo manual 2018: Guidelines for collecting, reporting and using data on innovation*. OECD. https://doi.org/10.1787/9789264304604-en

Osburn, H.G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*, *5*(3), 343–355. https://doi.org/10.1037/1082-989X.5.3.343

Pedrini, M.P. (2007). Human capital convergences in intellectual capital and sustainability reports. *Journal of Intellectual Capital*, *8*(2), 346-366. https://doi.org/10.1108/14691930710742880

Pedro, E. de M., Leitão, J., & Alves, H. (2020). Bridging intellectual capital, sustainable development and quality of life in higher education institutions. *Sustainability*, *12*(2), 1–27. https://doi.org/10.3390/su12020479

Preacher, K.J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research*, *48*(1), 28–56. https://doi.org/10.1080/00273171.2012.710386

R Core Team. (2021). *R: A language and environment for statistical comouting.* R Foundation for Statistical Computing. https://www.r-project.org/

Rae, G. (2007). A note on using stratified alpha to estimate the composite reliability of a test composed of interrelated nonhomogeneous items. *Psychological Methods*, *12*(2), 177–184. https://doi.org/10.1037/1082-989X.12.2.177

Ramírez, Y., & Gordillo, S. (2013). Recognition of intellectual capital importance in the university sector. *International Journal of Business and Social Research*, *3*(4), 27–41. https://doi.org/10.18533/ijbsr.v3i4.27

Ramírez, Y., & Gordillo, S. (2014). Recognition and measurement of intellectual capital in Spanish universities. *Journal of Intellectual Capital*, *15*(1), 173-188. https://doi.org/10.1108/JIC-05-2013-0058

Ramírez Córcoles, Y., & Tejada Ponce, Á. (2013). Cost-benefit analysis of intellectual capital disclosure: University stakeholders' view. *Revista de Contabilidad-Spanish Accounting Review*, *16*(2), 106–117. https://doi.org/10.1016/j.rcsar.2013.07.001

Ramirez, Y., Tejada, A., & Manzaneque, M. (2016). The value of disclosing intellectual capital in Spanish universities: A new challenge of our days. *Journal of Organizational Change Management*, *29*(2), 176–198. https://doi.org/10.1108/JOCM-02-2015-0025

Ren, J.Y. (2009). The empirical study on the relationship between corporate intellectual capital and corporate performance. *IE and EM 2009 - Proceedings 2009 IEEE 16th International Conference on Industrial Engineering and Engineering Management*, 2054–2058. https://doi.org/10.1109/ICIEEM.2009.5344238

Revelle, W. (2020). *psych: Procedures for personality and psychological research*. Northwestern University.

Robitzsch, A. (2021). *sirt: Supplementary item response theory models. R package version*

3.11-21. https://cran.r-project.org/package=sirt

Roos, J., Roos, G., Edvinsson, L., & Dragonetti, N. C. (1997). *Intellectual capital: Navigating in the new business landscape*. Macmillan Business.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2). http://dx.doi.org/10.18637/jss.v048.i02

Rstudio Team. (2021). *RStudio: Integrated development for R.* RStudio. https://www.rstudio.com/

Saint-Ogne, H. (1996). Tacit knowledge the key to the strategic alignment of intellectual capital. *Planning Review*, *24*(2), 10–16. https://doi.org/10.1108/eb054547

Sánchez, M.P., Elena, S., & Castrillo, R. (2009). Intellectual capital dynamics in universities: A reporting model. *Journal of Intellectual Capital*, *10*(2), 307–324. https://doi.org/10.1108/14691930910952687

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*, 23–74.

Schmitt, T.A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment*, *29*(4), 304–321. https://doi.org/10.1177/0734282911406653

Schneider, P. (2018). Managerial challenges of Industry 4.0: an empirically backed research agenda for a nascent field. *Review of Managerial Science*, *12*(3), 803–848. https://doi.org/10.1007/s11846-018-0283-2

Schumacker, R.E., & Lomax, R.G. (2016). *A beginner's guide to structural equation modeling*. Routledge.

Segars, A.H., & Grover, V. (1993). Re-examining perceived ease of use and usefulness: A confirmatory factor analysis. *MIS Quarterly: Management Information Systems*, *17*(4), 517–525. https://doi.org/10.2307/249590

Semenov, V. (2016). Institutional features of strategic management of intellectual capital reproduction. *2015 4th Forum Strategic Partnership of Universities and Enterprises of Hi-Tech Branches (Science. Education. Innovation)*, 52-54. https://doi.org/10.1109/IVForum.2015.7388251

Silva, T.M., & Ferreira, A. (2019). Intellectual capital sustainability in Brazilian public higher education. In F. Matos, V. Vairinhos, P. M. Selig, & L. Edvinsson (Eds.), *Intellectual capital management as a driver of sustainability: Perspectives for organizations and society*. Springer.

Sohrabi, B., Raeesi, I., & Khanlari, A. (2010). Intellectual capital components, measurement and management: A literature survey of concepts and measures. In P. L. Sáez, G. M. de Castro, J. E. N. López, & M. Delgado-Verde (Eds.), *Intellectual capital and technological innovation: Knowledge-based theory and practice* (pp. 1–38). Information Science Reference.

Steiger, J.H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences*, *42*(5), 893–898. https://doi.org/10.1016/j.paid.2006.09.017

Steiner, M.D., & Grieder, S.G. (2020). EFAtools: An R package with fast and flexible implementations of exploratory factor analysis tools. *Journal of Open Source Software*, *5*(53). https://doi.org/10.21105/joss.02521

Subramaniam, M., & Youndt, M.A. (2005). The influence of intellectual capital on the types of innovative capabilities. *Academy of Management Journal*, *48*(3), 450–463. https://doi.org/10.5465/amj.2005.17407911

Suciu, M., & Năsulea, D. (2019). Intellectual capital and creative economy as key drivers for competitiveness towards a smart and sustainable development: Challenges and opportunities for cultural and creative communities. In F. Matos, V. Vairinhos, P. M. Selig, & L. Edvinsson (Eds.), *Intellectual capital management as a driver of sustainability: Perspectives for organizations and society* (pp. 67–97). Springer.

Sultan, P., & Wong, H.Y. (2019). How service quality affects university brand performance, university brand image and behavioural intention: the mediating effects of satisfaction and trust and moderating roles of gender and study mode. *Journal of Brand Management*, *26*(3), 332–347. https://doi.org/10.1057/s41262-018-0131-3

Şencan, H. (2005). *Sosyal ve davranışsal ölçümlerde güvenilirlik ve geçerlilik [Reliability and validity in social and behavioral measures]*. Seçkin.

Tabachnick, B.G., & Fidell, L.S. (2014). *Using multivariate statistics*. Pearson.

Tajvidi, M., & Karami, A. (2015). *Product development strategy: Innovation capacity and entrepreneurial firm performance in high-tech SMEs*. Palgrave Macmillan.

Todericiu, R., & Stanit, A. (2016). Universities intellectual capital. *Management and Economics*, *4*(84), 348–356.

Todericiu, R., & Şerban, A. (2015). Intellectual capital and its relationship with universities. *Procedia Economics and Finance*, *27*(15), 713–717. https://doi.org/10.1016/s2212-5671(15)01052-7

Ullman, J.B. (2014). Structural equation modeling. In B. G. Tabachnick & L. S. Fidell (Eds.), *Using multivariate statistics* (pp. 731–836). Pearson.

Urban, B., & Joubert, G.C.D.S. (2017). Multidimensional and comparative study on intellectual capital and organisational performance. *Journal of Business Economics and Management*, *18*(1), 84–99. https://doi.org/10.3846/16111699.2016.1255990

Watkins, M.W. (2021). A step-by-step guide to exploratory factor analysis with R and RStudio. In *A Step-by-Step Guide to Exploratory Factor Analysis with R and RStudio*. Routledge. https://doi.org/10.4324/9781003120001

Welch, G.W. (2010). Confirmatory factor analysis. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 216–220). Sage.

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *dplyr: A grammar of data manipulation*.

Williams, B., Onsman, A., & Brown, T. (2010). Exploratory factor analysis: A five-step guide for novices. *Journal of Emergency Primary Health Care (JEPHC)*, *8*(3), 1–13. https://doi.org/10.33151/ajp.8.3.93

Wilson, F.R., Pan, W., & Schumsky, D.A. (2012). Recalculation of the critical values for Lawshe’s content validity ratio. *Measurement and Evaluation in Counseling and Development*, *45*(3), 197–210. https://doi.org/10.1177/0748175612440286

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

Youndt, M. A., & Snell, S. A. (2004). Human resource configurations, intellectual capital, and organizational performance. *Journal of Managerial Issues*, *16*(3), 337–360.

APPENDIX

Table A1. *Turkish version of Academic Intellectual Capital Scale.*

Faktör	#	Madde	Hiç doğru değil	Kısmen doğru	Yarı yarıya	Büyük ölçüde doğru	Kesinlikle doğru
Akademik İnsan Sermayesi	1	Üniversitemizde bilimsel araştırmaya odaklanmış güçlü bir akademik kültür vardır.	1	2	3	4	5
	2	Üniversitemizdeki öğretim elemanları, öğrencileri girişimciliğe teşvik eder.	1	2	3	4	5
	3	Üniversitemizdeki öğretim elemanları, yüksek akademik niteliklere sahiptir.	1	2	3	4	5
	6	Üniversitemiz, alanlarının en başarılı öğretim elemanlarına sahiptir.	1	2	3	4	5
	30	Üniversitemiz, özgün fikirleriyle bilinen öğretim elemanlarına sahiptir.	1	2	3	4	5
	31	Üniversitemizdeki öğretim elemanları, öğrencileri ekip çalışması yapmaya teşvik eder.	1	2	3	4	5
Akademik Yapısal Sermaye	4	Üniversitemiz, verilen eğitim içeriğini destekleyecek nitelikte dijital donanıma sahiptir.	1	2	3	4	5
	7	Üniversitemiz, verilen eğitim içeriğini destekleyecek nitelikte bina, donatı, vb. fiziki olanaklara sahiptir.	1	2	3	4	5
	10	Üniversitemizde ihtiyaçlara cevap verecek nitelikte bir bilgi yönetim sistemi (ders seçimi, not takibi vb.) kullanılır.	1	2	3	4	5
	15	Üniversitemiz, ihtiyaca cevap verecek nitelikte bir e-öğrenme platformuna sahiptir.	1	2	3	4	5
	23	Üniversitemizdeki bilgi yönetim sistemi (ders seçimi, not takibi vb.), öğretim elemanları tarafından etkin bir şekilde kullanılır.	1	2	3	4	5
	26	Üniversitemizdeki bilgi yönetim sistemi (ders seçimi, not takibi vb.), öğrenciler tarafından etkin bir şekilde kullanılır.	1	2	3	4	5
Akademik İlişkisel Sermaye	11	Üniversitemizde karar verilirken mezun öğrencilerin fikirleri dikkate alınır.	1	2	3	4	5
	14	Üniversitemizin, iş dünyasında faaliyet gösteren kurumlarla iş birliği protokolleri vardır.	1	2	3	4	5
	16	Üniversitemizin, sektördeki kuruluşlarla imzalanmış mezun işe alım protokolleri vardır.	1	2	3	4	5
	17	Üniversitemizin başka üniversitelerle iş birliği protokolleri vardır.	1	2	3	4	5
	18	Üniversitemiz bünyesinde işlevsel bir teknoloji transfer birimi vardır.	1	2	3	4	5
	19	Üniversitemizde, bilimsel anlayışı topluma yaymaya yönelik etkinlikler düzenlenir.	1	2	3	4	5
	22	Üniversitemizde çevre sorumluluğuna ilişkin etkinlikler düzenlenir.	1	2	3	4	5
	24	Üniversitemiz, yeni iş girişimi (start-up) firmalarını destekler.	1	2	3	4	5

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 1, 165–180

https://doi.org/10.21449/ijate.827950

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

Item Removal Strategies Conducted in Exploratory Factor Analysis: A Comparative Study

Meltem Acar Guvendir[1,*], **Yesim Ozer Ozkan**[2]

[1]Trakya University, Faculty of Education, Department of Educational Sciences, Turkiye
[2]Gaziantep University, Faculty of Education, Department of Educational Sciences, Turkiye

ARTICLE HISTORY

Received: Nov. 18, 2020
Revised: Dec. 07, 2021
Accepted: Jan. 28, 2022

Keywords:
Exploratory Factor Analysis,
Item removal strategies,
Crossloading items,
Construct validity,
Scale development.

Abstract: The aim of this study is to examine how the practice of different item removal strategies during exploratory factor analysis (EFA) phase of scale development change the number of factors, factor loadings, explained variance ratio, and reliability values (α and ω) explained. In the study, data obtained from 379 university students were used for the development of a 46-item scale. As the first item removal strategy, crossloading items on two factors and where the difference between factor loadings was less than .10 were identified. Then, items were removed one by one, starting with the item with the least difference between the loadings on the factors. As the second strategy, the items that loaded on two factors and where the difference between factor loadings was less than .10 were found, and these items were removed from the scale as a whole. As the third strategy, the items that gave high loading on more than two factors and where the difference between these factors was less than .10 were identified. The item removal process was started with these items. The study results show that the factor numbers obtained using three different strategies during the item removal process of EFA were the same; however, the number of items on the scale, the explained variance ratio, and the total scale, and reliability values differed. Furthermore, the items in the factors were not all the same. The study results underscore the importance of theoretical competence in the scale development process.

1. INTRODUCTION

Scales are measuring instruments that outline the criteria to be followed when classifying, sorting, or quantifying variables under investigation by researchers. Scales allow for the regulation of the data's quality. According to Netemeyer et al. (2003), while scale development continues to be a popular issue in the social sciences, scholars have proposed a variety of approaches to scale development. Murphy and Davidshofer (2005) divided the scale development process into three stages: instrument construction, instrument standardization, and instrument revision/updating. According to Clark and Watson (1995), scale development is a process that entails clearly defining the target construct, constructing an item pool, testing the pool's items on a representative sample, conducting correlation and factor analysis, and examining the dimension and discrimination validity. Crocker and Algina (1986) state that the scale development process involves determining the purpose of the scale scores, describing the

[*]CONTACT: Meltem ACAR GUVENDIR ✉ meltemacar@gmail.com 🖃 Trakya University, Faculty of Education, Department of Educational Sciences, Turkiye

e-ISSN: 2148-7456 **/© IJATE** 2022

behaviors, preparing the indicator table, creating the item pool, editing the items, administering pre-tests and pilot tests, item analysis, calculating validity and reliability, and standardization steps. Along with the phases outlined by Crocker and Algina (1986), Erkuş (2012) noted that the scale development process should include the following: selection of the scale development technique; written explanations and instructions; and, if required, repeating the applications. Although the process of scale construction is described differently, the fundamental objective is to generate a valid measure of the underlying psychological construct (Clark & Watson, 1995).

The term validity refers to the degree to which a measuring instrument accomplishes its objective. Thus, the measuring instrument is anticipated to measure only the characteristic it is intended to measure. According to Tavsancil (2002), the variables that contribute to the erosion of validity are linked to the development and implementation of scales. Construct validity refers to whether the scale measures precisely what it is intended to measure or whether the items on the scale accurately represent the theoretical or psychological construct (Erkuş et al., 2017). As Messick (1981) phrased it, construct validity is undoubtedly the core of validity. Factor analysis can be used to ascertain the scales' construct validity (Cronbach & Meehl, 1955). The determination of whether a test measures the intended or anticipated construct is a form of validity problem, which may be resolved using factor analytical methods (Stapleton, 1997).

Spearman pioneered the application of factor analysis in the early twentieth century (Ford, MacCallum & Tait, 1986). He claimed that intelligence had a one-factor structure by demonstrating that the individual's varied mental operations had a similar feature. In 1927, Spearman published his own study, "Talents of Man", in which he asserted that intelligence cannot be described by a single factor using the factor analysis technique. Rather than that, it was composed of two factors: general and specific. Following research that refined Spearman's approach concluded that multiple factors should be involved in the functioning of a complex phenomenon like intelligence (Özgüven, 1994). As a result, many researchers began to employ Spearman's approach in the subsequent time, and it developed into a tool for demonstrating the construct validity of data collecting instruments used, particularly in the domains of education and psychology.

Factor analysis is an important statistical operation in the social sciences since it elucidates the quality and validity of measurement. The primary objective of factor analysis is to reduce the number of dimensions (Brown, 2009). As Kerlinger (1979) expressed it, factor analysis is "one of the most powerful methods yet for reducing variable complexity to greater simplicity" (p. 180). Additionally, factor analysis confirms the scale's capacity to assess the construct being measured. Numerous statistical processes such as regression, correlation, discriminant analysis, and difference tests rely on factors extracted from the original data set (Albayrak, 2006). Factor analysis is used to uncover the invisible and immeasurable dimensions concealed underneath the numerous observable and measurable aspects (Johnson & Winchern, 2002). There are two types of factor analysis: exploratory and confirmatory factor analysis. The aim of exploratory factor analysis (EFA) is to identify variables based on their correlations (Kline, 2011; Stewens, 1996; Tabachncik & Fidell, 2001). A previously established hypothesis is tested using confirmatory factor analysis (CFA) based on the correlations between variables (Kline, 2011; Stewens, 1996; Tabachncik & Fidell, 2001). In other words, this method is a procedure for producing a latent variable (factor) using observed variables in a previously built model. It is frequently used in the construction of scales and validity studies, as well as to validate a preconceived construct.

To ensure the measuring tool's validity, each item must measure a single behavior. Therefore, in scale development studies, during the factor creation process using EFA, if an item's level of relationship with more than one factor is larger than the level of relationship with the other factor, the item should be counted under the factor with the higher level of relationship. According to Tabachnick and Fidell (2001), .32 is a reasonable rule of thumb for the bare minimum loading of a factor item, which corresponds to around 10% crossloading variation with the variance of the other factor items. A "crossloading" item has a loading factor of.32 or more on two or more variables concurrently. When considering whether to exclude a crossloading item from the study, the researcher should consider if there are sufficient strong loaders (.50 or greater) on each component to support the elimination. When there is significant crossloading, it is probable that the items were created insufficiently or that the a priori factor structure was flawed. According to Çokluk et al. (2010), crossloading items are those that have a high loading on more than one component and a difference of less than.10 between these loadings. According to Can (2018), the difference between factor loadings can be regarded as 0.15 if removing the items from the scale does not pose significant problems. To be more precise, these items do not measure a single behavior. As a result, it is critical for construct validity to exclude elements that have an approximate loading on more than one factor (Can, 2016; Kline, 2011; Stewens, 1996; Tabachnick & Fidell, 2001).

The values presented in these studies take a unique approach to item removal. When the examples given by Can (2016), Tabachnick and Fidell (2001), and Büyüköztürk (2007) are examined, the EFA eliminates all crossloading components. In his technical report on factor analysis, Samuels (2017) suggested that the item elimination process should be repeated until no crossloading items remain. Çokluk et al. (2010) discuss in additional detail whether the removal procedure will be carried out sequentially or totally via crossloading. They said that while doing EFA on a single factor, crossloading items must be deleted individually throughout the item extraction procedure. However, there is no definitive rule on whether crossloading elements should be eliminated from the analysis individually or entirely in an EFA for a multidimensional construct. Due to the fact that this condition will vary depending on the measured construct, the researcher will make this conclusion. If a researcher wishes to eliminate an item, it is advantageous to do so from the most crossloading to the least crossloading item (i.e., the item with the smallest difference between the two factor loadings). Raubenheimer (2004) proposed an alternative method for item removal in EFA. In his investigation, he began by removing elements that made a low contribution to the scale's reliability and kept removing items until the reliability value remained constant. He then switched to EFA when the reliability value remained constant. He concluded the procedure by eliminating items that had a high loading on more than one factor in the factor analysis. He then used direct EFA without removing low-reliability items as a second way. When he compared the results of the two methods, he concluded that the scale had a higher reliability value when EFA was performed after the items with low reliability were removed. Additionally, some items retained by the first method were eliminated by the second one. However, the second method required the removal of more items to enhance validity. As a result, researchers do not agree on the appropriate method for removing items. On the other hand, scale development studies in the literature frequently do not specify how to eliminate items during EFA.

Firstly, from which item on the scale should we begin the item removal? As mentioned above, the first strategy is to remove the items one by one, starting from the item that gives a high loading on two factors and where the difference between factor loadings is the smallest. The analysis is repeated after each item is removed from the test. The analysis is completed when no crossloading item is left. As a second strategy, all the crossloading items that give a high load on two factors and where the difference between factor loadings is less than .10 are determined, and these items are excluded from the analysis. Once the crossloading items are removed from the test, the analysis is repeated once again. After the analysis, if there are still crossloading items, they are all removed from the test and the analysis is repeated once more. In addition to these two options, as a third strategy, the items that give a high load on more than two factors and where the difference between these loadings is less than .10 are excluded from the test. After removing the items that give high loadings on more than two factors, the items are removed from the analysis one by one, starting with the item that gives high loading on two factors, and the difference between the loadings is the least. The analysis is repeated after each item is removed from the analysis. The analysis is terminated when no crossloading item remains. The reason for the item removal process starting with items that impose high loadings on more than two factors is that these items largely weaken the principle that each item should measure a single behavior.

Due to the fact that these three procedures will provide differing results, it is critical to examine the explained variance ratio acquired from each of them. According to Tabachnick and Fidell (2001), the variance ratio is calculated by dividing the sum of a factor's item's factor loading squares by the factor's overall item count. The high variance ratio explained demonstrates the designed scale's factor structure's robustness (Gorsuch, 1983). Scherer, Luther, Wiebe, and Adams (1988) state that the variance ratio in the social sciences should be between 40% and 60%. Büyüköztürk (2007), on the other hand, underlined that the explained variance ratio for unidimensional scales should be 30% or above, but it should be higher for unidimensional scales. As such, it is of importance to determine the explained variance ratio achieved by employing three distinct procedures during scale development research while removing items through EFA. There are a few studies in the literature (e.g. Raubenheimer, 2004) that use a variety of different EFA strategies on the same data set and report comparable results. There is no study in the literature that compares the three aforementioned strategies. Thus, this study will shed light on the literature, particularly on scale development studies, by comparing the factor numbers, loadings, explained variance ratio, and reliability values derived from three distinct item removal procedures used in EFA.

The purpose of this study is to examine how the use of different strategies to remove items during EFA in scale development studies change the number of factors, loadings, explained variance ratio and reliability values (α and ω) explained. For this general purpose of the research, answers to the following questions were sought:

1. What are the factor numbers, loadings, explained variance ratio, communality values, and reliability values obtained when items are removed starting from the most crossloading item to the least crossloading one while conducting EFA in the scale development studies?
2. What are the factor numbers, loadings, explained variance ratio, communality values, and reliability values when all the crossloading items are removed as a whole during EFA in scale development studies?
3. What are the factor numbers, loadings, explained variance ratio, communality values, and reliability values obtained when items are removed, starting from the items that load on more than two factors and where these loadings are close to each other during EFA in the scale development studies?

2. METHOD

2.1. Participants

The study collected data from 379 university students to develop a 46-item scale. The participants are enrolled at a university. While 70.3% of participants are female, 29.7% are male.

2.2. Data Analysis

While conducting EFA, the study compared the results obtained from three different item removal strategies. Firstly, data were tested in terms of EFA's assumptions. These are: the presence of missing data and extreme values, the adequacy of the sample size, and the suitability of the data to multivariate normality, whether the items are sufficiently correlated, and multicollinearity.

Before proceeding to EFA, it is essential to review the dataset for missing data. There were no missing data in this study's data set. Following that, the study looked for the existence of extreme values. Since the purpose of this study was to determine the number of factors based on the items, Mahalanobis distances were used to find the presence of multivariate extreme values. The Mahalanobis distance measure shows a subject's distance from the centroid, which is obtained using the means of all variables (Tabacknick & Fidell, 1996). By examining Mahalanobis distances, it was determined that 96 subjects could be considered extreme values in this study, and these subjects were eliminated from the analysis.

Numerous viewpoints have been expressed on the appropriate sample size for factor analysis. There are those who argue that the sample size should be determined by reducing the number of items and factors (Bryman & Cramer, 2001; Kline, 1994; Kass & Tinsley, 1979; Nunally, 1978), while others argue that it should be determined using absolute criteria (Comrey & Lee, 1992; Kline, 1994; Tabachnick & Fidell, 2001). According to Kline (1994), a sample size of 200 individuals is often adequate for factor analysis as an absolute criterion, although this number can be dropped to 100 when the number of factors is small yet open. Comrey and Lee (1992) claimed that 50 is a very small sample size, 200 is a medium sample size, 300 is a decent sample size, 500 is a very good sample size, and 1000 is the ideal sample size. Tabachnick and Fidell (2001) state that at least 300 participants are required for EFA. On the other hand, MacCallum et al. (1999) said that sample size is dependent on the characteristics of the data acquired, implying that precise sample size decisions are challenging. According to them, if communalities are strong and each component can be described by four or more items, the sample size can be small; but, if communalities are low, a large sample size will be required. In this case, reaching the largest possible sample is the best way for factor analysis since it cannot be known how high the communalities will be without analysis. The sample size was kept large to ensure that the EFA assumptions were met, and the Kayser Mayer Olkin (KMO) test was employed to assess the sample size's sufficiency. The calculated value was .839. Leech, Barrett, and Morgan (2005) claimed that a KMO value between 0.50-0.60 was insufficient, a value between 0.60-0.70 was poor, a value between 0.70-0.80 was moderate, a value between 0.80-0.90 was good, and a value over 0.90 was exceptional. As a consequence, the sample size in this study was sufficient for EFA based on the KMO value obtained.

The Bartlett's Test of Sphericity examines if the true correlation matrix differs significantly from the unit matrix. If the p value for this test is less than .05, it shows that the matrix of relationships between the items is different from the unit matrix without relations (Can, 2016). Due to the significance of the acquired value (.000), there is a significant difference between the true correlation matrix and the unit matrix in the current investigation.

Additionally, the data should have a multivariate normal distribution. To establish a multivariate normal distribution, the observations in the sample must exhibit a normal distribution across all variable combinations (Çokluk et al., 2010). Mardia's test can be used to check whether the data fit the multivariate normal distribution properly (Mardia, 1970). Due to the significance of Mardia's test result, it was found that the data did not follow the multivariate normal distribution assumption.

Additional measures are available to ascertain whether the items are sufficiently correlated. The first is the anti-image. The anti-image denotes the proportion of variance in an item that is unrelated to another item in the analysis. Obviously, all items should be highly correlated, in order to minimize an item's anti-images. The anti-image matrices address this issue (Sarstedt & Mooi, 2014). Correlations between anti-images are the inverse of partial correlations. In other words, it reflects the pairwise correlation that remains after other variables' influence is subtracted. A good factor solution is characterized by shared variance/covariance that extends beyond individual pairs of variables to a larger collection of variables. Anti-image correlation matrix diagonals should be greater than 0.5, which is associated with smaller off-diagonal partial pairwise correlations (Can, 2016; Costello & Osborne, 2005; Hauben et al., 2017; Spicer, 2005). In this study, the anti-image correlation matrix's diagonals were all greater than .5 (between .830 and .968). The aggregate data screen indicates that the situation is acceptable.

The data matrix's interrelationships should be sufficient. If the correlation coefficient is less than .33 as a result of an observation to be made in the correlation matrix, no factor analysis will be performed (Can, 2016). Given that the inter-item relationships were found to be between .377 and .684, it was concluded that the number of items with acceptable inter-item relationships was quite high. Additionally, the fact that the matrix's determinant is greater than .0001 indicates the possibility of factor analysis.

Another assumption of EFA is that there should be no problem of variable multicollinearity. To check for multicollinearity, tolerance and VIF values were examined in this direction. Tolerance values were found to range between .241 and .460, and VIF values were found to range between 2.423 and 4.151. Thus, tolerance values greater than .10 (Field, 2005; Mertler & Vannatta, 2005) and VIF values less than 10 (Albayrak, 2005), indicate that multicollinearity is not a concern.

The subsequent step was analysis, as the data set fulfilled th EFA's assumptions. During the EFA process, items were removed in ascending order from the most crossloading to the least crossloading. Then, crossloading items were found and eliminated, followed by a repetition of the analysis. The items were removed starting from the items that gave high load close to each other on more than two factors. The results section includes the factor numbers, loadings, and explained variance ratio derived using the three item elimination procedures. Direct oblimin was used as the factor rotation method in this procedure. Given the fact that oblique rotation creates a pattern matrix including the factor or item loadings, as well as a factor correlation matrix including the factor correlations. Oblique rotation methods such as Direct Oblimin and Promax are prevalent. Direct Oblimin aims to simplify the output's structure and mathematics (Gorsuch, 1983). As a result, the direct oblimin method was selected throughout the factor analysis procedure.

Unweighted Least Squares (ULS) was employed in the factor analysis process. Comrey (1962) developed unweighted least squares analysis in order to minimize the squares of the differences between observed and reproduced correlation matrices. Additionally, it is a subset of principal factor analysis in that it estimates the variance of common factors following analysis (Tabachnick & Fidell, 2001). This technique was used since the data in the research were categorical and violated the multivariate normality assumption.

Additionally, the correlation matrix employed in factor analysis might be vary depending on the number of categories in the scales. As the number of categories rises, the data may be regarded continuous and the Pearson correlation matrix can be used to do analyses. Finney and DiStefano (2013) claim that a data set is deemed continuous if it has six or more categories. Tabachnick and Fidell (2001) state that a data set is considered continuous if it contains seven or more categories. The tetrachoric correlation matrix, on the other hand, is used to analyze data with two categories, whilst the polychoric correlation matrix is used to study data with three, four, or five categories.

Correlation calculated to explain the relationships between unobserved variables is known as polychoric correlation (Basto & Pereira, 2012). Correlations are classified as polychoric if they are based on the premise that ordinal variables have a common continuous distribution (Ekström, 2011). When dealing with ordinal data, factor analysis should be conducted on the raw data matrix of polychoric correlations, not on Pearson correlations (Basto & Pereira, 2012). Because there were five categories in this study, factor analysis was done using the polychoric correlation matrix (5-point Likert scale-values between 0 and 4).

In factor analysis, when deciding on the number of factors, the eigenvalues-greater-than-one, the communalities explained by the items greater than .2, scatter plot (Cattell, 1966), the minimum average-partial correlation (minimum-average partial correlation), Bartlett's χ2 test (Bartlett, 1950), RMSEA-based maximum-likelihood method (Park et al., 2002), parallel analysis (Horn, 1965), and Velicer’s Minimum Average Partial Test (MAP) (Velicer, 1976) can be used. It should be noted, however, that each of these procedures produces a distinct factor number. Additionally, Hayton et al. (2004) state that examining eigenvalues is a commonly used approach for determining the number of factors. However, this technique is insufficient for capturing the true factor structure of the data since it overestimates the number of latent factors. Additionally, the scatter plot is subjective, as it is constructed based on the researcher's observation. As for Bartlett's χ^2 test, this test is based on χ^2 and is therefore sensitive to sample size. Although the RMSEA based maximum likelihood method is relatively accurate, it can only be used when the maximum likelihood estimator is used. Parallel analysis and the MAP test were used to determine the number of factors in this study, as well as communality and eigenvalues. The factor program and JASP were used for the analysis of this research.

3. RESULT / FINDINGS

To begin, parallel analysis and the MAP test were used to determine the number of factors in the data. While the parallel analysis indicated a two-factor structure, the MAP indicated a three-factor structure. Additionally, the eigenvalues were examined, and it was observed that four factors had eigenvalues greater than one. Table 1 contains the eigenvalues of the four components.

Table 1. *Eigenvalues.*

Variable	Eigenvalue	Proportion of Variance
1	26.386	.574
2	2.791	.061
3	1.881	.041
4	1.351	.029

According to the parallel analysis, there is a structure with two factors, three factors according to the MAP test, and four factors when the eigenvalues are examined. As a result, a three-factor structure was chosen.

In EFA, the item elimination procedure began with the most crossloading item and progressed to the least crossloading item. Table 2 shows the factor loadings for the items obtained prior to the item removal operation. According to Table 2, the 28th item has the highest crossloading, since the difference between the factor loadings is the smallest. As a result, the item removal procedure began with this item. After removing each item, the analysis was repeated. When no crossloading items remained, the analysis was concluded. The obtained factor loadings are summarized in Table 3.

Table 2. *Factor loadings for the items obtained before the item removal process using the first strategy.*

Variable	Factor 1	Factor 2	Factor 3
V1		.682	
V2		.823	
V3		.773	
V4		.745	
V5		.820	
V6		.737	.384
V7		.908	
V8		.853	
V9		.553	
V10		.682	
V11		.832	
V12		.587	
V13		.740	
V14		.755	
V15		.807	
V16		.761	
V17	.461	.384	
V18	.378	.462	
V19	.327	.369	.384
V20	.413	.434	
V21	.411	.467	
V22	.317	.474	
V23	.413	.428	
V24	.369	.411	
V25		.395	.459
V26		.537	
V27	.372	.332	
V28	.451	.465	
V29	.403		.515
V30	.576	.303	
V31	.680		
V32	.653		
V33	.739		
V34	.587		.362
V35	.672		
V36	.733		.419
V37	.820		
V38	.903		
V39	.774		
V40	.661		.413
V41	.843		
V42	.794		
V43	.794		
V44	.697		.361
V45	.815		
V46	.763		

*Factor loadings below .30 are not considered (Kline, 1994) and are not shown in the table.

** While the item shown in red in the table has a higher loading than one in all three factors, the item shown in green shows the most crossloading item and the items shown in blue show the other crossloading items. In addition, all items shown in blue, red, and green are crossloading. In the first method, the item throwing process began with the item shown in green, and in the third method, it began with the item shown in red. During the item removal process, all blue, red, and green items were discarded in the second method.

Table 3. *Factor loadings for the items obtained using the first strategy.*

Variable	Factor 1	Factor 2
V1	.796	
V2	.889	
V3	.814	
V4	.642	
V5	.881	
V6	.441	
V7	.905	
V8	.841	
V9	.703	
V10	.707	
V11	.740	
V12	.789	
V13	.645	
V14	.704	
V15	.846	
V16	.736	
V18	.594	
V20	.737	
V21	.761	
V22	.537	
V26	.694	
V29		.704
V31	.333	.536
V33	.550	.312
V36		.935
V37		.689
V38		.661
V39		.732
V40		.903
V41		.728
V42		.639
V43		.603
V44		.862
V45	.355	.499

*Factor loadings below .30 are not considered (Kline, 1994) and are not shown in the table. In addition, at least .40 factor loadings (Pett, Lackey, & Sullivan, 2003) were taken as basis for items to be placed under one factor.

According to Table 3, when the item removal process was carried out towards the least crossloading item starting from the most crossloading one, a two-factor structure consisting of 34 items was obtained. Communality values of the items are between .431 and .770. The analysis was repeated 11 times as the analysis was performed after removing each item. The factor loadings of the items are between .441 and .935. The explained variance ratio was found to be 64.844. The explained variance ratio according to the factors is 57.296, 7.547, respectively. There are 22 items in factor 1 and 12 items in factor 2.

Secondly, all crossloading items were identified and all of them were removed as a whole. The item removal procedure was completed by completely deleting crossloading items. The analysis was repeated until there were no remaining crossloading items. Table 4 details the factor loadings of the items.

Table 4. *Factor loadings for the items obtained using the second strategy.*

Variable	Factor 1	Factor 2
V1	.763	
V2	.899	
V3	.786	
V4	.720	
V5	.870	
V6	.622	
V7	.939	
V8	.841	
V9	.625	
V10	.652	
V11	.763	
V12	.682	
V13	.682	
V14	.702	
V15	.797	
V16	.711	
V22	.447	.325
V26	.567	
V30		.603
V31		.713
V32		.634
V33		.589
V35		.629
V37		.875
V38		.921
V39		.862
V40		.732
V41		.906
V42		.826
V43		.810
V45		.721
V46		.627

*Factor loadings .30 are not considered (Kline, 1994) and are not shown in the table.

Identifying all crossloading items and removing all of them as a whole was used as the second strategy during the item removal of EFA. Ten crossloading items were detected in the first step, and the study was redone once these items were eliminated. Two crossloading items were excluded from the analysis in the second step. One crossloading item was deleted in the third and fourth rounds. When the analysis was repeated, the results revealed a two-factor construct with 32 items. The items had communality values ranging from.440 to.794. Factor loadings for the items ranged from .447 to .939. The explained variance ratio was determined to be 65.460.

The explained variance ratio for the factors is 58.639 and 6.821, respectively. Factor 1 has 18 items, whereas Factor 2 contains 14 items.

As the last and third item removal strategy, items were deleted starting with those that had a high load on more than two factors and were situated close to one another. The process of item removal began with this crossloading item. Following that, the 25th item was deleted since it loaded on three factors again, and the item removal procedure was finished by deleting each item one by one, starting with the most crossloading item and ending with the least crossloading item. The analysis was repeated nine times until there were no remaining crossloading items. Factor loadings of the items are given in Table 5.

Tablo 5. *Factor loadings for the items obtained using the third strategy.*

Variable	Factor 1	Factor 2
V1	.791	
V2	.901	
V3	.811	
V4	.667	
V5	.879	
V6	.496	
V7	.914	
V8	.836	
V9	.674	
V10	.687	
V11	.741	
V12	.759	
V13	.650	
V14	.693	
V15	.823	
V16	.715	
V18	.534	
V20	.670	
V21	.690	
V22	.504	
V27		.461
V30	.351	.544
V31		.626
V32	.326	.506
V35	.326	.544
V36		.928
V37		.802
V38		.775
V39		.794
V40		.868
V41		.808
V42		.709
V43		.694
V44		.846
V45		.596
V46	.315	.494

*Factor loadings values below .30 are not considered (Kline, 1994) and are not shown in the table. In addition, at least .40 factor loadings (Pett, Lackey, & Sullivan, 2003) were taken as basis for items to be placed under one factor.

Finally, the items were removed starting from the items that gave high load close to each other on more than two factors. Subsequently, after the items that gave high loadings close to each other on more than two factors (where the difference between factor loadings was at least .10) were removed from the test, the items with the lowest difference between the factor loadings giving high load on two factors were determined and these items were excluded from the test one by one. As a result, the analysis was repeated nine times and a two-factor structure consisting of 36 items was obtained. The communality values of the items were between .421 and .781. The explained variance ratio was found to be 64.418. The explained variance ratio according to the factors was 57.422 and 6.996. The factor loadings were between .461 and .928. There are 20 items in factor 1 and 16 items in factor 2.

Cronback Alfa and McDonald's Omega reliability values according to the total items and factors with each method of the scale are given in Table 6.

Table 6. *Cronbach alfa and McDonald's omega values.*

	1st method		2nd method		3rd method	
	α	ω	α	ω	α	ω
Factor 1	.962	.962	.952	.952	.958	.959
Factor 2	.943	.944	.953	.953	.956	.957
Total	.971	.971	.971	.971	.973	.973

*Reliability values were written according to the factor order obtained according to the 1st, 2nd and 3rd methods

Table 6 shows that α and ω values are close to each other according to the factors and the scale obtained by using all the three strategies. The α and ω values of the scale obtained in the 3nd strategy are higher than the values obtained in the 1st and 2nd strategies.

4. DISCUSSION and CONCLUSION

EFA is used to reveal the structures of variables that are composed of different components, whose structure is not fully known, but whose existence is also present (Can, 2016). Each item in a measurement tool must measure a single feature. This is an indispensable rule for validity. Therefore, in EFA, crossloading items that give high loading on more than one factor and where the difference between these loadings is less than .10 are removed from the test (Can, 2016; Çokluk et al., 2010; Kline, 2011; Stewens, 1996; Tabachncik & Fidell, 2001). Various procedures are used to identify and eliminate these crossloading elements. Relevant research illustrating the processes of EFA demonstrates that items are removed using a variety of strategies (Büyüköztürk, 2007; Can, 2016; Çokluk et al., 2010; Tabachncik & Fidell, 2001). Accordingly, this research examined how the use of different item removal strategies during EFA in scale development studies changed the number of factors, loadings, variance ratio and reliability values (α and ω) explained.

The study's results suggest that while the factor numbers produced using three distinct procedures throughout the item removal phase of EFA were identical, the scale's item number, explained variance ratio by the factors and the overall scale, and reliability values varied. Additionally, the items in the factors were not identical. For instance, an item that remained in the scale following the second strategy was removed during the first item removal strategy. Additionally, the second strategy had the highest explained variance rate. When reliability values were analyzed, the third strategy produced the greatest values for both Cronbach's alpha and omega coefficients. When the number of items is considered, the third approach yielded the most items. The outcomes of all three strategies exhibit a high degree of resemblance in terms of factor loadings and communality values. The second strategy resulted in the highest factor loading and communality values.

The study's findings emphasize the importance of researchers having a firm grasp of the theoretical framework when determining which item removal approach to use throughout the EFA. In particular, the item selection and removal strategy should be matched to the objective of scale development and to the theoretical conceptualization of the target construct. In their groundbreaking studies on measurement and validity, Cronbach and Meehl (1955) and Loevinger (1957) stated the prominence of theory in measurement and stressed that the latent construct should be grounded in a theoretical framework. A well-grounded theory begins with conceptualizations based on a thorough review of the literature which serves two important purposes (Netemeyer et al., 2003). First, such a review will serve to clarify the nature and range of the content of the target construct. Second, a study of the literature may assist in identifying shortcomings with existing measures and determining whether the suggested scale is indeed necessary (Clark & Watson, 1995). Additional crossloading items can be reviewed and revised if necessary. The findings of the current research show that theoretical knowledge and literature review may serve a third purpose in scale development by influencing researchers' decisions on item removal strategies that they will follow during EFA.

EFA is an analytical statistical method, and in EFA, the process of removing items from the test proceeds mechanically. As a result, if the researcher is theoretically competent about the construct being measured, s/he will also be competent in the item removal process. What the researchers should aim for is that the structure in the theoretical framework overlaps with the data at hand (Tabachnick & Fidell, 2001). Henson and Roberts (2006) stated that very few researchers considered the expected number of factors in a theory. According to them, even though factor analysis is an important process in determining the number of factors in scale development, the theoretically recommended and expected number of factors is also very important. Consequently, the items include pieces of a theory, and EFA is done to check whether the items conform to the theory or not.

Bornstein (1996) suggests that before proceeding to EFA, the researcher should review the items and determine if possessing each item is indeed theoretically relevant. Additionally, the researcher should indicate the degree to which a theoretically significant item adequately describes the conceptual framework. Regarding the item removal procedure in factor analysis, Ziegler (2014) noted that the researcher's theoretical underpinning for the construct being examined comes before the methodologies used during item removal. The initial objective is to establish a theoretical framework for the construct to be examined, and statistical approaches become more significant after the theoretical framework is established. As a consequence, there may be items that the researcher believes are critical in explaining the scale's structure. As a result, experimenting with alternative methods of removing the item may prevent such items from being instantly removed from the test. Even if researchers choose an appropriate approach, they should analyze the items they seek to remove first and then establish whether the item is acceptable for the psychological structure being described. In this instance, researchers with a firm grasp of the theoretical background can select the most appropriate strategy for item removal during EFA.

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). **Ethics Committee Number:** Trakya University, 19.01.2022 - 01/18.

Authorship Contribution Statement

Meltem Acar Guvendir: Methodology, Investigation, Resources, Visualization, Software, Formal Analysis, and Writing -original draft. **Yesim Ozer Ozkan:** Introduction and discussion, Writing -original draft, and Validation.

Orcid

Meltem Acar Guvendir https://orcid.org/0000-0002-3847-0724
Yeşim Ozer Ozkan https://orcid.org/0000-0002-7712-658X

REFERENCES

Albayrak, A.S. (2006). *Uygulamalı çok değişkenli istatistik teknikleri* [*Applied multivariate statistical techniques*]. Asil Yayın Dağıtım.

Albayrak, A.S. (2005). Çoklu doğrusal bağlantı halinde en küçük kareler tekniğinin alternatifi yanlı tahmin teknikleri ve bir uygulama [*Alternative to the minimum square technique: a multi-linear connection balanced estimating techniques and an application*]. *Zonguldak Karaelmas Üniversitesi Sosyal Bilimler Dergisi, 1*(1), 105-126.

Bartlett, M.S. (1950). Tests of significance in factor analysis. *British Journal of Psychology, 3,* 77-85.https://doi.org/10.1111/j.2044-8317.1950.tb00285.x

Basto, M., & Pereira, J.M. (2012). An SPSS R-Menu for ordinal factor analysis. *Journal of statistical software, 46*(4), 1-29. https://doi.org/10.18637/jss.v046.i04

Bornstein, R.F. (1996). Face validity in psychological assessment: Implications for a unified model of validity. *American Psychologist, 51*(9), 983-984. https://doi.org/10.1037/0003-066X.51.9.983

Brown, J.D. (2009). Statistics Corner Questions and answers about language testing statistics: Principal components analysis and exploratory factor analysis, In. Definitions, differences, and choices. *Shiken: JALT Testing & Evaluation SIG Newsletter, 13*(1), 19 - 23. https://hosted.jalt.org/test/PDF/Brown30.pdf

Büyüköztürk, Ş. (2007). Veri Analizi El Kitabı [*Data analysis handbook for social sciences*]. Ankara: Pegem Yayınları.

Bryman, A., & Cramer, D. (2011). *Quantitative data analysis with IBM SPSS 17, 18 and 19: A guide for social scientists.* Routledge-Cavendish/Taylor & Francis Group.

Can, A. (2016). *SPSS ile bilimsel araştırma sürecinde nicel veri analizi* [*Quantitative data analysis in the process of scientific research with SPSS*]. Ankara: Pegem Akademi.

Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245-276. https://doi.org/10.1207/s15327906mbr0102_10

Clark, L.A., & Watson, D. (2016). Constructing validity: Basic issues in objective scale development. In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (pp. 187-203). American Psychological Association. https://doi.org/10.1037/14805-012

Comrey, A.L. (1962). The minimum residual method of factor analysis. *Psychological Reports, 11*(1), 15-18. https://doi.org/10.2466/pr0.1962.11.1.15

Comrey, A.L., & Lee, H.B. (1973). *A first course in factor analysis.* Academic Press.

Costello, A.B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research, and evaluation, 10*(7), 1-9. https://doi.org/10.7275/jyj1-4868.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. CBS College Publishing.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957.

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2010). *Sosyal bilimler için çok değişkenli istatistik* [*Multivariate statistics for social sciences*]. Pegem Akademi.

Erkuş, A. (2012). *Psikolojide ölçme ve ölçek geliştirme* [*Measurement and scale development in psychology*]. Pegem Akademi Yayınları.

Erkuş, A., Sünbül, Ö., Sünbül, S.Ö., Yormaz, S., & Aşiret, S. (2017). *Psikolojide ölçme ve ölçek geliştirme II* [*Measurement and scale development in psychology II*]. Pegem Akademi.

Ekström, J. (2011). A Generalized Definition of the Polychoric Correlation Coefficient. *UCLA: Department of Statistics, UCLA*. https://escholarship.org/uc/item/583610fv

Field, A. (2005). *Discovering statistics using SPSS.* (2nd ed.). London: Sage

Finney, S.J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. G. R. Hancock ve R. O. Mueller (Ed.), *Structural equation modeling: A second course* (2nd ed., 439– 492). Charlotte.

Ford, J.K., MacCallum, R.C., & Tait, M. (1986). The Application of exploratory factor analysis in applied psychology: A Critical review and analysis. *Personnel Psychology, 39*(2), 291-314. https://doi.org/10.1111/j.1744-6570.1986.tb00583.x

Gorsuch, R.L. (1983). *Factor analysis* (2nd ed.). Hillside, NJ: Lawrence Erlbaum Associates

Hauben, M., Hung, E., & Hsieh, W.Y. (2017). An exploratory factor analysis of the spontaneous reporting of severe cutaneous adverse reactions. *Therapeutic advances in drug safety*, *8*(1), 4-16. https://doi.org/10.1177/2042098616670799

Hayton, J.C., Allen, D.G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods, 7*(2), 191-205. https://doi.org/10.1177/1094428104263675

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. https://doi.org/10.1177/0013164405282485

Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179-185. https://doi.org/10.1007/BF02289447

Johnson, R.A., & Wichern, D.W. (2002). *Applied multivariate statistical analysis.* Upper Saddle River.

Kass, R.A., & Tinsley, H.E.A. (1979). Factor analysis. *Journal of Leisure Research, 11,* 120-138. https://doi.org/10.1080/00222216.1979.11969385

Kerlinger, F.N. (1979). *Behavioral research: A conceptual approach.* Rinehart & Winston.

Kline, P. (1994). *An easy guide to factor analysis.* Routledge.

Kline, R. B. (2011). *Principles and practice of structural equation modeling (3rd Edition).* The Guilford Press.

Leech, N.L., Barrett, K.C., & Morgan, G.A. (2005) *SPSS for Intermediate Statistics, Use and Interpretation. 2nd Edition*. Lawrence Erlbaum.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*(3), 635–694. https://doi.org/10.2466/pr0.1957.3.3.635

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

Mardia, K.V. (1970). Measures of multivariate skewnees and kurtosis with applications. *Biometrika, 57*(3), 519-530. https://doi.org/10.2307/2334770

Mertler, C.A., & Vannatta, R.A. (2005). *Advanced and multivariate statistical methods: Practical application and interpretation (3th ed.).* Pyrczak Publishing.

Messick, S. (1981). Evidence and ethics in the evaluation of tests 1. *ETS Research Report Series, 1981*(1), 1-41. https://doi.org/10.1002/j.2333-8504.1981.tb01244.x

Murphy, K.R., & Davidshofer, C.O. (2005). *Psychological testing: principles and applications.* Pearson Education International.

Netemeyer, R.G., Bearden, W.O., & Sharma, S. (2003). *Scaling procedures.* SAGE Publications, Inc.

Nunally, J.C. (1978). *Psychometric theory.* McGraw Hill.

Özgüven, E. (1994). *Psikolojik testler* [*Psychological tests*]. Yeni Doğuş Matbaası.

Park, H.S., Dailey, R., & Lemus, D. (2002). The use of exploratory factor analysis and principal components analysis in communication research. *Human Communication Research, 28*(4), 562-577. https://doi.org/10.1111/j.1468-2958.2002.tb00824.x

Pett, M.A., Lackey, N.R., & Sullivan, J.J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. SAGE.

Raubenheimer, J. (2004). An item selection procedure to maximize scale reliability and validity. *SA Journal of Industrial Psychology, 30*(4), 59-64. https://doi.org/10.4102/sajip.v30i4.168

Samuels, P. (2017). *Advice on Exploratory Factor Analysis. Technical Report.* Centre for Academic Success, Birmingham City University.

Scherer, R.F., Luther, D.C., Wiebe, F.A., & Adams, J.S. (1988). Dimensionality of coping: Factor stability using the ways of coping questionnaire. *Psychological Reports, 62*(3), 763-770. https://doi.org/10.2466/pr0.1988.62.3.763

Sarstedt M., & Mooi E. (2014). Factor analysis. *In: A concise guide to market research. springer texts in business and economics*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-53965-7_8

Spicer J. (2005). *Making sense of multivariate data analysis: An Intuitive approach*. SAGE.

Stapleton, C.D. (1997). *Basic concepts and procedures of confirmatory factor analysis* [*Paper presentation*]. The Annual Meeting of the South West Educational Research Association. Austin.

Stewens, J. (1996). *Appied multivariate statistics for the social science (Third Edition).* Lawrence Erlbaum Associates.

Şencan, H. (2005). *Sosyal ve davranışsal ölçümlerde güvenirlik ve geçerlilik* [*Reliability and validity in social and behavioral assessments*]. Seçkin Yayıncılık.

Tabachnick, B.G., & Fidell, L.S. (1996). *Using multivariate statistics*. Harper Collins.

Tabachnick, B.G., & Fidell, L.S. (2001). *Using multivariate statistics. 4th Edition, Allyn and Bacon.* MA.

Tavşancıl E. (2002). *Tutumların ölçülmesi ve SPSS ile veri analizi* [*Measurement of attitudes and data analysis with SPSS*]. Nobel Yayınevi.

Velicer, W.F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*(3), 321-327. https://doi.org/10.1007/BF02293557

Ziegler, M. (2014). Comments on item selection procedures. *European Journal of Psychological Assessment, 30*(1), 1-2. https://doi.org/10.1027/1015-5759/a000196

IJATE

International Journal of Assessment Tools in Education

2022, Vol. 9, No. 1, 181–202

https://doi.org/10.21449/ijate.1062613

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

The Effectiveness of Cognitive Behavioral Therapy Including Updating the Early Life Experiences and Images with the Empty Chair Technique on Social Anxiety

Volkan Avsar [1,*], **Seher A. Sevim** [2]

[1]Recep Tayyip Erdogan University, Faculty of Education, Department of Educational Sciences, Guidance and Psychological Counseling, Rize, Turkiye

[2]Independent Researcher

ARTICLE HISTORY

Received: Sep. 30, 2021

Revised: Jan. 16, 2022

Accepted: Jan. 22, 2022

Keywords:

Cognitive behavioral therapy,

Social anxiety,

Negative life experiences,

Imagery,

Empty chair.

Abstract: Specific memories of early negative life experiences (ENLE) and images play an important role in the cause and persistence of social anxiety. In this study, we aimed to investigate the effectiveness of individual Cognitive Behavior Therapy (iCBT), which includes updating a specific memory of ENLE and related images using the empty chair technique, on the social anxiety of university students. In addition to this, how iCBT applied affects the students' general self-efficacy and psychological well-being is also examined. The current study was carried out with a total of eight university students, six female and two male. Participants attended iCBT sessions for 12 weeks. Changes in participants were evaluated with Liebowitz Social Anxiety Scale (LSAS), General Self-Efficacy Scale (GSES), and Flourishing Scale (FS) before iCBT, and one week, three months and six months after iCBT ended. The findings show that participants' social anxiety decreased both statistically and clinically in post-tests and follow-up measures. There were statistically and clinically significant increases in general self-efficacy and psychological well-being. In addition, the changes in the participants' social anxiety, general self-efficacy and psychological well-being post-test and follow-up measures have a large effect size. This study shows that iCBT, which includes updating of specific memories of ENLE and images, is effective in reducing social anxiety and increasing general self-efficacy and psychological well-being. These findings show that iCBT, which includes updating a specific memory of ENLE and related images, using the empty chair technique is an effective method reducing for university students' social anxiety.

1. INTRODUCTION

Social anxiety is a common mental health problem that causes significant functional impairment (Aderka et al., 2012). Its yearly prevalence in the general population is between 2% and 7.4% (Fehm et al., 2008; Kessler et al., 2012), while its lifetime prevalence is between 3.1% and 12.1% (Faravelli et al., 2000; Ruscio et al., 2008). The pervasiveness rate among university students ranges from 11.6% to 36.3% (Baptista et al., 2012; Regis et al., 2018). Among university students in Turkey, the yearly rate is between 7.9% to 20.9%, and the lifetime rate is between 9.6% to 21.7% (Gultekin & Dereboy, 2011; Izgic et al., 2004). This problem, which is also common among students, negatively affects their relationships, functionality, quality of

*CONTACT: Volkan AVSAR ✉ volkan.avsar@erdogan.edu.tr Recep Tayyip Erdogan University, Faculty of Education, Department of Educational Sciences, Guidance and Psychological Counseling, Rize, Turkiye

life, education and career, and adaptation to university, preventing them from benefiting from university life adequately (Brook & Willoughby, 2015; Ghaedi et al., 2010; Nordstrom et al., 2014).

Social anxiety is negatively related to people's general self-efficacy (Rudy et al., 2012). General self-efficacy includes the individual's belief in coping skills in the face of difficulties (Luszczynska et al., 2005; Scherbaum et al., 2006; Schwarzer, 1994). In other words, people with high general self-efficacy have confidence in their ability to cope with the difficulties they face. According to Bandura (1977a), having a functional coping skill contributes positively to the sense of self-efficacy. For this reason, individuals' experiences in which they can use their skills successfully will increase their sense of self-efficacy (Bandura, 1977b). Such experiences include the use of cognitive and behavioral interventions (behavioral such as exposure), which are the basic components of Cognitive Behavior Therapy (CBT) in anxiety-provoking situations for individuals with social anxiety (Gordon et al., 2014; Holaway & Heimberg, 2004). Belief in self-efficacy is also associated with cognitive change and behavioral change (Bandura, 1997). It is stated that exposure to certain situations particularly increase self-efficacy (Biran & Wilson, 1981). As a result, it can be said that cognitive and behavioral interventions to be applied for social anxiety may also lead to an increase in general self-efficacy feelings.

Studies show that people with social anxiety also have low psychological well-being (Fava et al., 1998; Kashdan et al., 2006; Wang et al., 2014; Wersebe et al., 2018). Psychological well-being point to the experience of life going well (Huppert, 2009) and realization of one's true potential (Ryff, 1989b). Psychological well-being is the combination of feeling good and functioning effectively. In other words, it includes experiencing both positive emotions of happiness and contentment, and negative or painful emotions that are a natural part of life and being functional enough to cope with them (Huppert, 2009). According to (Ryff, 1989a), effective functioning is important for psychological well-being. The reason is that when negative emotions last excessively or for too long and the functionality of the person in their daily life is negatively affected, psychological well-being is compromised (Huppert, 2009). Social anxiety is a persistent problem (Kring & Johnson, 2014). People with social anxiety show significant functional impairments (Aderka et al., 2012; Maner & Kenrick, 2010). In addition, these individuals have dysfunctional cognitive and behavioral strategies (McManus et al., 2008; Mellings & Alden, 2000). It can be said that these situations, which prevent them from realizing their potential, also affect their psychological well-being. Consequently, it can be said that since the cognitive and behavioral interventions that will be applied to people so that they can cope with their social anxiety will improve their functionalities, it may also increase their psychological well-being.

The fact that social anxiety starts from an early age and develops slowly (American Psychiatric Association, 2013), becomes persistent when not intervened, and causes significant deterioration in academic, professional, and social functionality by affecting the later years of life draw the attention to childhood years (Vassilopoulos, 2012). Childhood years is a period when social life begins, life-oriented learning is intense, and memories that may affect the later stages of life are formed. Negative life experiences in childhood have a unique effect on the onset of social anxiety (Magee, 1999). People with social anxiety develop a series of assumptions, dysfunctional beliefs, and negative self-images about themselves and their social world based on their early negative life experiences in these years (Clark, 2001). Studies have shown that negative early life experiences are also common among university students with high social anxiety (Binelli et al., 2012), and that these people mostly have negative and social anxiety-related memories (Krans et al., 2014).

Clark and Wells' (1995) social anxiety model has simplified the understanding of the cognitive and behavioral components that cause the maintaining of social anxiety. This model claims that the self-focused attention, safety behaviors, and their dysfunctional beliefs about themselves and their social world as the main components that maintain social anxiety. Another important point that the model focuses on is mental imagery. People with social anxiety generally have negative self-images in relation to negative self-beliefs (Wild & Clark, 2015). Such people, who focus on how they make an impression on others, especially in feared social situations, experience extremely negative images of themselves. Using these images, they make biased and erroneous inferences about how they are seen by others (Clark & Wells, 1995). Hackmann et al. (1998) confirm that people with social anxiety experience negative, distorted, observer-perspective images in social environments where their anxiety is activated. Socially anxious people with negative images experience high levels of anxiety, believe that their anxiety is seen by others, and evaluate their performance negatively. They also use more safety behaviors, have more negative thoughts, and are more self-focused (Hirsch et al., 2004; Makkar & Grisham, 2011).

Negative images commonly seen in people with social anxiety play a causal role in the development and maintaining of social anxiety (Hirsch et al., 2003; Hirsch et al., 2006). Negative images are related to negative life experiences during the onset of social anxiety (Hackmann et al., 2000; Kuo et al., 2011; Stopa & Jenkins, 2007; Wild & Clark, 2011). Therefore, social anxiety interventions should focus on updating the mental representation images of early negative life experiences (Knutsson et al., 2020).

People with social anxiety commonly experience negative, recurrent, and intrusive images in anxiety-provoking social situations (Hackmann et al., 2000; Wild et al., 2007, 2008). In other words, images are activated again and again following the onset of social anxiety. Even if the person experiences positive experiences in social situations in the later years of their life, they cannot develop new perspectives and update themselves due to the activation of images (Hackmann et al., 2000). In order to cope with social anxiety, these negative images, which play a key role in the maintaining of social anxiety, need to be updated (Wild & Clark, 2011). Updating the images will have a strong effect on the emotional processing and change of meaning of early life experiences (Arntz, 2011). In summary, recurrent and intrusive memories of earlier negative life experiences and images associated with present social anxiety is seen. Therefore, it is important to modify and update the early negative life experiences and the images which are mental representations of these early life experiences.

There are many studies in the literature showing that CBT, which includes early negative life experiences and updating the images, is effective for social anxiety (Knutsson et al., 2020; Lee & Kwon, 2013; Nilsson et al., 2012; Norton & Abbott, 2016; Norton et al., 2021; Romano et al., 2020; Wild et al., 2007, 2008). During the counseling process, CBT also makes use of techniques from other theories (Harwood et al., 2010; Howes & Parrott, 1991). While updating negative life experiences and images, the "empty chair" technique from Gestalt Therapy (GT) can also be used. It is stated that this technique can be combined with CBT in order to increase the affect, reveal the cognitions of the participants, and for purposes of cognitive restructuring (Arnkoff, 1981; Kellogg, 2004). According to Paivio et al. (2001), the impact of early negative life experiences and emotions can be reduced by internalizing new information and developing new perspectives. With the empty chair technique, people can express their feelings, thoughts and needs to people with whom they have problems, even if it is imaginary. Eventually, this contributes to the development of new understandings and new perspectives about themselves, the people with whom they have problems and their negative lives. It also helps to finish unfinished business (Greenberg, 2010). It is thought that an individual CBT (iCBT), in which Clark and Wells' (1995) social anxiety model is included, and specific memories of early negative life experiences and related images are updated with the empty chair technique, will be effective in coping with social anxiety.

The aim of this study is to investigate the effectiveness of iCBT, in which a specific memory of early negative life experiences and related images will be modified and updated using the empty chair technique, on the social anxiety of university students. In addition, the effect of iCBT applied on university students' general self-efficacy and psychological well-being is also examined. The research hypotheses formed in line with the aims of the research are as follows:

1. The social anxiety pre-test scores of the participants are higher than the post-test, follow-up 1 and follow-up 2 scores.
2. The general self-efficacy pre-test scores of the participants are lower than the post-test, follow-up 1 and follow-up 2 scores.
3. The psychological well-being pre-test scores of the participants are lower than the post-test, follow-up 1 and follow-up 2 scores.

2. METHOD

2.1. Research Model

This study was carried out in the one group pre-test – post-test design, which is one of the poor experimental designs. The current study design is shown at Table 1.

Table 1. *Study design.*

Participants	Pre-test	X	Post-test	Follow-up 1	Follow-up 2
P	O_1	Intervention	O_2	O_3	O_4

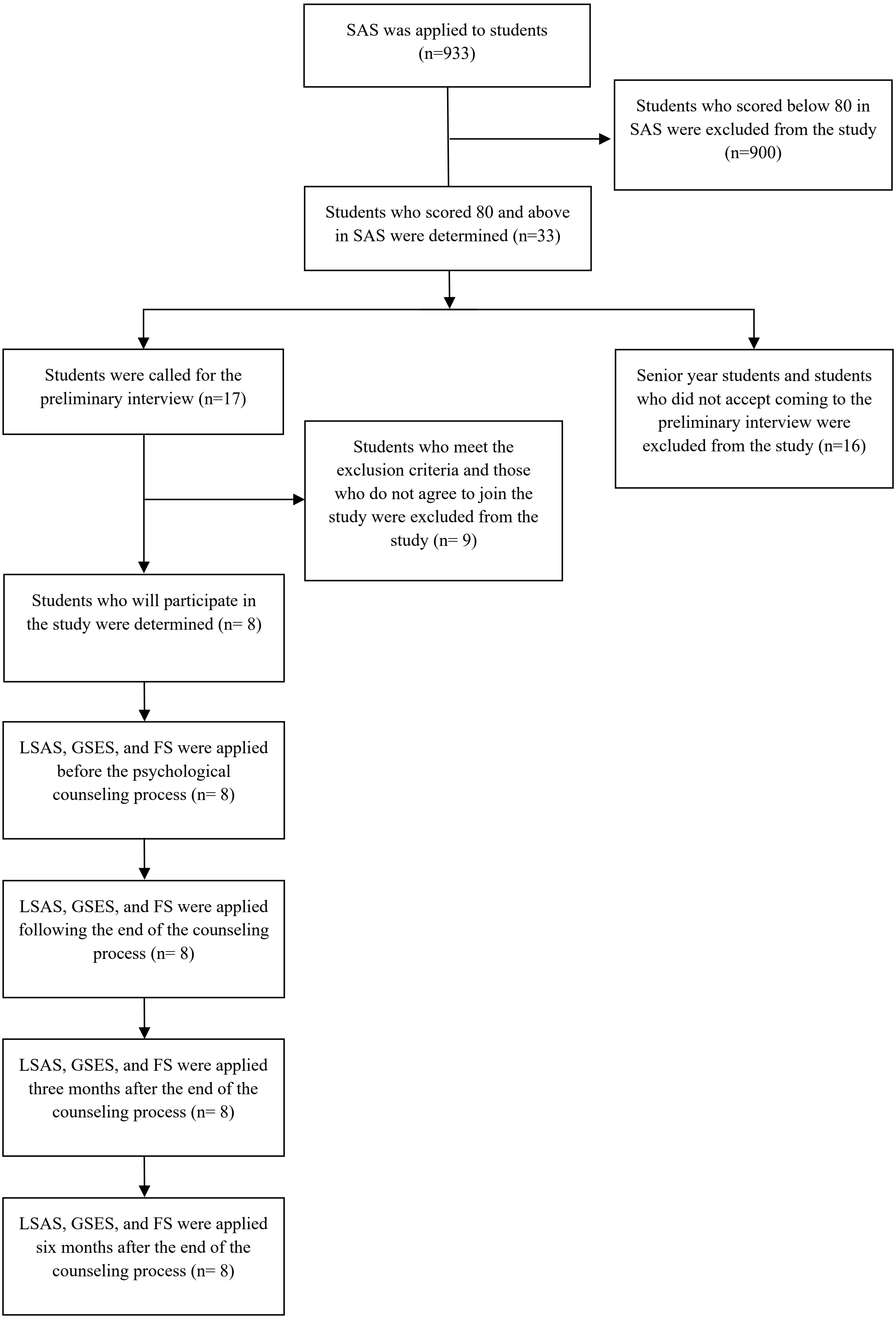
2.2. Participants

This study was approved by the Blinded University Ethics Committee (85434274-050.04.04/6260). In order to determine the participants, the necessary legal permission was obtained from the university where the application will be made (72940495-900/1203). Following the permission, the Social Anxiety Scale (SAS) developed by Özbay and Palancı (2001), was applied to 933 undergraduate students studying in seven different departments (see Figure 1). While settling the participants who will participate in the study, 33 students with high social anxiety were contacted for a preliminary interview. Among those, students who are in their last year were not included in the pre-interview so as not to prevent the possible effects of situations such as graduation, worries around finding a job, and preparing for national exams.

While giving information about the research process to 17 students who participated in the preliminary interview, the researcher also collected information about the students' personal problems, eligibility for work and motivation level. The information obtained in the preliminary interview was evaluated in terms of inclusion and exclusion criteria, which were created to identify the participants. The inclusion criteria of the study consisted of: (a) A score of 80 or more in SAS, which is two standard deviations above the mean, (b) Having the motivation to cope with social anxiety, (c) Accepting the recording of the sessions, (d) Signing the Informed Consent form, (e) To be able to participate actively and continuously in the study process. Individuals were excluded if: (a) There is another psychological problem accompanying social anxiety, (b) There are people with a serious physical health or psychological problem in their family or close circle, (c) The participant has received or is still receiving counseling near the time of the study, (d) They use medication due to social anxiety.

As a result of the preliminary interview, eight students who met the inclusion criteria were included in the study. Detailed information about the process (meeting place and time, duration and number of sessions, confidentiality, recording sessions etc.) was given to these people. In addition, the Liebowitz Social Anxiety Scale (LSAS), General Self-Efficacy Scale (GSES) and Flourishing Scale (FS) were applied to the participants before the counseling process started, and one week, three months and six months after the counseling process ended.

Figure 1. *The flowchart of determining of the participants and the process.*


SAS was applied to students (n=933)
Students who scored below 80 in SAS were excluded from the study (n=900)
Students who scored 80 and above in SAS were determined (n=33)
Students were called for the preliminary interview (n=17)
Senior year students and students who did not accept coming to the preliminary interview were excluded from the study (n=16)
Students who meet the exclusion criteria and those who do not agree to join the study were excluded from the study (n= 9)
Students who will participate in the study were determined (n= 8)
LSAS, GSES, and FS were applied before the psychological counseling process (n= 8)
LSAS, GSES, and FS were applied following the end of the counseling process (n= 8)
LSAS, GSES, and FS were applied three months after the end of the counseling process (n= 8)
LSAS, GSES, and FS were applied six months after the end of the counseling process (n= 8)

2.3. Counseling Process and Sessions

For the participants to cope with social anxiety, iCBT procedure which consists of 12 face-to-face sessions are structured. The fact that the participants were university students was influential in limiting iCBT to 12 sessions. It was aimed to conclude the iCBT in one semester in order to avoid secondary variables that may appear during the semester break and affect the research results. The structured sessions are presented to three academicians who are experts in the field of CBT, their opinions are collected and the final version of the iCBT procedure is formed. The duration of each session is 50 minutes. However, the first session in which the participants' issues are evaluated as well as the fourth session in which a specific memory of negative life experiences and related images are re-evaluated are determined as 90 minutes as per their content. The application of the structured counseling process is carried out by the first author of the research, who has a CBT training. Throughout the procedure, training and regular supervision from experienced CBT supervisors is received.

In Session 1, it is aimed to evaluate, conceptualize the problem and create a therapeutic alliance. Within that framework, when the participants' history of the problem is recorded the beliefs, thoughts, emotions, behaviors and physiological symptoms are identified. A specific memory of early negative life experiences and images that may be associated with the participants' problems are also pinpointed. How the problem affects the life of participants is evaluated.

In Session 2, the focus is on psycho-education, goal setting and conceptualization of the problem. Participants are primarily informed about the general structure of the sessions and agenda setting. Then, the objectives that the participants want to achieve at the end of the process are determined. Then, the formulations of the participants are carried out using the social anxiety model of Clark and Wells (1995).

In Session 3, the focus is on how the participants' automatic thoughts, cognitive distortions and safety behaviors relate to their problem and their affect on the maintenance of this issue.

In Session 4, a specific memory of early negative life experiences related to the participants' problem, the meaning they ascribe to this memory and images are identified, and are re-evaluated using the empty chair technique.

In the first step of addressing negative experiences, participants' memories from their childhood years that affect them and cause them to experience negative emotions when they remember them were discussed in general terms. Memories that participants told were listed under titles, and the participants were asked to determine the memory that they deem is related to their issue and one that affect them the most. Later on, their beliefs, thoughts, emotions, images related to this experience and the extent to which they are affected by this experience was discussed. In the second step, participants' discovery of how their negative experiences affect the later years of their lives, their relationships, and how it is related to their problem was focused on.

In the third step, memories specified by the participants were addressed in two stages, using the empty chair technique. Memories participants focused on in the first stage of the application were revived as if they were being experienced again at that current time. With the help of empty chair technique, participants were facilitated to express their emotions, thoughts and images by creating dialogues with the people they experienced the issue. After the participants expressed their emotions, thoughts and images, they took the place of the person who caused the negative experience by taking the empty chair facing them. They responded as that person in response to what they said as themselves. In this way, the participants had the opportunity to express their feelings, thoughts and images by experiencing them. Also, by taking the place of the person in front of them, they had the opportunity to evaluate their life and images from the perspective of the other person in a way that they had not evaluated before.

In the second stage of the application, participants evaluated their experience as their current adult self and as a child who had the experience. Within this framework, their reactions as a child who encountered that situation at the time were elaborated. How the child self-evaluated the issue and the limitations of their reactions as a child were discussed. Later, it was focused how their current adult self would evaluate the situation and how they would react. During this application, the participants were facilitated to discover the difference between children and adults in terms of giving reaction. In the fourth and final step after the participants re-evaluated a specific memory of early negative life experiences and images, it was focused whether they discovered something about their life they had not noticed before or had just noticed, or whether they gained new perspectives. It was addressed whether there was any change or difference in the meaning they attribute to their memories, images, emotional processing, and beliefs.

In Sessions 5, 6 and 7, the beliefs of the participants about their problems were determined based on the agenda, and cognitive interventions were carried out for the determined beliefs. During these interventions, what the participants learned in the fourth session was also utilized.

In Session 8, the avoidance and safety behaviors of the participants were discussed, and the rationality of exposure and habituation, its place and importance in coping with the problem were evaluated.

In Sessions 9, 10, and 11, cognitive and behavioral interventions were carried out simultaneously. Cognitive and behavioral techniques were used together to intervene on the agenda items determined together with the participants. In these sessions, participants were exposed to these situations they avoid that are related to their problem during the session as well as their daily life. Interventions were made during the session, particularly through behavioral experiments, video feedback, and real and gradual exposure.

In Session 12, the process was concluded.

2.4. Data Collection Instruments

2.4.1. *Liebowitz social anxiety scale*

LSAS, which was developed by Liebowitz (1987) and which consists of 24 items, evaluates the anxiety or avoidance levels in different social and performance situations. Each item in the scale is given an individual score from one to four for anxiety and avoidance dimensions. Adapted to Turkish culture by Soykan et al. (2003), the reliability of LSAS is calculated to be 0.97 for the whole scale via test and re-test method. The Cronbach Alpha reliability coefficient in terms of internal consistency is found to be 0.98 for the whole scale, 0.96 for the anxiety dimension, and 0.95 for the avoidance dimension. In this study, the Cronbach Alpha reliability coefficient of LSAS was calculated to be 0.95 for the whole scale, 0.91 for the anxiety dimension, and 0.89 for the avoidance dimension.

2.4.2. *Social anxiety scale*

SAS was developed by Özbay and Palancı (2001) to identify the social anxiety related problems of university students. The scale consists of 30 items and has a five-point rating within the range of 0-4. For the reliability of the scale, the Cronbach Alpha internal consistency coefficient was calculated as 0.89. In this study, the Cronbach Alpha reliability coefficient of SAS was calculated to be 0.93.

2.4.3. *General self-efficacy scale*

GSES was developed by Schwarzer and Jerusalem (1995) to determine people's efficacy belief in coping with stressful and challenging life events. The scale consists of 10 items and has a four-point rating. The scale was adapted to Turkish culture by Aypay (2010). The reliability of the scale was researched via test and re-test method and the Cronbach Alpha internal consistency coefficient. Test re-test reliability coefficient was calculated as 0.80, and the

Cronbach Alpha reliability coefficient was calculated as 0.83. In this study, the Cronbach Alpha reliability coefficient of GSES was calculated to be 0.89.

2.4.4. *Flourishing scale*

FS, which was developed by Diener et al. (2010) to measure the psychological well-being, consists of eight items. The items in the scale has seven-point rating. The scale was adapted to Turkish culture by Telef (2013). The reliability of the scale was researched via test and re-test method and the Cronbach Alpha internal consistency coefficient. Test re-test reliability coefficient was calculated as 0.86, and the Cronbach Alpha internal consistency coefficient was calculated as 0.80. In this study, the Cronbach Alpha reliability coefficient of FS was calculated to be 0.88.

2.5. Statistical Analyses

For the purpose of the study, whether there is a statistically significant difference between the mean scores of the pre-test, post-test and follow-up scores obtained from the scales applied to the participants was analyzed via the Friedman test. Furthermore, in order to analyze whether there was a statistically significant difference between the pretest-posttest, pretest-follow-up 1 and pretest-follow-up 2 scores of the participants, the Wilcoxon Signed Rank Test was used. The effect size of the existing difference was calculated with the formula in equation 1 (Rosenthal, 1991).

$$r = \frac{|z|}{\sqrt{N}} \qquad \text{(Equation 1)}$$

In addition, the clinical significance of the change in participants was evaluated in this study. Two conditions need to be met in order to determine the clinical significance of the change observed in the participants as a result of the intervention. According to the first of the two conditions, the participants in the dysfunctional population before the intervention must be transferred to the functional population after the intervention. While determining this transition, the cut-off score of the measurement tool used before and after the intervention is used. In other words, the post-test scores of the person need to surpass the cut-off point threshold. In cases where there is no cut-off point for the measurement tools, a cut-off point is calculated for the measurement tools. Since there were no cut-off scores for the measurement tools used in this study, cut-off scores were generated (Bauer et al., 2004; Jacobson & Truax, 1991). In accordance with this, the cut-off scores for LSAS were 84; 26 for GSES; and 43 for FS. According to the second condition, the change in the person who exceeds the cut-off score should be statistically reliable. And this is evaluated using the "*Reliable Change Index (RCI)*" (Jacobson et al., 1984; Jacobson & Truax, 1991).

If the RCI is above ∓ 1.96, it shows that the change observed in the person is reliable, while a range of ∓ 1.96 indicates that the change is unreliable (Jacobson & Truax, 1991). The change observed in people using RCI and cut-off scores is evaluated in four categories (Atkins et al., 2005):

a. *Recovered* (Meets both RCI and cut-off score criteria)
b. *Improved* (Meets RCI criteria but not cut-off score criteria)
c. *Unchanged* (Does not meet neither RCI nor cut-off score criteria)
d. *Deteriorated* (Meets RCI criteria but deteriorated)

3. RESULTS

3.1. Participants' Baseline Characteristics

The demographics of the participants is given in Table 2. When examined, it is seen in Table 2 that most of the participants are women (6 female, 2 male participants). Additionally, the participants come from different departments and grade levels.

Table 2. *Participants' baseline characteristics.*

Variable	Level	Frequency (f)
Gender	Female	6
	Male	2
Age	18	3
	19	1
	20	2
	21	2
Grade Level	1	4
	2	3
	3	1
Department	School Teaching	3
	Psychological Counseling and Guidance	2
	Science Teaching	1
	Turkish Teaching	1
	Computer and Instructional Technologies Teaching	1
Total		8

3.2. Treatment Effects

The average and standard deviations of the participants' pre-test scores from LSAS, GSES and FS and their post-test, follow-up 1 (3 months), follow-up 2 (6 months) scores are given in Table 3.

Table 3. *Descriptive statistics of study variables across counseling conditions.*

	Pre-treatment		Post-treatment		Follow-up 1		Follow-up 2	
Measures	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
LSAS	108.12	12.13	45.62	19.54	37.87	8.37	38.62	10.88
GSES	18.00	4.10	33.12	4.54	31.62	3.77	30.87	5.43
FS	24.50	9.33	46.75	4.26	47.37	3.58	46.12	7.21

Note. LSAS = Liebowitz Social Anxiety Scale; GSES = General Self-Efficacy Scale; FS = Flourishing Scale.

When Table 3 is examined, it is seen that the social anxiety of the participants decreased after the iCBT and in the follow-up measurements compared to the pre-test measurements. Their general self-efficacy and psychological well-being, however, increased after iCBT and in follow-up measurements compared to pre-test measurements. Whether the difference between the scores is statistically significant is determined via the Wilcoxon Signed Rank Test, and the acquired results are given in Table 4. Also, the values obtained from the comparison of the measurements taken at three different times (Pre-test – Post-test/Pre-test – Follow-up 1/Pre-test – Follow-up 2) from three different measurement tools (LSAS, GSES, and FS) are the same and are included in Table 4. In addition, the RCI results calculated to reveal the clinical significance of the changes in the social anxiety, general self-efficacy and psychological well-being of the participants are included in Table 5, Table 6 and Table 7, respectively.

Table 4. *The comparison of the participants' pre-test scores and post-test, follow-up 1 and follow-up 2 scores.*

Measures		n	Pre-test–Post-test/Pre-test–Follow-up 1/Pre-test–Follow-up 2 $\bar{X}rank$	 $\sum rank$	 z	 p	 r
LSAS	Negative Ranks	8	4.50	36.00	*-2.52***	*0.012**	*0.89*
	Positive Ranks	0	0.00	0.00			
	Ties	0					
	Total	8					
GSES	Negative Ranks	0	0.00	0.00	*-2.52****	*0.012**	*0.89*
	Positive Ranks	8	4.50	36.00			
	Ties	0					
	Total	8					
FS	Negative Ranks	0	0.00	0.00	*-2.52****	*0.012**	*0.89*
	Positive Ranks	8	4.50	36.00			
	Ties	0					
	Total	8					

Note. $^{*}p<.05$; **Based on positive ranks; ***Based on negative ranks; *r*: Effect size

When Table 4 is examined, it is seen that the difference between the participants' LSAS pre-test scores and post-test, follow-up 1 and follow-up 2 scores is statistically significant (z=-2.52, p<.05). The participants' social anxiety pre-test scores are higher than their post-test, follow-up 1 and follow-up 2 scores. The effect size of this existing difference is large r = 0.89; 95% confidence interval [CI: .49, .98].

It is seen that the difference between the participants' GSES pre-test scores and post-test, follow-up 1 and follow-up 2 scores is statistically significant (z=-2.52, p<.05). The participants' general self-efficacy pre-test scores is lower than their post-test, follow-up 1 and follow-up 2 scores. The effect size of this existing difference is large r = 0.89; 95% CI [.49, .98].

It is seen that the difference between the participants' FS pre-test scores and post-test, follow-up 1 and follow-up 2 scores is statistically significant (z=-2.52, p<.05). The participants' psychological well-being pre-test scores are lower than their post-test, follow-up 1 and follow-up 2 scores. The effect size of this existing difference is large r=0.89; 95% CI [.49, .98]. According to the results obtained, the hypotheses of the research are accepted.

Table 5. *RCI results for clinical significance of difference between scores from LSAS.*

P	$X_{pre\text{-}test}$	$SD_{pre\text{-}test}$	$X_{post\text{-}test}$	$X_{follow\text{-}up1}$	$X_{follow\text{-}up2}$	$X_{post\text{-}test-pre\text{-}test}$	$X_{follow\text{-}up1-pre\text{-}test}$	$X_{follow\text{-}up2-pre\text{-}test}$	r_x	$RCI_{post\text{-}test}$	$RCI_{follow\text{-}up1}$	$RCI_{follow\text{-}up2}$
P1	101	12.13	43	41	37	-58	-60	-64	*.97**	-19.52	-20.19	-21.54
P2	112		84	52	51	-28	-60	-61		-9.42	-20.19	-20.53
P3	126		34	41	43	-92	-85	-83		-30.96	-28.60	-27.93
P4	96		44	42	38	-52	-54	-58		-17.50	-18.17	-19.52
P5	114		66	40	31	-48	-74	-83		-16.15	-24.90	-27.93
P6	109		34	31	27	-75	-78	-82		-25.24	-26.25	-27.59
P7	89		27	26	56	-62	-63	-33		-20.86	-21.20	-11.10
P8	118		33	30	26	-85	-88	-92		-28.60	-29.61	-30.96

Note. P: Participant; r_X: LSAS reliability co-efficient;

*The reliability co-efficient of LSAS is taken from Soykan et al.'s (2003) research.

When Table 5 is examined, it is seen that the social anxiety scores of all the participants in the post-test, follow-up 1 and follow-up 2 tests are equal to or below the cut-off score. When RCI scores from Table 5 is examined, it is seen that RCI scores of all the participants exceed ±1.96. This finding shows that there is a statistically reliable change in social anxiety when the changes between the scores of the participants from LSAS from the pre-test to the follow-up 2 are evaluated.

Consequently, the changes in all participants' social anxiety post-test, follow-up 1 and follow-up 2 test scores are clinically significant because they meet the cut-off score and RCI conditions together. The change in the participants' social anxiety is seen in *recovered* category.

Table 6. *RCI results for clinical significance of difference between scores from GSES.*

P	$X_{pre\text{-}test}$	$SD_{pre\text{-}test}$	$X_{post\text{-}test}$	$X_{follow\text{-}up1}$	$X_{follow\text{-}up2}$	$X_{post\text{-}test-pre\text{-}test}$	$X_{follow\text{-}up1-pre\text{-}test}$	$X_{follow\text{-}up2-pre\text{-}test}$	r_x	$RCI_{post\text{-}test}$	$RCI_{follow\text{-}up1}$	$RCI_{follow\text{-}up2}$
P1	18	4.1	26	27	32	8	9	14	*.80*[*]	3.08	3.47	5.39
P2	18		29	34	33	11	16	15		4.24	6.17	5.78
P3	25		36	32	30	11	7	5		4.24	2.70	1.92
P4	23		34	31	32	11	8	9		4.24	3.08	3.47
P5	17		30	26	29	13	9	12		5.01	3.47	4.62
P6	15		38	35	35	23	2	20		8.87	7.71	7.71
P7	13		33	31	19	20	18	6		7.71	6.94	2.31
P8	15		39	37	37	24	22	22		9.25	8.48	8.48

Note. P: Participant; r_x: GSES reliability co-efficient;

[*]The reliability co-efficient of GSES is taken from Aypay's (2010) research.

When Table 6 is examined, it is seen that the general self-efficacy scores of all the participants in the post-test, follow-up 1 and follow-up 2 tests (except for P7) are equal to or above the cut-off score. Only the P7 follow-up 2 test scored below the cut-off point. When the RCI scores in Table 6 are examined, it is seen that the changes in the post-test and follow-up 1 tests of all participants are clinically significant. In follow-up 2 test, the changes in the general self-efficacy of participants other than P3 and P7 are clinically significant.

Consequently, the changes in the general self-efficacy post-test and follow-up 1 tests of all participants are clinically significant. The changes observed are seen in the *recovered* category. In the follow-up 2 test, the changes in the general self-efficacy of the participants except for P3 and P7 were clinically significant and seen in the *recovered* category. In the follow-up 2 test, the change in P3 does not meet the RCI condition and the change in P7 does not meet the cut-off score condition, hence it is not clinically significant. In other words, although the change in the P3 general self-efficacy exceeds the cut-off score in the follow-up 2 test, it is not statistically reliable. Nonetheless, the change in P7 in the follow-up 2 test is statistically reliable but the change is included in the *improved* category because it does not exceed the cut-off score.

When Table 7 is examined, it is seen that the psychological well-being scores of the majority of the participants in the post-test, follow-up 1 and follow-up 2 tests are equal to or above the cut-off score. P1 and P5 among the participants got a score below the cut-off score in the post-test, and P7 got a score below the cut-off in the follow-up 2 test. When the RCI scores in Table 7 are examined, it is seen that the RCI scores of the participants other than P3 in the post-test, follow-up 1 and follow-up 2 exceed ∓ 1.96. RCI scores of Participant 3 did not exceed ∓ 1.96 in any of the three measurements.

Table 7. *RCI results for clinical significance of difference between scores from FS.*

P	$X_{pre\text{-}test}$	$SD_{pre\text{-}test}$	$X_{post\text{-}test}$	$X_{follow\text{-}up1}$	$X_{follow\text{-}up2}$	$X_{post\text{-}test-pre\text{-}test}$	$X_{follow\text{-}up1-pre\text{-}test}$	$X_{follow\text{-}up2-pre\text{-}test}$	r_x	$RCI_{post\text{-}test}$	$RCI_{follow\text{-}up1}$	$RCI_{follow\text{-}up2}$
P1	28	9.33	42	50	51	14	22	23	*.86**	2.83	4.45	4.65
P2	31		45	46	48	14	15	17		2.83	3.03	3.44
P3	39		47	45	44	8	6	5		1.62	1.21	1.01
P4	27		44	44	45	17	17	18		3.44	3.44	3.64
P5	20		42	43	47	22	23	27		4.45	4.65	5.46
P6	12		53	51	51	41	39	39		8.30	7.90	7.90
P7	12		49	47	30	37	35	18		7.49	7.08	3.64
P8	27		52	53	53	25	26	26		5.06	5.26	5.26

Note. P: Participant; r_x: FS reliability co-efficient;
*FS reliability co-efficient is taken from Telef's (2013) research.

Consequently, the changes in the psychological well-being post-test, follow-up 1 and follow-up 2 tests of P2, P4, P6 and P8 are clinically significant. The changes observed are seen in the *recovered* category. The change in P3's psychological well-being is clinically insignificant because it does not meet the RCI condition in any of the three measurements. In other words, although the change in the P3's psychological well-being exceeds the cut-off score in all three measurements, it is not statistically reliable. Additionally, the changes in the psychological well-being of P1 and P5 in the post-test, and of P7 in the follow-up 2 test met the RCI condition but did not meet the cut-off score condition. Hence, the changes observed in these participants are statistically reliable but since they do not exceed the cut-off score, they are included in the *improved* category.

4. DISCUSSION and CONCLUSION

In this study, the effect of iCBT, which includes a specific memory of early negative life experiences and updating related images using the empty chair technique, on social anxiety was researched. The social anxiety post-test scores of the participants decreased statistically significantly compared to the pre-test scores. The decrease continued in the follow-up measurements. In accordance with this, iCBT is effective and has permanent effect. The results gathered are consistent with experimental studies using iCBT with socially anxious university students (Shorey & Stuart, 2012; Tsitsas & Paschali, 2014) and participants from different age groups (Datta & Das, 2016; Goldin et al., 2013; Goldin et al., 2012; Leigh & Clark, 2016; Narr & Teachman, 2017; Pinjarkar et al., 2015; Priyamvada et al., 2009; Weiss et al., 2011; Wootton et al., 2018; Yoshinaga, Kobori, et al., 2013; Yoshinaga, Ohshima, et al., 2013). In addition, the study of Leigh and Clark (2016), which includes interventions for early negative life experiences just like current study, is important in that respect. Meta-analysis and review studies, similar to the results of this research, reveal that iCBT is effective in coping with social anxiety (Gil et al., 2001; Mayo-Wilson et al., 2014; Taylor, 1996).

In this study, cognitive and behavioral techniques were used in social anxiety intervention processes. Additionally, participants' negative childhood experiences, the meaning they attributed to these experiences and their images were re-evaluated using the empty chair technique. It could be said that research findings which interfere with the early negative life experiences of individuals with social anxiety via imagery rescripting technique support this study (Frets et al., 2014; Knutsson et al., 2020; Lee & Kwon, 2013; Nilsson et al., 2012; Norton & Abbott, 2016; Reimer & Moscovitch, 2015; Wild et al., 2007, 2008). These studies have proven that re-evaluating negative experiences even for a single session is effective in reducing social anxiety, changing the meaning of negative experiences with the effect of images, and weakening negative core beliefs. It is also seen that this effect continues in the follow-up measurements as well. Wild and Clark (2011) emphasize that images that are associated with a specific memory of early negative life experiences have an important role in the maintenance of social anxiety. They explain that these images have to be intervened in order to cope with social anxiety. In this study, it can be stated that re-evaluating the participants' a specific memory of early negative life experiences and related images with the empty chair technique is effective in reducing their anxiety.

The decrease in all participants' social anxiety is clinically significant and can be seen in *recovered* category. These results appear to be consistent with the study results in which clinical significance is calculated (Bogels et al., 2014; Clark et al., 2006; Goldin et al., 2013; Mörtberg et al., 2007; Shorey & Stuart, 2012; Stangier et al., 2003; Wootton et al., 2018). The fact that the study was conducted with a group that was not clinically diagnosed may have facilitated the participants' effective benefiting and *recovery* from an intervention process that lasted for a total of 12 weeks. Different from the studies above, in this study, a specific memory of early negative life experiences and images of the participants were re-evaluated experientially using the empty chair technique. This might be another reason why the participants benefited from this process. To sum, the results gathered with regards to the clinical significance of the decrease in the participants' social anxiety present a crucial proof that show the intervention is effective. Additionally, the fact that the effect obtained continues to increase in both follow-up measurements supports that the effect of the study is permanent.

iCBT applied in this study is also effective in the increase of the participants' general self-efficacy. This result from the study is consistent with the experimental studies where iCBT is used (Jafari et al., 2012; Keshi & Basavarajappa, 2013). Jafari et al. in particular (2012) comes to be conclusion that iCBT is effective in the increasing of general self-efficacy and that the effect continues in the follow-up measurement. Another resemblance between these studies and current study is that the participants were selected from a non-clinical sample. This may be one of the factors in increasing general self-efficacy as a result of the intervention. Additionally, it can be said that in this study the interventions for early negative life experiences and images have an effect on the increase in the general self-efficacy of the participants.

The increase observed in the general self-efficacy post-test and follow-up 1 test of all participants are clinically significant. These increases can be seen in the *recovered* category. In the follow-up 2 test, the increase in two participants was not clinically significant. Although the change in one of these participants (P7) was statistically reliable, it was not clinically significant. For this reason, the change in their general self-efficacy is included in the *improved* category. When this participant was contacted for the follow-up 2 test, they stated that they had "*serious family problems*". It can be said that this situation had an impact on the change observed in the follow-up 2 test. The results obtained show that applied iCBT has a clinically significant effect on increasing general self-efficacy. Also, follow-up measurements collected support that this effect is permanent. In the literature review, no study was found in which clinical significance of general self-efficacy was reported. It is thought that this result about general self-efficacy is original, and will contribute to the literature.

It is seen that the iCBT applied in this study is effective in the increase of the participants' psychological well-being as well. This result is consistent with the research results in the literature. Studies show that CBT and approaches based on CBT increase the psychological well-being (Fava et al., 1998; Freeman et al., 2014; Ruini et al., 2006). In a meta-analysis study, it is emphasized that face-to-face applications in which behavioral interventions were used were effective in increasing psychological well-being (Weiss et al., 2016). Similarly, in this study too, it was found that iCBT decreases social anxiety while it also increases psychological well-being. In addition, there are studies in the literature showing that early negative life experiences have negative effects on well-being (Melkman, 2017; Mosley-Johnson et al., 2018). In Ryff's (2014) model, these results are seen to be related to the self-acceptance dimension, which includes the individual's positive feelings and evaluations about their past life and themselves. In other words, early negative life experiences create an obstacle in people's self-acceptance and affect their psychological well-being. In this study, it can be said that the interventions made for a specific memory of early negative life experiences and images of the participants have an effect on increasing their psychological well-being.

It was seen that the increase in the psychological well-being of half of the participants is clinically significant. These increases can be seen in the *recovered* category. In the follow-up test measurements, the number of participants whose change in their psychological well-being is clinically significant increased. And this shows that the effect of the iCBT applied continues to increase. The change in psychological well-being of only one participant (P3) was not clinically significant in all three measurements. In the post-test, the change in two participants (P1 and P5) was clinically insignificant. Although the psychological well-being of these participants increased compared to their pre-test scores, the change in post-test measurements was not clinically significant since the score was one point below the cut-off score. However, the psychological well-being of both participants increased in the follow-up measurements and became clinically significant. In the follow-up 2 test, on the other hand, the change in one participant (P7) was not clinically significant, as it did not exceed the cut-off score. It can be said that the follow-up 2 measurement of this participant when they were experiencing "*serious family problems*" had an effect on their psychological well-being as well as their general self-efficacy. According to the results obtained, iCBT applied has a clinically significant effect on increasing the psychological well-being of the participants. Also, follow-up measurements collected support that this effect continues to increase. In the literature review, no study was found in which clinical significance of psychological well-being was reported. It is thought that this result about psychological well-being is original and will contribute to the literature.

The changes observed in the post-test and follow-up 1 and 2 measurements of social anxiety, general self-efficacy, and psychological well-being in this study have a large effect size. These results are consistent with many studies in which CBT was used in relation to social anxiety (Clark et al., 2006; Clark et al., 2003; Gaudiano & Herbert, 2003; Mayo-Wilson et al., 2014; Mörtberg et al., 2007; Stangier et al., 2003; Wootton et al., 2018; Yoshinaga, Ohshima, et al., 2013). Post-test results regarding general self-efficacy are similar to the study by Keshi and Basavarajappa (2013) in which they used iCBT to increase general self-efficacy. Post-test results on psychological well-being are consistent with the study of Freeman et al. (2014), while they are different from the findings of Weiss et al. (2016). In the meta-analysis study of Weiss et al. (2016), it was seen that the post-test effect sizes were medium and small as well as large. The most important reason for this may be the inclusion of studies involving different theoretical approaches and participants with different characteristics (selected from clinical and non-clinical samples) in the meta-analysis.

This study shows that the applied iCBT is effective in reducing the social anxiety of university students in terms of statistical, clinical significance and effect size. It has been determined that the effect of the study continues to increase in the follow-up measurements taken. These results contribute to research results revealing the impact of early negative life experiences and images on coping with social anxiety (e.g., Hackmann et al., 2000; Wild & Clark, 2011; Wild et al., 2007, 2008). In addition, it was revealed that the iCBT applied contributed to the decrease in the social anxiety of the participants, while also contributing to the increase in their general self-efficacy and psychological well-being. This provides further evidence that the applied counseling process is effective.

The effectiveness of the iCBT applied in this study can be re-tested with applications that will involve more socially anxious university students in the future. By making necessary updates on the configured iCBT, its effectiveness can be investigated in different age groups or people with different problems accompanying social anxiety as well. Additionally, the applications on the non-clinical sample in this study can be tested on clinical samples in terms of its effectiveness in the future. It was not possible to form a control group as a limited number of socially anxious participants were reached in this study. In future studies, the effectiveness of this study can be researched once again with participants that include a control group. In new applications that will be made, the effectiveness of the sessions and the structured psychological counselor can be investigated in different dimensions by taking measurements after each session. In this way, the progress the participants make and in which sessions can be determined.

Acknowledgments

This study is based on a summary of the doctoral dissertation entitled "The Effect of Individual Counseling Based on Cognitive Behavior Therapy on Social Anxiety, Self-Efficacy, and Psychological Well-Being" prepared in the Guidance and Psychological Counseling Program, Ankara University.

Availability of Data

Data are openly available at the research Open Science Framework page (https://osf.io/5kbr7/).

Declaration of Conflicting Interests and Ethics

The authors have no competing interests to declare that are relevant to the content of this article. All scientific responsibility of the manuscript belongs to the authors. This study was approved by the **Blinded University Ethics Committee**, under the code of 85434274-050.04.04/6260.

Authorship Contribution Statement

Volkan Avsar: Investigation, Visualization, Structuring and application of the iCBT process, Participants' recruitment, Drafting manuscript. **Seher A. Sevim**: Methodology and Supervision.

Orcid

Volkan Avsar https://orcid.org/0000-0001-9427-9425
Seher A. Sevim https://orcid.org/0000-0002-4914-2486

REFERENCES

Aderka, I.M., Hofmann, S.G., Nickerson, A., Hermesh, H., Gilboa-Schechtman, E., & Marom, S. (2012). Functional impairment in social anxiety disorder. *Journal of Anxiety Disorders*, *26*(3), 393-400. https://doi.org/10.1016/j.janxdis.2012.01.003

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (5th ed.)*. https://doi.org/10.1176/appi.books.9780890425596

Arnkoff, D.B. (1981). Flexibility in practicing cognitive therapy. In G. Emery, S.D. Hollon, & R.C. Bedrosian (Eds.), *New directions in cognitive therapy* (1st ed., pp. 203-223). The Guilford Press.

Arntz, A. (2011). Imagery rescripting for personality disorders. *Cognitive and Behavioral Practice*, *18*(4), 466-481. https://doi.org/10.1016/j.cbpra.2011.04.006

Atkins, D.C., Bedics, J.D., McGlinchey, J.B., & Beauchaine, T.P. (2005). Assessing clinical significance: Does it matter which method we use? *Journal of Consulting and Clinical Psychology*, *73*(5), 982-989. https://doi.org/10.1037/0022-006x.73.5.982

Aypay, A. (2010). Genel özyeterlik ölçeği'nin (GÖYÖ) Türkçe'ye uyarlama çalışması [The adaptation study of general self-efficacy (GSE) scale to Turkish]. *İnönü Üniversitesi*

Eğitim Fakültesi Dergisi, *11*(2), 113–131. https://dergipark.org.tr/tr/download/article-file/92265

Bandura, A. (1977a). Self–efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215. https://doi.org/10.1016/0146-6402(78)90002-4

Bandura, A. (1977b). *Social learning theory* (1st ed.). Prentice Hall.

Bandura, A. (1997). *Self–efficacy: The exercise of control* (1st ed.). W: H. Freeman & Company.

Baptista, C.A., Loureiro, S.R., Osorio, F.D., Zuardi, A.W., Magalhaes, P.V., Kapczinski, F., Santos, A., Freitas-Ferrari, M.C., & Crippa, J.A.S. (2012). Social phobia in Brazilian university students: Prevalence, under-recognition and academic impairment in women. *Journal of Affective Disorders*, *136*(3), 857-861. https://doi.org/10.1016/j.jad.2011.09.022

Bauer, S., Lambert, M.J., & Nielsen, S.L. (2004). Clinical significance methods: A comparison of statistical techniques. *Journal of Personality Assessment*, *82*(1), 60-70. https://doi.org/10.1207/s15327752jpa8201_11

Binelli, C., Ortiz, A., Muñiz, A., Gelabert, E., Ferraz, L., Filho, A.S., Crippa, J.A.S., Nardi, A. E., Subira, S., & Martin-Santos, R. (2012). Social anxiety and negative early life events in university students. *Revista Brasileira De Psiquiatria*, *34*(1), 69-74. https://doi.org/10.1590/s1516-44462012000500006

Biran, M., & Wilson, G.T. (1981). Treatment of phobic disorders using cognitive and exposure methods: A self-efficacy analysis. *Journal of Consulting and Clinical Psychology*, *49*(6), 886-899. https://doi.org/10.1037/0022-006x.49.6.886

Bogels, S.M., Wijts, P., Oort, F.J., & Sallaerts, S.J.M. (2014). Psychodynamic psychotherapy persus cognitive behavior therapy for social anxiety disorder: An efficacy and partial effectiveness trial. *Depression and Anxiety*, *31*(5), 363-373. https://doi.org/10.1002/da.22246

Brook, C.A., & Willoughby, T. (2015). The social ties that bind: Social anxiety and academic achievement across the university years. *Journal of Youth and Adolescence*, *44*(5), 1139-1152. https://doi.org/10.1007/s10964-015-0262-8

Clark, D.M. (2001). A cognitive perspective on social phobia. In W. R. Crozier & L. E. Alden (Eds.), *International handbook of social anxiety: Concepts, research and interventions relating to the self and shyness* (1st ed., pp. 405-430). John Wiley & Sons Ltd.

Clark, D.M., Ehlers, A., Hackmann, A., McManus, F., Fennell, M., Grey, N., Waddington, L., & Wild, J. (2006). Cognitive therapy versus exposure and applied relaxation in social phobia: A randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *74*(3), 568-578. https://doi.org/10.1037/0022-006x.74.3.568

Clark, D.M., Ehlers, A., McManus, F., Hackmann, A., Fennell, M., Campbell, H., Flower, T., Davenport, C., & Louis, B. (2003). Cognitive therapy versus fluoxetine in generalized social phobia: A randomized placebo-controlled trial. *Journal of Consulting and Clinical Psychology*, *71*(6), 1058-1067. https://doi.org/10.1037/0022-006x.71.6.1058

Clark, D.M., & Wells, A. (1995). A cognitive model of social phobia. In R. G. Heimberg, M. R. Liebowitz, D.A. Hope, & F.R. Schneier (Eds.), *Social phobia: Diagnosis, assessment and treatment* (1st ed., pp. 69-93). The Guilford Press.

Datta, S., & Das, S. (2016). Cognitive behaviour therapy in social phobia: A case study. *SIS Journal of Projective Psychology & Mental Health*, *23*(3), 88-95.

Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D.W., Oishi, S., & Biswas-Diener, R. (2010). New well–being measures: Short scales to assess flourishing and positive and negative feelings. *Social Indicators Research*, *97*(2), 143-156. https://doi.org/10.1007/s11205-009-9493-y

Faravelli, C., Zucchi, T., Viviani, B., Salmoria, R., Perone, A., Paionni, A., Scarpato, A., Vigliaturo, D., Rosi, S., D'adamo, D., Bartolozzi, D., Cecchi, C., & Abrardi, L. (2000). Epidemiology of social phobia: A clinical approach. *European Psychiatry*, *15*(1), 17-24. https://doi.org/10.1016/S0924-9338(00)00215-7

Fava, G.A., Rafanelli, C., Cazzaro, M., Conti, S., & Grandi, S. (1998). Well-being therapy. A novel psychotherapeutic approach for residual symptoms of affective disorders. *Psychological Medicine*, *28*(2), 475-480. https://doi.org/10.1017/S0033291797006363

Fehm, L., Beesdo, K., Jacobi, F., & Fiedler, A. (2008). Social anxiety disorder above and below the diagnostic threshold: prevalence, comorbidity and impairment in the general population. *Social Psychiatry and Psychiatric Epidemiology*, *43*(4), 257-265. https://doi.org/10.1007/s00127-007-0299-4

Freeman, D., Pugh, K., Dunn, G., Evans, N., Sheaves, B., Waite, F., Černis, E.R.L., & Fowler, D. (2014). An early phase II randomised controlled trial testing the effect on persecutory delusions of using CBT to reduce negative cognitions about the self: The potential benefits of enhancing self confidence. *Schizophrenia Research*, *160*(1–3), 186-192 https://doi.org/10.1016/j.schres.2014.10.038

Frets, P.G., Kevenaar, C., & van der Heiden, C. (2014). Imagery rescripting as a stand-alone treatment for patients with social phobia: A case series. *Journal of Behavior Therapy and Experimental Psychiatry*, *45*(1), 160-169. https://doi.org/10.1016/j.jbtep.2013.09.006

Gaudiano, B.A., & Herbert, J.D. (2003). Preliminary psychometric evaluation of a new self-efficacy scale and its relationship to treatment outcome in social anxiety disorder. *Cognitive Therapy and Research*, *27*(5), 537-555. https://doi.org/10.1023/A:1026355004548

Ghaedi, G., Tavoli, A., Bakhtiari, M., Melyani, M., & Sahragard, M. (2010). Quality of life in college students with and without social phobia. *Social Indicators Research*, *97*(2), 247-256. https://doi.org/10.1007/s11205-009-9500-3

Gil, P.J.M., Carrillo, F.X.M., & Meca, J.S. (2001). Effectiveness of cognitive-behavioural treatment in social phobia: A meta-analytic review. *Psychology in Spain*, *5*(1), 17-25.

Goldin, P.R., Ziv, M., Jazaieri, H., Hahn, K., Heimberg, R., & Gross, J.J. (2013). Impact of cognitive behavioral therapy for social anxiety disorder on the neural dynamics of cognitive reappraisal of negative self-beliefs: Randomized clinical trial. *Jama Psychiatry*, *70*(10), 1048-1056. https://doi.org/10.1001/jamapsychiatry.2013.234

Goldin, P.R., Ziv, M., Jazaieri, H., Werner, K., Kraemer, H., Heimberg, R.G., & Gross, J.J. (2012). Cognitive reappraisal self-efficacy mediates the effects of individual cognitive-behavioral therapy for social anxiety disorder. *Journal of Consulting and Clinical Psychology*, *80*(6), 1034-1040. https://doi.org/10.1037/a0028555

Gordon, D., Wong, J., & Heimberg, R.G. (2014). Cognitive–behavioral therapy for social anxiety disorder: The state of the science. In J.W. Weeks (Ed.), *The wiley blackwell handbook of social anxiety disorder* (1st ed., pp. 477–497). John Wiley & Sons.

Greenberg, L.S. (2010). Emotion-focused therapy: A clinical synthesis. *The Journal of Lifelong Learning in Psychiatry*, *81*(1), 32-42. https://doi.org/10.1176/foc.8.1.foc32

Gultekin, B.K., & Dereboy, I.F. (2011). Üniversite öğrencilerinde sosyal fobinin yaygınlığı ve sosyal fobinin yaşam kalitesi, akademik başarı ve kimlik oluşumu üzerine etkileri [The prevalence of social phobia, and its impact on quality of life, academic achievement, and identity formation in university students]. *Turk Psikiyatri Dergisi*, *22*(3), 150-158. https://www.turkpsikiyatri.com/PDF/C22S3/150-158

Hackmann, A., Clark, D.M., & McManus, F. (2000). Recurrent images and early memories in social phobia. *Behaviour Research and Therapy*, *38*(6), 601-610. https://doi.org/10.1016/S0005-7967(99)00161-8

Hackmann, A., Surawy, C., & Clark, D.M. (1998). Seeing yourself through others' eyes: A study of spontaneously occurring images in social phobia. *Behavioural and Cognitive Psychotherapy*, *26*, 3-12. https://doi.org/10.1017/S1352465898000022

Harwood, T.M., Beutler, L.E., & Charvat, M. (2010). Cognitive–behavioral therapy and psychotherapy integration. In K. S. Dobson (Ed.), *Handbook of cognitive–behavioral therapies* (3rd ed., pp. 94-130). The Guilford Press.

Hirsch, C., Meynen, T., & Clark, D. (2004). Negative self-imagery in social anxiety contaminates social interactions. *Memory*, *12*(4), 496 506. https://doi.org/10.1080/09658210444000106

Hirsch, C.R., Clark, D.M., Mathews, A., & Williams, R. (2003). Self-images play a causal role in social phobia. *Behaviour Research and Therapy*, *41*(8), 909-921. https://doi.org/10.1016/S0005-7967(02)00103-1

Hirsch, C.R., Mathews, A., Clark, D.M., Williams, R., & Morrison, J.A. (2006). The causal role of negative imagery in social anxiety: A test in confident public speakers. *Journal of Behavior Therapy and Experimental Psychiatry*, *37*(2), 159-170. https://doi.org/10.1016/j.jbtep.2005.03.003

Holaway, R.M., & Heimberg, R.G. (2004). Cognitive–behavioral therapy for social anxiety disorder: A treatment review. In B. Bandelow & D.J. Stein (Eds.), *Social anxiety disorder* (1st ed., pp. 207–220).

Howes, J.L., & Parrott, C.A. (1991). Conceptualization and flexibility in cognitive therapy. In T.M. Vallis, J.L. Howes, & P.C. Miller (Eds.), *The challenge of cognitive therapy: Applications to nontraditional populations* (pp. 25-42). Springer.

Huppert, F.A. (2009). Psychological well-being: Evidence regarding its causes and consequences. *Applied Psychology: Health and Well Being*, *1*(2), 137-164. https://doi.org/10.1111/j.1758-0854.2009.01008.x

Izgic, F., Akyuz, G., Dogan, O., & Kugu, N. (2004). Social phobia among university students and its relation to self-esteem and body image. *The Canadian Journal of Psychiatry*, *49*(9), 630-634. https://doi.org/10.1177/070674370404900910

Jacobson, N.S., Follette, W.C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy*, *15*(4), 336-352. https://doi.org/10.1016/S0005-7894(84)80002-7

Jacobson, N.S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*(1), 12-19. https://doi.org/10.1037/0022-006x.59.1.12

Jafari, M., Shahidi, S., & Abedin, A. (2012). Comparing the effectiveness of cognitive behavioral therapy and stages of change model on improving abstinence self-efficacy in Iranian substance dependent adolescents. *Iranian Journal of Psychiatry and Behavioral Sciences*, *6*(2), 7-15.

Kashdan, T.B., Julian, T., Merritt, K., & Uswatte, G. (2006). Social anxiety and posttraumatic stress in combat veterans: Relations to well-being and character strengths. *Behaviour Research and Therapy*, *44*(4), 561-583. https://doi.org/10.1016/j.brat.2005.03.010

Kellogg, S. (2004). Dialogical encounters: Contemporary perspectives on "chairwork" in psychotherapy. *Psychotherapy: Theory, Research, Practice, Training*, *41*(3), 310-320. https://doi.org/10.1037/0033-3204.41.3.310

Keshi, A.K., & Basavarajappa. (2013). Effectiveness of cognitive behavior therapy on self-efficacy among high school students. *Asian Journal of Management Sciences & Education*, *2*(4), 68–79.

Kessler, R.C., Petukhova, M., Sampson, N.A., Zaslavsky, A.M., & Wittchen, H.U. (2012). Twelve-month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the United States. *International Journal of Methods in Psychiatric Research*, *21*(3), 169-184. https://doi.org/10.1002/mpr.1359

Knutsson, J., Nilsson, J.E., Eriksson, Å., & Järild, L. (2020). Imagery rescripting and exposure in social anxiety: A randomized trial comparing treatment techniques. *Journal of Contemporary Psychotherapy*, *50*(3), 233-240. https://doi.org/10.1007/s10879-019-09448-1

Krans, J., de Bree, J., & Bryant, R.A. (2014). Autobiographical memory bias in social anxiety. *Memory*, *22*(8), 890-897. https://doi.org/10.1080/09658211.2013.844261

Kring, A.M., & Johnson, S.L. (2014). *Abnormal Psychology* (12th ed.). John Wiley & Sons, Inc.

Kuo, J.R., Goldin, P.R., Werner, K., Heimberg, R.G., & Gross, J.J. (2011). Childhood trauma and current psychological functioning in adults with social anxiety disorder. *Journal of Anxiety Disorders*, *25*(4), 467-473. https://doi.org/10.1016/j.janxdis.2010.11.011

Lee, S.W., & Kwon, J.H. (2013). The efficacy of imagery rescripting (IR) for social phobia: A randomized controlled trial. *Journal of Behavior Therapy and Experimental Psychiatry*, *44*(4), 351-360. https://doi.org/10.1016/j.jbtep.2013.03.001

Leigh, E., & Clark, D.M. (2016). Cognitive therapy for social anxiety disorder in adolescents: A development case series. *Behavioural and Cognitive Psychotherapy*, *44*(1), 1-17. https://doi.org/10.1017/S1352465815000715

Liebowitz, M.R. (1987). Social phobia. *Modern Problems of Pharmacopsychiatry*, *22*, 141-173. https://doi.org/10.1159/000414022

Luszczynska, A., Gutierrez-Dona, B., & Schwarzer, R. (2005). General self-efficacy in various domains of human functioning: Evidence from five countries. *International Journal of Psychology*, *40*(2), 80-89. https://doi.org/10.1080/00207590444000041

Magee, W.J. (1999). Effects of negative life experiences on phobia onset. *Social Psychiatry and Psychiatric Epidemiology*, *34*(7), 343-351. https://doi.org/10.1007/s001270050154

Makkar, S.R., & Grisham, J.R. (2011). Social anxiety and the effects of negative self-imagery on emotion, cognition, and post-event processing. *Behaviour Research and Therapy*, *49*(10), 654-664. https://doi.org/10.1016/j.brat.2011.07.004

Maner, J.K., & Kenrick, D.T. (2010). When adaptations go awry: Functional and dysfunctional aspects of social anxiety. *Social Issues and Policy Review*, *4*(1), 111-142. https://doi.org/10.1111/j.1751-2409.2010.01019.x

Mayo-Wilson, E., Dias, S., Mavranezouli, I., Kew, K., Clark, D.M., Ades, A.E., & Pilling, S. (2014). Psychological and pharmacological interventions for social anxiety disorder in adults: A systematic review and network meta-analysis. *The Lancet Psychiatry*, *1*(5), 368-376. https://doi.org/10.1016/S2215-0366(14)70329-3

McManus, F., Sacadura, C., & Clark, D.M. (2008). Why social anxiety persists: An experimental investigation of the role of safety behaviours as a maintaining factor. *Journal of Behavior Therapy and Experimental Psychiatry*, *39*(2), 147-161. https://doi.org/10.1016/j.jbtep.2006.12.002

Melkman, E.P. (2017). Childhood adversity, social support networks and well-being among youth aging out of care: An exploratory study of mediation. *Child Abuse & Neglect*, *72*, 85–97. https://doi.org/10.1016/j.chiabu.2017.07.020

Mellings, T.M.B., & Alden, L.E. (2000). Cognitive processes in social anxiety: The effects of self-focus, rumination and anticipatory processing. *Behaviour Research and Therapy*, *38*(3), 243-257. https://doi.org/10.1016/S0005-7967(99)00040-6

Mosley-Johnson, E., Garacci, E., Wagner, N., Mendez, C., Williams, J.S., & Egede, L.E. (2018). Assessing the relationship between adverse childhood experiences and life satisfaction, psychological well–being, and social well–being: United States longitudinal cohort 1995–2014. *Quality of Life Research, Advance online publication*. https://doi.org/10.1007/s11136-018-2054-6

Mörtberg, E., Clark, D.M., Sundin, Ö., & Wistedt, A.Å. (2007). Intensive group cognitive treatment and individual cognitive therapy vs. treatment as usual in social phobia: A randomized controlled trial. *Acta Psychiatrica Scandinavica*, *115*(2), 142-154. https://doi.org/10.1111/j.1600-0447.2006.00839.x

Narr, R.K., & Teachman, B.A. (2017). Using advances from cognitive behavioral models of anxiety to guide treatment for social anxiety disorder. *Journal of Clinical Psychology*, *73*(5), 524-535. https://doi.org/10.1002/jclp.22450

Nilsson, J.E., Lundh, L.G., & Viborg, G. (2012). Imagery rescripting of early memories in social anxiety disorder: An experimental study. *Behaviour Research and Therapy*, *50*(6), 387-392. https://doi.org/10.1016/j.brat.2012.03.004

Nordstrom, A.H., Goguen, L.M.S., & Hiester, M. (2014). The effect of social anxiety and self–esteem on college adjustment, academics, and retention. *Journal of College Counseling*, *17*(1), 48-63. https://doi.org/10.1002/j.2161-1882.2014.00047.x

Norton, A.R., & Abbott, M.J. (2016). The efficacy of imagery rescripting compared to cognitive restructuring for social anxiety disorder. *Journal of Anxiety Disorders*, *40*, 18-28. https://doi.org/10.1016/j.janxdis.2016.03.009

Norton, A.R., Abbott, M.J., Dobinson, K.A., Pepper, K.L., & Guastella, A.J. (2021). Rescripting social trauma: A pilot study investigating imagery rescripting as an adjunct to cognitive behaviour therapy for social anxiety disorder. *Cognitive Therapy and Research*, *45*(6), 1180-1192. https://doi.org/10.1007/s10608-021-10221-9

Özbay, Y., & Palancı, M. (2001). *Sosyal kaygı ölçeği: Geçerlik ve güvenirlik çalışması [Social anxiety scale: The validity and reliability study]* VI. Ulusal Psikolojik Danışma ve Rehberlik Kongresi [VI. National Psychological Counseling and Guidance Congress], Ankara, Turkey.

Paivio, S.C., Hall, I.E., Holowaty, K.A.M., Jellis, J.B., & Tran, N. (2001). Imaginal confrontation for resolving child abuse issues. *Psychotherapy Research*, *11*(4), 433-453. https://doi.org/10.1093/ptr/11.4.433

Pinjarkar, R.G., Sudhir, P.M., & Math, S.B. (2015). Brief cognitive behavior therapy in patients with social anxiety disorder: A preliminary investigation. *Indian Journal of Psychological Medicine*, *37*(1), 20-25. https://doi.org/10.4103/0253-7176.150808

Priyamvada, R., Kumari, S., Prakash, J., & Chaudhury, S. (2009). Cognitive behavioral therapy in the treatment of social phobia. *Industrial Psychiatry Journal*, *18*(1), 60-63. https://doi.org/10.4103/0972-6748.57863

Regis, J.M.O., Ramos-Cerqueira, A.T.A., Lima, M.C.P., & Torres, A.R. (2018). Social anxiety symptoms and body image dissatisfaction in medical students: Prevalence and correlates. *Jornal Brasileiro de Psiquiatria*, *67*, 65-73. https://doi.org/10.1590/0047-2085000000187

Reimer, S.G., & Moscovitch, D.A. (2015). The impact of imagery rescripting on memory appraisals and core beliefs in social anxiety disorder. *Behaviour Research and Therapy*, *75*, 48-59. https://doi.org/10.1016/j.brat.2015.10.007

Romano, M., Moscovitch, D.A., Huppert, J.D., Reimer, S.G., & Moscovitch, M. (2020). The effects of imagery rescripting on memory outcomes in social anxiety disorder. *Journal of Anxiety Disorders*, *69*. https://doi.org/10.1016/j.janxdis.2019.102169

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Revised ed., Vol. 6). Sage Publications.

Rudy, B.M., Davis Ill, T.E., & Matthews, R.A. (2012). The relationship among self-efficacy, negative self-referent cognitions, and social anxiety in children: A multiple mediator model. *Behavior Therapy*, *43*(3), 619-628. https://doi.org/10.1016/j.beth.2011.11.003

Ruini, C., Belaise, C., Brombin, C., Caffo, E., & Fava, G.A. (2006). Well-being therapy in school settings: A pilot study. *Psychotherapy and Psychosomatics*, *75*(6), 331-336. https://doi.org/10.1159/000095438

Ruscio, A.M., Brown, T.A., Chiu, W.T., Sareen, J., Stein, M.B., & Kessler, R.C. (2008). Social fears and social phobia in the USA: results from the National Comorbidity Survey Replication. *Psychological Medicine*, *38*(1), 15-28. https://doi.org/10.1017/S0033291707001699

Ryff, C.D. (1989a). Beyond ponce de leon and life satisfaction: New directions in quest of successful ageing. *International Journal of Behavioral Development*, *12*(1), 35–55. https://doi.org/10.1177/016502548901200102

Ryff, C.D. (1989b). Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *Journal of Personality and Social Psychology*, *57*(6), 1069-1081. https://doi.org/10.1037/0022-3514.57.6.1069

Ryff, C.D. (2014). Psychological well–being revisited: Advances in the science and practice of eudaimonia. *Psychotherapy and Psychosomatics*, *83*(1), 10-28. https://doi.org/10.1159/000353263

Scherbaum, C.A., Cohen-Charash, Y., & Kern, M.J. (2006). Measuring general self-efficacy: A comparison of three measures using item response theory. *Educational and Psychological Measurement*, *66*(6), 1047-1063. https://doi.org/10.1177/0013164406288171

Schwarzer, R. (1994). Optimism, vulnerability, and self-beliefs as health-related cognitions: A systematic overview. *Psychology and Health*, *9*(3), 161-180. https://doi.org/10.1080/08870449408407475

Schwarzer, R., & Jerusalem, M. (1995). Generalized self–efficacy scale. In J. Weinman, S.C. Wright, & M. Johnston (Eds.), *Measures in health psychology: A user's portfolio* (1st ed., pp. 35–37). NFER–Nelson.

Shorey, R.C., & Stuart, G.L. (2012). Manualized cognitive-behavioral treatment of social anxiety disorder: A case study. *Clinical Case Studies*, *11*(1), 35-47. https://doi.org/10.1177/1534650112438462

Soykan, C., Ozguven, H. D., & Gencoz, T. (2003). Liebowitz social anxiety scale: The Turkish version. *Psychological Reports*, *93*(3), 1059-1069. https://doi.org/10.2466/Pr0.93.7.1059-1069

Stangier, U., Heidenreich, T., Peitz, M., Lauterbach, W., & Clark, D.M. (2003). Cognitive therapy for social phobia: Individual versus group treatment. *Behaviour Research and Therapy*, *41*(9), 991-1007. https://doi.org/10.1016/S0005-7967(02)00176-6

Stopa, L., & Jenkins, A. (2007). Images of the self in social anxiety: Effects on the retrieval of autobiographical memories. *Journal of Behavior Therapy and Experimental Psychiatry*, *38*(4), 459-473. https://doi.org/10.1016/j.jbtep.2007.08.006

Taylor, S. (1996). Meta-analysis of cognitive-behavioral treatments for social phobia. *Journal of Behavior Therapy and Experimental Psychiatry*, *27*(1), 1-9. https://doi.org/10.1016/0005-7916(95)00058-5

Telef, B.B. (2013). Psikolojik iyi oluş ölçeği: Türkçeye uyarlama, geçerlik ve güvenirlik çalışması [The Adaptation of psychological well-being into Turkish: A validity and reliability study]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *28*(3), 374-384. https://dergipark.org.tr/tr/download/article-file/87222

Tsitsas, G.D., & Paschali, A.A. (2014). A Cognitive–behavior therapy applied to a social anxiety disorder and a specific phobia: Case study. *Health Psychology Research*, *2*(1603), 78-82. https://doi.org/10.4081/hpr.2014.1603

Vassilopoulos, S.P. (2012). Social anxiety and memory biases in middle childhood: A preliminary study. *Hellenic Journal of Psychology*, *9*(2), 114-131.

Wang, J.L., Jackson, L.A., Gaskin, J., & Wang, H.Z. (2014). The effects of Social Networking Site (SNS) use on college students' friendship and well-being. *Computers in Human Behavior*, *37*, 229-236. https://doi.org/10.1016/j.chb.2014.04.051

Weiss, B.J., Singh, J.S., & Hope, D.A. (2011). Cognitive–behavioral therapy for immigrants presenting with social anxiety disorder: Two case studies. *Clinical Case Studies*, *10*(4), 324-342. https://doi.org/10.1177/1534650111420706

Weiss, L.A., Westerhof, G.J., & Bohlmeijer, E.T. (2016). Can we increase psychological well–being? The effects of interventions on psychological well-being: A meta-analysis of randomized controlled trials. *Plos ONE*, *11*(6). https://doi.org/10.1371/journal.pone.0158092

Wersebe, H., Lieb, R., Meyer, A.H., Miche, M., Mikoteit, T., Imboden, C., Hoyer, J., Bader, K., Hatzinger, M., & Gloster, A. T. (2018). Well-being in major depression and social phobia with and without comorbidity. *International Journal of Clinical and Health Psychology*, *18*(3), 201-208. https://doi.org/10.1016/j.ijchp.2018.06.004

Wild, J., & Clark, D.M. (2011). Imagery rescripting of early traumatic memories in social phobia. *Cognitive and Behavioral Practice*, *18*(4), 433-443. https://doi.org/10.1016/j.cbpra.2011.03.002

Wild, J., & Clark, D.M. (2015). Experiential exercises and imagery rescripting in social anxiety disorder: New perspectives on changing beliefs. In N. C. Thoma & D. McKay (Eds.), *Working with emotion in cognitive-behavioral therapy: Techniques for clinical practice* (1st ed., pp. 216-236). The Guilford Press.

Wild, J., Hackmann, A., & Clark, D.M. (2007). When the present visits the past: Updating traumatic memories in social phobia. *Journal of Behavior Therapy and Experimental Psychiatry*, *38*(4), 386-401. https://doi.org/10.1016/j.jbtep.2007.07.003

Wild, J., Hackmann, A., & Clark, D.M. (2008). Rescripting early memories linked to negative images in social phobia: A pilot study. *Behavior Therapy*, *39*(1), 47-56. https://doi.org/10.1016/j.beth.2007.04.003

Wootton, B. M., Hunn, A., Moody, A., Lusk, B. R., Ranson, V. A., & Felmingham, K. L. (2018). Accelerated outpatient individual cognitive behavioural therapy for social anxiety disorder: A preliminary pilot study. *Behavioural and Cognitive Psychotherapy*, *46*(6), 690-705. https://doi.org/10.1017/S1352465818000267

Yoshinaga, N., Kobori, O., Iyo, M., & Shimizu, E. (2013). Cognitive behaviour therapy using the Clark & Wells model: A case study of a Japanese social anxiety disorder patient. *The Cognitive Behaviour Therapist*, *6*(3), 1-30. https://doi.org/10.1017/S1754470X13000081

Yoshinaga, N., Ohshima, F., Matsuki, S., Tanaka, M., Kobayashi, T., Ibuki, H., Asano, K., Kobori, O., Shiraishi, T., Ito, E., Nakazato, M., Nakagawa, A., Iyo, M., & Shimizu, E. (2013). A preliminary study of individual cognitive behavior therapy for social anxiety disorder in Japanese clinical settings: A single–arm, uncontrolled trial. *BMC Research Notes*, *6*(74), 1-8. https://doi.org/10.1186/1756-0500-6-74

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 1, 203–219

https://doi.org/10.21449/ijate.894688

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

Development and Validation of a Reading Self-Efficacy Scale

Gulten Kosar[1,*], **Yunus Emre Akbana**[2], **Levent Yakar**[3]

[1]Hatay Mustafa Kemal University, Faculty of Edu., Dep. of English Lang. Teaching, Hatay, Turkiye
[2]Kahramanmaras Sütçü Imam University, Faculty of Edu., Dep. of English Lan. Teac., Kahramanmaras, Turkiye
[3]Kahramanmaras Sütçü Imam University, Faculty of Edu., Dep. of Edu. Sciences, Kahramanmaras, Turkiye

ARTICLE HISTORY

Received: Mar. 10, 2021
Revised: Nov. 20, 2021
Accepted: Jan. 24, 2022

Keywords:

College students,
English language teaching,
Reading self-efficacy.

Abstract: Reading self-efficacy performs a fundamental role in gaining academic achievement in college education. Review of the related literature unveils that it needs to be enriched by conducting further research on college students' reading self-efficacy. The paucity of investigations into college students' reading self-efficacy could have a connection with the lack of a comprehensive reading self-efficacy scale targeting exclusively measuring it. For this reason, this study aims at developing and validating a reading self-efficacy scale which could be used to measure college students' reading self-efficacy. The data was collected from three distinct groups consisting of a total of 430 students of the departments of English language teaching and English language and literature studying at state universities in Turkey. The findings obtained from exploratory factor analysis revealed that the scale had a unidimensional structure and the ones provided by confirmatory factor analysis confirmed the structure of the scale. The developed and validated 16-item reading self-efficacy scale could prompt the university teachers of reading to undertake studies with an eye to examining their students' reading self-efficacy.

1. INTRODUCTION

Reading comprehension is a fundamental skill to achieve success at tertiary level (Meniado, 2016). A reader's prior knowledge (Kintsch, 1998), cognitive skills and repertoire of reading strategies (Duke & Pearson, 2002), reading proficiency (Mills et al., 2006), the text, reading context and motivation (Afflerbach & Cho, 2010), and the online/offline reading modes (Forzani et al., 2021) are among the determining factors in comprehending a reading. Being literate in a second language (L2) or foreign language (FL) requires an individual to overcome several challenges as opposed to reading in L1 due to the intricacies of L2/FL reading and the complex cognitive processes (Graham et al., 2020; Murad Sani & Zain, 2011). For example, linguistic, affective and motivational factors affect reading comprehension in L2/FL (Grabe & Stoller, 2019; Li & Wang, 2010). Additionally, other key factors such as comprehension strategies (Taylor, 2014), language proficiency (Fung & Macaro, 2019), self-regulatory reading strategies (Macaro & Erler, 2008), self-efficacy and strategy use (Zimmerman, 2013), and strategy instruction for developing self-efficacy (Gu, 2019) are vital to reading achievement in L2/FL (Graham et al., 2020).

*CONTACT: Gülten KOŞAR ✉ gulten.kosar@mku.edu.tr Hatay Mustafa Kemal University, Faculty of Education, Department of English Language Teaching, Hatay, Turkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

Achievement in reading is pertinent to reading self-efficacy (Tavakoli & Koosha, 2016). The association between these two constructs becomes perfectly clear considering the positive correlation between them; that is to say, the higher one's reading self-efficacy, the higher their reading achievement is (Hedges & Gable, 2016). High level of self-efficacy is revealed to contribute to learners' growth in reading (Capara et al., 2008). Several studies have reported the higher self-efficacy beliefs, the better results in developing reading comprehension (Bağcı, 2019; Barkley, 2006; Cho et al., 2015; Forzani et al., 2021; Hedges & Gable, 2016; McLean & Poulshock, 2018; Peura et al., 2019; Ronimus et al., 2020; Soland & Sandilos, 2020; Tremblay & Gardner, 1995; Unrau et al., 2018). Basically, attitudes towards reading could influence reading self-efficacy of an individual; for instance, individuals who enjoy reading are shown to have higher levels of reading self-efficacy in comparison to the ones not enjoying reading (Burrows, 2012; Carroll & Fox, 2017).

Efficacy beliefs can range from weak to strong due to the specificity and the level of task difficulty (Bandura, 1997). Self-efficacy is defined as an individual's own judgments of their perceived self-capabilities to perform tasks or actions at the designated levels (Bandura, 1997; Schunk, 2003). According to Unrau et al. (2018), reading self-efficacy is "readers' perceptions of competence in their ability to successfully complete reading tasks" (p. 168). To measure reading self-efficacy, Carroll and Fox (2017) argue that there is still a strong need for "… a specific single measure of reading self-efficacy" (p. 2) because previous studies do not purely focus on such a single measure of self-efficacy. As Bandura (1997) outlines, three levels of specificity exist to be resorted to in the measurement of efficacy beliefs: general, intermediate and specific. The specificity level of general refers to the general beliefs of one's efficacy and the specificity level of intermediate describes the performances under a domain of activity while that of specific addresses the completion of specific tasks. Drawing on Peura et al.'s (2019) study evaluating specificity levels of reading self-efficacy scales, it could be argued that the scale this study develops falls into the specificity level of intermediate. Furthermore, to prepare a good self-efficacy scale, Bandura (1997, 2006) suggested several principles. To list, the items should be phrased in a clear and unambiguous language; written in the form of *can do* statements; prepared for a specific domain with the content of tasks; challenging to avoid ceiling effect; asking one's competences; arranged randomly or from the least to the most challenging employing content focusing on beliefs rather than self-worth, locus of control or outcome expectancies; and further, it should employ predictive power by testing what it intends to measure and include only necessary sub-skills or one operative efficacy.

There have been studies extensively focusing on L1 reading self-efficacy and its relation to reading attitudes and reading motivation (Hedges & Gable, 2016). For example, Ghonsooly and Elahi (2010) investigated the relationships among reading self-efficacy, reading anxiety and reading achievement. To do so, the authors designed a reading self-efficacy scale based on three other scales available in the literature and adapted it in Persian language. The scale had 11 items with a four-factor structure involving the dimensions of students' ability in reading English texts, student's inability in reading English texts, practice and skill and enjoying group work. The Cronbach's alpha coefficient for their scale was satisfying (α = 0.78), but the number of participants was only 150 second-year college students. In addition, Peura et al. (2019) developed reading self-efficacy and reading comprehension scale for young learners, focusing on reading fluency. There are also several studies measuring L1 reading self-efficacy construct in Turkish context. For instance, Kula and Budak (2020) developed a scale to measure reading self-efficacy perceptions of 525 primary school students. The scale had a single-factor structure and consisted of 29 items written in the form of *can do* statements. The scale was designed as a 3-point ikert scale rather than a 5-point one. Besides, Karabay (2013) developed "Critical Reading Self-Efficacy Perception Scale" with the participation of 650 pre-service Turkish language teachers. The scale involved 41 items with three sub-dimensions; evaluation, research, and visual themes. In addition, Şahin and Öztahtalı (2019) developed "Effective Reading Self-Efficacy Perception Scale" with 677 high-school students. The scale had a four-factor structure with the dimensions of comprehension, breathing, pausing and appearance, and consisted of 21 items with a 5-point Likert scale but not in the form of *can do* statements.

Regarding target language reading self-efficacy scales, Ahmadian and Pasand's (2017) scale, which was originally developed by Zare and Davoudi Mobarakeh (2011), is satisfactory with regard to content, validity and favorable particularly for L2/FL learners with intermediate level of English. Secondly, Kakaew and Damnet's (2017) reading self-efficacy scale is available in the literature but it does not align with Bandura's suggestions because the scale incorporates only three *can do* statements and measures constructs other than reading self-efficacy such as self-worth and locus of control. Thirdly, the reading self-efficacy scale used in Boakye's (2015) study was originally developed from the works of Grabe and Stoller (2002), and Guthrieet al. (2000). However, it might fail to measure reading self-efficacy because it addresses the common reading problems among students at risk of academic failure. Additionally, it fails to measure competency beliefs because it employs items focusing on self-worth rather than the context governing the challenge of the tasks. Fourthly, the scale used in McLean and Poulshock's (2018) study was originally developed by Burrows (2012), which fits into Bandura's guidelines except for its being developed with non-English major students. Finally, the L2 reading self-efficacy questionnaire developed by Mullins (2018) in Spanish aligns with Bandura's principles in that it employs *can do* statements, developed for a specific target group (*viz.* novice learners), involves unambiguous items, and has potentially rich response use (*viz.* 100-Likert scale) but Mullins (2019) criticizes the first 8 items which focus on the components of reading rather than the domain of reading self-efficacy.

Reading self-efficacy is a significant construct that needs to be measured accurately, bringing to the forefront the need for the instruments specifically developed to assess it. Review of literature on reading self-efficacy instruments prompted the researchers to conduct research to develop and validate an FL/L2 reading self-efficacy scale with an eye to measuring the level of reading self-efficacy of college students particularly majoring in foreign language teaching and literature departments. In view of the lack of a reading self-efficacy scale developed and validated through the participation of students at different years of study in the aforementioned majors, the deficiencies in the existing reading self-efficacy scales pinpointed in the preceding paragraph and the comprehensiveness of the items in the scale developed and validated in this study, it could be argued that this research could contribute a lot to the literature on reading self-efficacy in FL/L2.

2. METHOD

2.1. Study Groups

Three different data sets were collected from three distinct groups to perform exploratory factor analysis (EFA), confirmatory factor analysis (CFA) and test-retest reliability.

2.1.1. *First group*

The first group was comprised of a cohort of 180 participants in the selection of whom convenience sampling was used. 100 (56%) of the participants were the students of English language teaching department and the remaining 80 (44%) participants were the students studying at the department of English language and literature at four state universities in Turkey. 46 (26%) of the participants were first-year students (female students = 28, male students = 18, mean age = 18.8). 50 (28%) students in the first group were second-year students (female students = 31, male students = 19, mean age = 19.2), and 42 (23%) students were third-year students (female students = 27, male students = 15, mean age = 21.2), and the rest of the students in the first group were 42 (23%) fourth-year students (female students = 25, male students = 17, mean age = 22.4). Students' level of proficiency in English was not a parameter taken into consideration in choosing participants. The data collected from the first group was used for conducting EFA.

2.1.2. *Second group*

Second group involved 153 students of the departments of English language teaching and English language and literature studying at four state universities in Turkey, a group different from the first one and selected through convenience sampling. 38 (25%) students were first-year students (female students = 26, male students = 12, mean age = 18.7), and 44 (29%) students were second-year students (female students = 28, male students = 16, mean age = 19.8). 35 (22%) students were third-year students (female students = 19, male students = 16, mean age = 21.1), and the remaining 36 (24%) participants were fourth-year students (female students = 23, male students =13, mean age = 22.8). The data gathered from the second group was used for performing CFA.

2.1.3. *Third group*

Third group consisted of 65 (67%) first-year and 32 (33%) fourth-year students of English language teaching department studying at two state universities in Turkey (female students = 70, male students = 27, mean age = 20.59). Convenience sampling was used to select the third group participants, akin to the sampling used in the selection of the first and second group participants. The data collected from the group was analyzed to measure test-retest reliability coefficient.

2.2. Data Collection Tool

Since this study aims to design a reading self-efficacy scale with the participation of college students majoring at the departments of English language teaching and English language and literature, the development of the data collection tool was commenced by reading books on teaching reading (Blachowicz & Ogle, 2008; Butterworth & Thwaites, 2013; Carter, 2011; Dallon & Ratner, 2002; Guthrie & Taboada, 2004; McGuinness, 2005; Palincsar & Brown, 1986; Smith, 2012). Aside from reading books on the teaching of reading, the questionnaires and scales having been developed to measure reading self-efficacy thus far and the research having been undertaken on reading self-efficacy were analyzed (*e.g.* Ahmadian & Pasand, 2017; Barber et al., 2015; Boakye, 2015; Ferrara, 2007; Mullins, 2018; Peura et al., 2009; Solheim, 2011; Venegas, 2018). While reading the literature pertaining to reading self-efficacy, factor structures and response scales were reviewed in order to seek for potential dimensions. According to this review, only two studies mentioned the factor structure of the reading self-efficacy scale. Soolheim (2011) reported the reading self-efficacy scale had one dimension and 5-step Likert continuum (From 1 "*I don't agree*" to 5 "*I agree*") and Peura et al. (2019) identified a three-hypothetical-factor structure in the scale and related the dimensions to the general, intermediate and specific levels of specificity as outlined by Bandura (1997). Peura et al. (2019) used a 7-point scale for responses (From 1 "*I'm totally certain I can't*" to 7 "*I'm totally certain I can*"). The factor structure of the 5- point Likert scale (From 1 "*Strongly agree*", to 5 "*Strongly disagree*") in Ghonsooly and Elahi' (2010) study was found to include the dimensions of students' ability in reading English texts, student's inability in reading English texts, practice and skill and enjoying group work. A single-factor structure was identified in Kula and Budak's (2020) 3-point Likert scale (1 "*Doesn't fit me*", 2 "*Fits me a little*", 3 "*Fits me completely*"). Also, evaluation, research, and visual themes were detected in Karabay's (2013) 5-point Likert scale (From 1 "*Never*" to 5 "*Always*") and finally comprehension, breathing, pausing and appearance were presented to be the four dimensions in Şahin and Öztahtalı's (2019) 5-point Likert scale starting with "*Always* (1)" and moving towards "*Never* (5)". Nonetheless, in their examination of other scales, Peura et al. (2019) found that most of the reading self-efficacy studies did not report the factor structure and the level of specificity of reading self-efficacy scales. This encouraged the researchers to state the factor structure and specificity level of the developed and validated reading self-efficacy scale in this paper. Additionally, in light of the literature on reading self-efficacy, an item pool including 16 items in *can do statements* and 5-point Likert scale (*Not at all* (1), *Slightly* (2), *Somewhat* (3), *Fairly well* (4), *Very well* (5) was generated. The items were produced in light of Bandura's (1997, 2006) principles of developing items in a reading self-efficacy scale.

The generated items were e-mailed to three pre-service English language teacher educators teaching Reading Skills and Critical Reading and Writing courses for more than 10 years at three different state universities in an effort to ensure content validity. The teacher educators judged that all the constructed items served for measuring college students' reading self-efficacy and were adequate for accurate measurement of it. As a result of the expert opinions, the scale was administered to 20 students and the feedback collected from them indicated no need for amendments in the items (See Appendix). Once the ethical approval was obtained from Hatay Mustafa Kemal University Social and Human Sciences Research and Publication Ethics Board (document no. 902-01-FR 006), the data from three groups were collected first in the third week of October, 2020; then, data from the third group were gathered again in the first week of November, 2020 to perform test-retest reliability coefficient. The data were obtained through Google Forms ensuring students' anonymity and consent.

2.3. Data Analysis

The three distinct sets of data were subjected to a different analysis. First of all, data 1 was analyzed through performing EFA with a view to figuring out the construct/s measured by the scale and determining the scale items. Then, in order to verify the unfolded construct, CFA was performed on data 2. Finally, test-retest reliability analysis was conducted using data 3 to explore the stability of the confirmed scale through test-retest technique.

Outliers and normal distribution, the required assumptions for performing EFA, were checked out in data 1. The investigation into outliers showed that all the participants' standard *Z* scores were between the range -2.65 and 2.14, and thus, between the range ±3,29, indicating the absence of outliers (Tabachnick & Fidell, 2007). In addition, the analysis of box graph demonstrated there were not outliers. Multivariate outliers were scanned with Mahalanobis distance, and no outlier was identified. Examination on normal distribution revealed that skewness and kurtosis values were -.10 and -.19, respectively, showing that they were between the range ±1. Furthermore, the division results of the skewness and kurtosis values to their standard errors were -.52 and -.53, respectively. These values did not surpass the critical range ±1.96 values. The analysis of the histogram graphic at this stage unraveled that it had normal distribution values. Normality assumption was also measured by Kolmogrov-Smirnov analysis, resulting in reaching the decision that the data had normal distribution in view of the insignificance of the statistics ($p > 0.05$) (Çokluk et al., 2014). In order to check the multivariate normality assumption of the items, Mardia's test was administered through MVN package (Korkmaz et al., 2014) from the R software (R Core Team, 2021). The Mardia skewness and kurtosis were detected to be different from normal distribution ($p < .001$). Since the multivariate normality assumption was not provided, Principal Axis Factoring extraction technique was selected for extraction (Fabrigar et al., 1999).

EFA was carried out using the responses of 180 students given to 16 item candidates in the scale to find out their dimensions. The sample size indicates the existence of a sufficient data set considering the suggestion of Kline (1994) and Bryman and Cramer (2001) as to gathering data 10 times bigger than the number of items (as cited in Çokluk et al., 2014). Barlett and Kaiser-Meyer-Olkin (KMO), two of the tests of sphericity, were conducted to be able to perform EFA. Following the appropriateness of the results, factor and the loadings of the items on the factor were discovered using Principal Axis Factoring method of EFA. To determine the number of factors, the results obtained from the analyses of eigenvalue, scree plot and parallel analysis methods (Timmerman & Lorenza-Seva, 2011) were evaluated together. Since a single-factor structure was identified, no rotation was used. Besides, Cronbach's Alpha and McDonalds' coefficient Omega were calculated. The stated analyses of item candidates were done by the psych package (Revelle, 2020) on the software programs R (R Core Team, 2021), SPSS 25 and on Factor 10.5 (Lorenza-Seva & Ferrando, 2006).

Data 2 was exposed to CFA using Lisrel software *v*. 8.8 (Jöreskog & Sörbom, 2006) to confirm the scale, the factor of which was determined by EFA. It was expected to establish model-data fit at this stage. To that end, in addition to calculating Root Mean Square Error of Approximation (RMSEA) value, Comparative Fit Index (CFI), Standardized Root Mean Square Residual (SRMR) indices, recommended by Kline (2015), χ^2/sd, Normed Fit Index (NFI), Goodness of Fit Index (GFI) values were calculated. If the χ^2/sd value is smaller than 2, it indicates the perfect fit and if it is 2-3, it signifies good fit. Providing RMSEA and SRMR values are smaller than .05, it indicates perfect fit, and if it is between .05-.08, it means there is mediocre fit (Browne & Cudeck, 1993). If CFI, NFI and GFI fit indices are bigger than .95, it points to high fit, and if they are between .95-.90, it indicates acceptable fit (Çokluk et al., 2014; Hu & Bentler, 1999; Tabachnick & Fidell, 2007). On the condition that the fit indices are at desired levels, it will be shown that the structure that is tested is confirmed.

As well as the factor analyses, the difference in the level of self-efficacy of the first- and fourth-year students studying at the same department, which is assumed to be different from each other, was analyzed running independent samples t-test. The probable significant difference between the levels of reading self-efficacy of the first- and fourth-year students could be deemed to be evidence for construct validity.

To explore the stability of the confirmed scale, data 3 was collected from the same sample through test-retest. The analysis of Pearson correlation coefficient between the test and the retest will enable the investigation of the reliability of the scale results.

3. FINDINGS

In this section, results of the validity of the scale (content and construct validity), EFA and CFA, and reliability (internal consistency and stability) analyses have been presented.

3.1. Findings Regarding the Validity Analyses

In this section, results of the validity of the scale (content and construct validity), EFA and CFA, and reliability (internal consistency and stability) analyses will be presented.

3.1.1. *Content validity*

As was mentioned in the methodology section, the generated 16 items were e-mailed to three pre-service English language teacher educators teaching the courses of Reading Skills and Critical Reading and Writing to make sure the items catered for unearthing college students' level of reading self-efficacy. The three teacher educators contended that all the items could contribute to the measurement of reading self-efficacy of college students, and therefore, there was no need to exclude any items from the scale.

3.1.2. *EFA*

After testing the assumptions necessary for performing EFA and seeing that they were met, Barlett and Kaiser-Meyer-Olkin (KMO) analyses were carried out so as to check out the appropriateness of the data for factor analysis, the findings of which are given in Table 1 below.

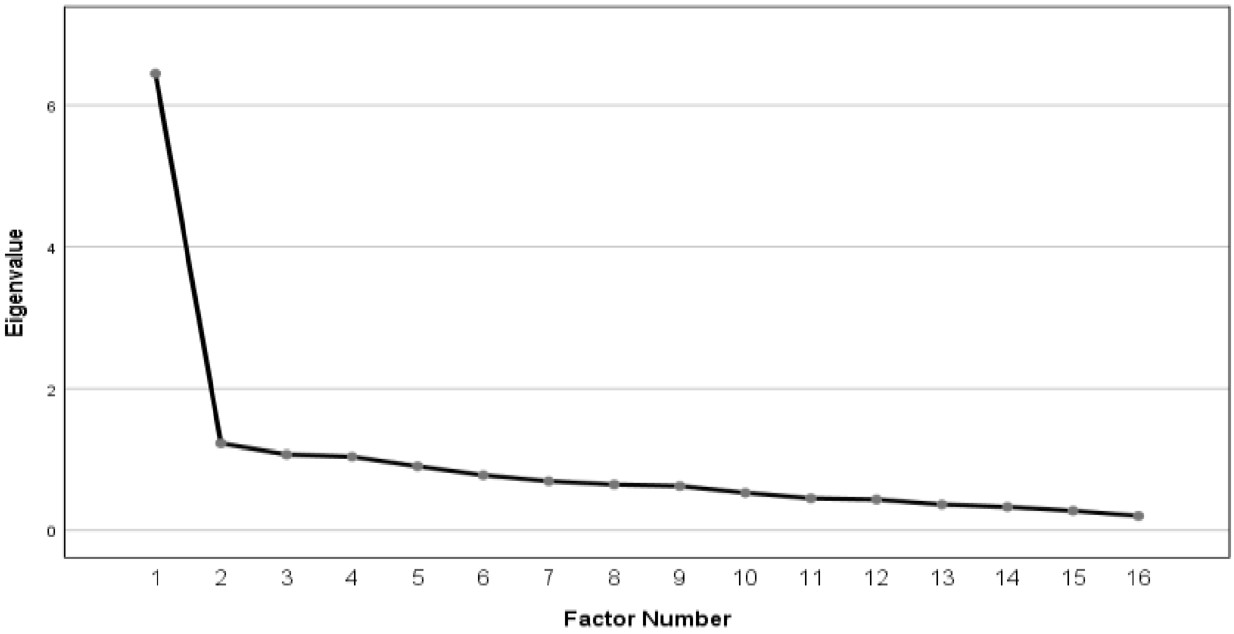
Table 1. *Results of barlett and kaiser-meyer-olkin (KMO) analyses.*

Statistic		Value
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.861
Bartlett's Test of Sphericity	χ^2	1170.25
	df	120
	p	.000

Table 1 displays that data 1 was appropriate for factor analysis in view of the fact that the attained value was between .8-9 (Çokluk et al., 2014). Likewise, because Bartlett's Test of Sphericity was significant, it indicated the data could be factorized (Tabachnick & Fidell, 2007); as a result, EFA was performed.

The analyses of eigenvalues, scree plot and parallel analysis methods (Timmerman & Lorenza-Seva, 2011) were done to determine the number of construct/s measured by the scale. The number of factors was ascertained in light of the findings in Figure 1 and Table 2.

Figure 1. *Scree plot.*



Taking into account Figure 1 demonstrating the eigenvalue of the factors determined by running EFA, it was induced that the scale had one dimension in that the slope took on almost a horizontal shape with the second factor.

Table 2. *Results of EFA with respect to the number of factors.*

Factor	Eigenvalue	Exploratory variance	Parallel Analysis Real Data Variance	Parallel Analysis Random Variance
1	6.45	40.31	45.7	12.8
2	1.22	7.65	8.2	11.6
3	1.07	6.68	7.2	10.6
4	1.04	6.49	6.2	9.7

As depicted in Table 2, 16-item scale had four factors whose eigenvalues surpass 1. Considering the probability that more factors than the ones existing in reality could be extracted on the condition that eigenvalue is more than 1, the eigenvalues of consecutive potential factors were compared. Whereas the eigenvalue of the first factor was 5.29 times more than that of the second factor, this value decreased to 1.14 in the next consecutive factor, which also showed that the scale was a unidimensional-scale. The findings obtained from conducting parallel analysis developed by Timmerman and Lorenza-Seva (2011) revealed that the real data variance was bigger than the random variance solely in the first factor, and in the other variances, it was smaller. This finding also suggested the scale was one-dimensional. When the findings obtained from all the methods were evaluated together, it was concluded that the scale had one dimension. It was also revealed that the factor explained 40% of the total variance.

Following taking the decision that the scale had one dimension, EFA was re-conducted for the one-dimensional situation and the factor loadings of the items on the dimension were calculated, which are given in Table 3 below.

Table 3. *Factor loadings of the items on the dimension.*

Item	Factor Loading
14	.710
13	.685
02	.662
16	.661
01	.660
11	.652
09	.618
04	.598
06	.593
5	.557
12	.555
7	.543
15	.538
08	.537
03	.523
10	.514

Table 3 shows that the loadings of the items on the factor varied between .51-.71. These values were above the critical loading values, .32 (Tabachnick & Fidell, 2007) and .4 (Kim-Yin, 2004, as cited in Çokluk et al., 2014), mentioned in the literature. Therefore, loadings of the factors were ensured over .5 as outlined by Hair et al. (2019). For this reason, it is decided that none of the items would be excluded from the scale.

3.1.3. *CFA*

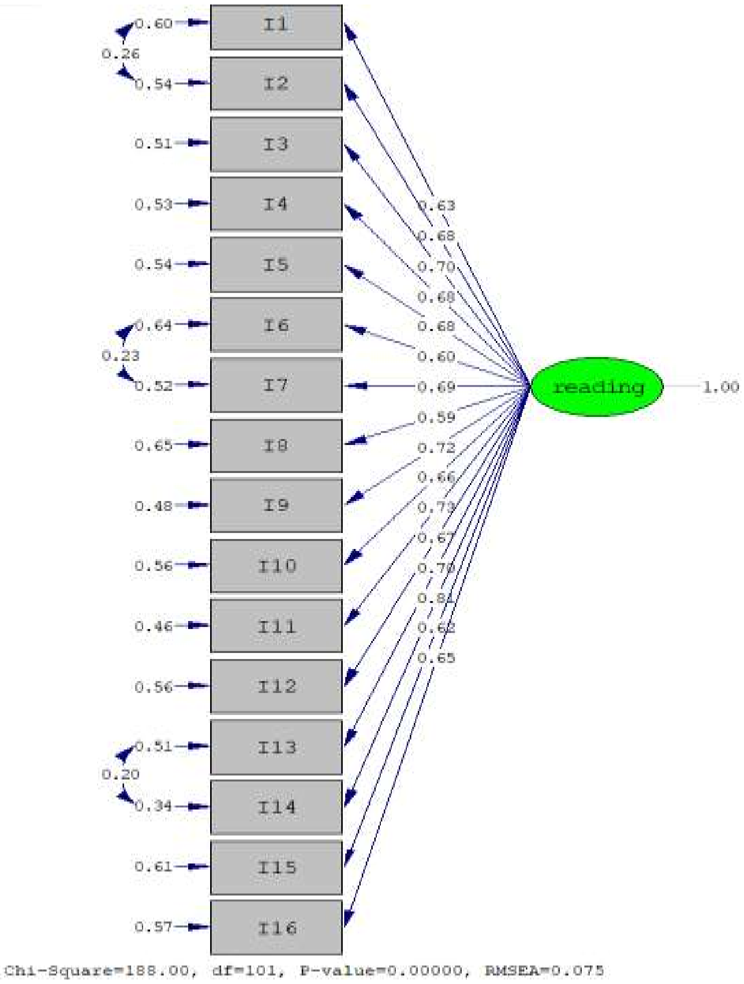
CFA was performed in an attempt to confirm the construct of the scale, figured out by virtue of EFA. The findings with respect to the CFA performed on data 2, different from data 1, are presented in Figure 2 and Table 4 below.

Table 4. *Goodness of fit indices of the scale.*

Statistics	Before Modification	After Modification	Related Range	Meaning
χ^2/sd	2.57	1.86	<2	Perfect fit
RMSEA	.1	.0755	.05<RMSEA<.08	Acceptable
SRMR	.059	.051	.05<SRMR<.08	Good fit
CFI	.96	.98	CFI>.95	Perfect fit
NFI	.93	.95	NFI≥.95	Perfect fit
GFI	.82	.87	GFI<.9	Weak fit

As depicted in Table 4, out of a total of six fit indices, perfect fit was observed in three indices, with 1.86, which is lower than 2 in χ^2/sd statistics, with 98 in CFI and .95 showing an equal value of the .95 cut-off in NFI statistics. Acceptable fit was observed within the range of .05 and .08 in RMSEA statistics with an index value of .075. Good fit was observed with .051 falling into the range between .05 and .08 in SRMR statistics while the weak fit was only seen in GFI statistics with a value of .87, which is lower than .9. Since the GFI value was close to good fit value and the expected values were provided by the other indices, it can be concluded that the model-data fit was ensured in data 2, and the construct determined by the EFA was verified by the CFA.

Figure 2. *Factor loadings of the items revealed by CFA results.*



As demonstrated in Figure 2, the factor loadings of all the items were above .4. Seeing that factor loading *t* values of all the items were *p* = 0.01, surpassing the critical value 2.56, it was confirmed that their existence in the scale was significant. Analyzing the modification suggestions produced by the software program, error covariance was determined for the items 1 and 2, items 6 and 7, and items 13 and 14, involving similar expressions. The other modifications were not carried out as they were not grounded upon theoretical basis. Model-data fit indices of the scale tested by CFA are given in Table 4 below.

The analysis of the construct validity was completed analyzing the total scores and correlation of the items, displayed in Table 5 below.

Table 5. *Total scores and correlation of the items.*

Item	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
1	.622	.891
2	.625	.891
3	.494	.895
4	.559	.893
5	.523	.894
6	.558	.893
7	.514	.895
8	.510	.895
9	.590	.892
10	.489	.896
11	.619	.891
12	.530	.894
13	.645	.890
14	.669	.889
15	.502	.896
16	.630	.891

Table 5 illustrates that the total scores of the items with their correlation coefficients changed between .489-.669. Because all the values were bigger than .3 (Field, 2009), it is deduced that all the items were closely related to the construct.

The findings obtained from the analysis of data 3 conducted to compare the level of self-efficacy of the first- and fourth-year students are presented in Table 6.

Table 6. *The independent t-test results of the difference in the level of the first- and fourth-year students' reading self-efficacy.*

Group	*N*	Mean	*SD*	*t*	*p*
First-year	30	58.03	10.36	-4.436	.000
Fourth-year	35	68.60	8.84		

Table 6 shows that the level of fourth-year participants' reading self-efficacy was statistically higher than that of fourth-year participants' reading self-efficacy. Taking into consideration the point that it is something expected, it could be concluded that the scale accurately measures the construct.

3.2. Findings Regarding the Reliability

In view of the Cronbach's Alpha internal consistency coefficient obtained from data 1 was .90. This finding exhibits this scale has a high level of internal consistency (George & Mallary, 2003). McDonalds' Omega coefficient, another reliability measurement method, performed on data 1 was .90. This finding shows that the scale as a whole has a high level of reliability. Furthermore, when Table 5 is interpreted in regard to the effect of the items on reliability, it could be seen that the exclusion of each item reduced the reliability. This might be viewed as the positive influence of the items on reliability. Lastly, the stability coefficient of the scale was revealed to be .76 through employing the test-retest method in data 3 over a time interval of two weeks, indicating a high level of relationship. In this regard, the scale has an adequate level of stability.

4. DISCUSSION and CONCLUSION

The findings obtained from EFA indicated the appropriateness of the data for carrying out factor analysis, and that the reading self-efficacy scale had only one dimension. Another finding yielded from EFA is linked with the factor loadings of the items on the dimension, suggesting exclusion of any of the items from the scale was unnecessary. CFA results showed that all the items in the scale were significant and closely related to the construct and confirmed the construct figured out by EFA.

Collecting data from the students of the departments of English language teaching and English language and literature at different years of study afforded an opportunity to the researchers to investigate whether or not there is a statistically significant difference between the levels of reading self-efficacy of the first- and fourth-year students, which also functioned as a medium for checking out if the scale served for measuring the construct. The findings obtained from independent-samples t-test analysis revealed the existence of a statistically significant difference between the levels of first- and fourth-year students' reading self-efficacy, providing strong evidence for the result that the scale measured the construct. The findings on Cronbach's Alpha internal consistency coefficient and McDonalds' Omega coefficient values showed that the developed scale had a high level of internal consistency and reliability. The finding on the stability coefficient value of the scale demonstrated that it has an adequate level of stability.

This study set out to fill the gap in the literature in developing and validating a reading self-efficacy scale through collecting data from the students at different years of study in the departments of English language teaching and English language and literature. The developed and validated five-point Likert reading self-efficacy scale differs from the ones existing in the literature considering the number of the collected data (Ahmadian & Pasand, 2017; Boakye, 2015; Burrows, 2012; Kakaew, & Damnet, 2017; McLean, & Poulshock, 2018; Mullins, 2018; Zare & Davoudi Mabarakeh, 2011), the comprehensiveness of the items directly related to reading self-efficacy as opposed to other scales (Boakye, 2015; Kakaew & Damnet, 2017). Additionally, as opposed to the previously developed scales (Boakye, 2015; Kakaew & Damnet, 2017), the items in this scale have been produced in accord with Bandura's (1997, 2006) principles of generating items in a reading self-efficacy scale. Thereupon, it could be argued that this study is highly likely to enable appropriately measuring reading self-efficacy of college students. Nevertheless, it should be noted that it could be adapted for measuring the level of L2 reading self-efficacy of students of different ages and at different levels of education although the reading self-efficacy scale was developed and validated through gathering data from college students. Moreover, using the reading self-efficacy scale could help identify students' strengths and weaknesses in their reading skills, which could enable reading teachers to become conscious of the areas to be developed in students' reading abilities. Considering the importance of being aware of what should be improved, the data to be provided by the scale can lead the teacher to properly scaffold their students to enhance their reading comprehension, and in turn, to increase the level of their reading self-efficacy. In view of the comprehensiveness of the items in the scale, it could be suggested it be adapted to measure younger students' reading self-efficacy.

Acknowledgments

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

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research and publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s). **Ethics Committee Number**: Hatay Mustafa Kemal University Social and Human Sciences Research and Publication Ethics Board (Document numbered 902-01-FR 006 and dated October 1, 2020).

Authorship Contribution Statement

Gulten Kosar: Structuring the research, collecting data and writing the original draft. **Yunus Emre Akbana**: Finding resources, collecting data and revising the draft. **Levent Yakar:** Analyzing the data.

Orcid

Gulten Kosar https://orcid.org/0000-0002-4687-4382
Yunus Emre Akbana https://orcid.org/0000-0002-5707-3564
Levent Yakar https://orcid.org/0000-0001-7856-6926

REFERENCES

Afflerbach, P., & Cho, B.Y. (2010). Determining and describing reading strategies: Internet and traditional forms of reading. In H.S. Waters & W. Schneider (Eds.), *Metacognition, strategy use, and instruction* (pp. 201–225). Guilford.

Ahmadian, M., & Pasand, P.G. (2017). Efl learners' use of online metacognitive reading strategies and its relation to their self-efficacy in reading. *Reading Matrix: An International Online Journal, 17*(2), 117–132. http://mail.readingmatrix.com/files/17-097to04m.pdf

Bağcı, H. (2019). An investigation of Turkish language and Turkish language and literature teacher candidates' critical reading self-efficacy (The case of Mehmet Akif Ersoy University). *Advances in Language and Literary Studies, 10*(4), 14–20. http://dx.doi.org/10.7575/aiac.alls.v.10n.4p.14

Barber, A.T., Buehl, M.M., Kidd, J.K., Sturtevant, E.G., Nuland, L.R., & Beck, J. (2015). Reading engagement in social studies: Exploring the role of a social studies literacy intervention on reading comprehension, reading self-efficacy, and engagement in middle school students with different language backgrounds. *Reading Psychology, 36*(1), 31–85. https://doi.org/10.1080/02702711.2013.815140

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

Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares, & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 307–337). Information Age Publishing.

Barkley, J.M. (2006). Reading education: is self-efficacy important? *Reading Improvement, 43*(4), 194–211. https://eric.ed.gov/?id=EJ765527

Blachowicz, C., & Ogle, D. (2008). *Reading comprehension: Strategies for independent learners*. The Guilford Press.

Boakye, N.A.N.Y. (2015). The relationship between self-efficacy and reading proficiency of first-year students: An exploratory study. *Reading & Writing: Journal of the Reading Association of South Africa, 6*(1), 1–9. https://doi.org/10.4102/rw.v6i1.52

Browne. M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Sage.

Burrows, L. (2012). *The effects of extensive reading and reading strategies on reading self-efficacy* [Unpublished Doctoral dissertation, Temple University]. http://dx.doi.org/10.34944/dspace/868

Butterworth, J., & Thwaites, G. (2013). *Thinking skills critical thinking and problem solving*. Cambridge University Press.

Caprara, G.V., Fida, R., Vecchione, M., Del Bove, G., Vecchio, G.M., Barbaranelli, C., Bandura, A. (2008). Longitudinal analysis of the role of perceived self-efficacy for self-regulated learning in academic continuance and achievement. *Journal of Educational Psychology, 100*(3), 525–534. https://doi.org/10.35569/biormatika.v6i02.773

Carroll, J.M., & Fox, A.C. (2017). Reading self-efficacy predicts word reading but not comprehension in both girls and boys. *Frontiers in Psychology*, *7*, 2056. 1–9. https://doi.org/10.3389/fpsyg.2016.02056

Carter, C.E. (2011). *Mindscapes: Critical reading skills and strategies.* Cengage Learning.

Cho, E., Roberts, G.J., Capin, P., & Roberts, G. (2015). Cognitive attributes, attention, and self-efficacy of adequate and inadequate responders in a fourth grade reading intervention. *Learning Disabilities Research & Practice, 30*(4), 159-170. https://doi.org/10.1111/ldrp.12088

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2014). *Sosyal bilimler için çok değişkenli istatistik SPSS ve LISREL uygulamaları [Multivariate statistics for social sciences: SPSS and LISREL applications]* (3rd ed.). Pegem Akademi.

Dallon, B., & Ratner, W. (2002). *Reading between the lines: Improve your scores on English & social studies tests*. Learning Express, LLC.

Duke, N.K., & Pearson, P.D. (2002). Effective practices for developing reading comprehension. In A. E. Farstrup & S.J. Samuels (Eds.), *What research has to say about reading instruction* (3rd ed., pp. 205–242). International Reading Association.

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. https://doi.org/10.1037/1082-989X.4.3.272

Ferrara, S.L.N. (2007). Reading fluency and self-efficacy: A case study. *International Journal of Disability, Development and Education, 52*(3), 215-231. https://doi.org/10.1080/10349120500252858

Forzani, E., Leu, D.J., Yujia Li, E., Rhoads, C., Guthrie, J.T., & McCoach, B. (2021). Characteristics and validity of an instrument for assessing motivations for online reading to learn. *Reading Research Quarterly, 56(4)*, 1-20, https://doi.org/10.1002/rrq.337

Fung, D., & Macaro, E. (2019). Exploring the relationship between linguistic knowledge and strategy use in listening comprehension. *Language Teaching Research*. 1–25. https://doi.org/10.1177/1362168819868879

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference.* (11.0 update) (4th ed.). Allyn & Bacon.

Ghonsooly, B., & Elahi, M. (2010). Learners' self-efficacy in reading and its relation to foreign language reading anxiety and reading achievement. *Journal of English Language Teaching and Learning, 53*(217), 45-67. https://www.sid.ir/FileServer/JE/1323201021703.pdf

Grabe, W., & Stoller, F. (2002). *Teaching and researching reading*. Pearson Education.

Grabe, W., & Stoller, F. L. (2019). *Teaching and researching reading*. Routledge.

Guthrie, J.T., & Taboada, A. (2005). Fostering the cognitive strategies of reading comprehension. In J.T. Guthrie, A. Wigfield, & K.C. Perencevich (Eds.), *Motivating reading comprehension: Concept-oriented reading instruction* (pp. 87–113). Lawrence Erlbaum Associates, Inc.

Guthrie, J.T., Wigfield, A., & VonSecker, C. (2000). Effects of integrated instruction on motivation and strategy use in reading. *Journal of Educational Psychology, 92*(2), 331–341. https://doi.org/10.1037/0022-0663.92.2.331

Graham, S., Woore, R., Porter, A., Courtney, L., & Savory, C. (2020). Navigating the challenges of L2 reading: Self-efficacy, self-regulatory reading strategies, and learner profiles. *The Modern Language Journal*, *104*(4), 693-714. https://doi.org/10.1111/modl.12670

Gu, Y. (2019). Approaches to learning strategy instruction. In A. U. Chamot & V. Harris (Eds.), *Learning strategy instruction in the language classroom: Issues and implementation* (pp. 22–37). Multilingual Matters.

Hair, J.F., Anderson, R.E., Tatham, R.L., & Black, W.C. (2019). *Multivariate data analysis* (8th ed.). Prentice Hall, Inc.

Hedges, J.L., & Gable, R. (2016). The relationship of reading motivation and self-efficacy to reading achievement. *K-12 Education*, *31*(1). https://scholarsarchive.jwu.edu/k12_ed/31

Hu, L.T., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: A Multidisciplinary Journal, 6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Jöreskog, K.G., & Sörbom, D. (2006). *LISREL 8.8 for windows* [Computer software]. Scientific Software International.

Kakaew, J., & Damnet, A. (2017). Learning strategies model to enhance Thai undergraduate students' self-efficacy beliefs in EIL textual reading performance. *Advances in Language and Literary Studies, 8*(6), 19–27. http://dx.doi.org/10.7575/aiac.alls.v.8n.6p.19

Karabay, A. (2013). Eleştirel okuma özyeterlik algı ölçeğinin geliştirilmesi [The development of critical reading self-efficacy perception scale]. *Turkish Studies – International Periodical For The Languages, Literature, and History of Turkish and Turkic*, *8*(13), 1107–1122. http://dx.doi.org/10.7827/TurkishStudies.5389

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge.

Kline, R.B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Kula, S.S., & Budak, Y. (2020). Self-efficacy perceptions scale for reading comprehension of 4th grade students in primary school: Validity and reliability study. *Bartın University Journal of Faculty of Education*, *9*(1), 106-120. http://dx.doi.org/10.14686/buefad.536885

Li, Y., & Wang, C. (2010). An empirical study of reading self-efficacy and the use of reading strategies in the Chinese EFL context. *Asian EFL Journal*, *12*(2), 144–162. http://70.40.196.162/PDF/June-2010.pdf#page=144

Lorenzo-Seva, U., & Ferrando, P.J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavioral Research Methods, Instruments and Computers, 38*(1), 88-91. https://doi.org/10.3758/BF03192753

Macaro, E. (2019). Language learner strategies and individual differences. In A. U. Chamot & V. Harris (Eds.), *Learning strategy instruction in the language classroom: Issues and implementation* (pp. 68-80). Multilingual Matters. https://doi.org/10.21832/9781788923415-011

McGuinness, D. (2005). *Language development and learning to read*. The MIT Press.

McLean, S., & Poulshock, J. (2018). Increasing reading self-efficacy and reading amount in efl learners with word-targets. *Reading in a Foreign Language*, *30*(1), 76–91. https://files.eric.ed.gov/fulltext/EJ1176293.pdf

Meniado, J.C. (2016). Metacognitive reading strategies, motivation, and reading comprehension performance of Saudi EFL students. *English Language Teaching*, *9*(3), 117–129. http://dx.doi.org/10.5539/elt.v9n3p117

Mills, N., Pajares, F., & Herron, C. (2006). A reevaluation of the role of anxiety: Self-efficacy, anxiety, and their relation to reading and listening proficiency. *Foreign Language Annals*, *39*(2), 276–295. https://doi.org/10.1111/j.1944-9720.2006.tb02266.x

Mullins, L.A. (2018). *Personalized texts and second language reading: A study in self-efficacy.* [Unpublished Doctoral dissertation, Georgia State University]. https://scholarworks.gsu.edu/mse_diss/70/

Mullins, L.A. (2019). Evaluating target language reading self-efficacy scales: Applying principles gleaned from Bandura's writings. *Reading Matrix: An International Online Journal*, *19*(2), 1–12. http://www.readingmatrix.com/files/21-1jkbwrqn.pdf

Murad Sani, A., & Zain, Z. (2011). Relating adolescents' second language reading attitudes, self efficacy for reading, and reading ability in a non-supportive ESL setting. *The Reading Matrix*, *11*(3), 243-254. http://www.readingmatrix.com/articles/september_2011/sani_zain.pdf

Palincsar, A.S., & Brown, A.L. (1986). Interactive teaching to promote independent learning from text. *The Reading Teacher, 39*(8), 771–777. http://www.jstor.org/stable/20199221

Peura, P.I., Viholainen, H.J., Aro, T.I., Räikkönen, E.M., Usher, E.L., Sorvo, R.M., ... & Aro, M.T. (2019). Specificity of reading self-efficacy among primary school children. *The Journal of Experimental Education*, *87*(3), 496-516. https://doi.org/10.1080/00220973.2018.1527279

R Core Team. (2021). *R: A language and environment for statistical computing* [Computer software manual]. R foundation for statistical computing. https://www.R-project.org/

Revelle, W. (2020). *Psych: procedures for personality and psychological research* (Version 2.1.9) [Computer software] https://cran.r-project.org/web/packages/psych/index.html

Ronimus, M., Eklund, K., Westerholm, J., Ketonen, R., & Lyytinen, H. (2020). A mobile game as a support tool for children with severe difficulties in reading and spelling. *Journal of Computer Assisted Learning*, *36*(6), 1011–1025. https://doi.org/10.1111/jcal.12456

Schunk, D.H. (2003). Self efficacy for reading and writing: Influence of modeling, goal-setting and self-evaluation. *Reading and Writing Quarterly, 19*(2), 159-172. https://doi.org/10.1080/10573560308219

Smith, F. (2012). *Understanding reading: A psycholinguistic analysis of reading and learning to read*. Routledge.

Soland, J., & Sandilos, L.E. (2020). English language learners, self-efficacy, and the achievement gap: understanding the relationship between academic and social-emotional growth. *Journal of Education for Students Placed at Risk (JESPAR)*, *26(*1), 1–25. https://doi.org/10.1080/10824669.2020.1787171

Solheim, O.J. (2011). The impact of reading self-efficacy and task value on reading comprehension scores in different item formats. *Reading Psychology, 32*(1), 1–27. https://doi.org/10.1080/02702710903256601

Şahin, E., & Öztahtalı, İ. (2019).Etkili okuma özyeterlik algı ölçeğinin geliştirilmesi: geçerlilik ve güvenirlik çalişması [The development of effective reading self efficacy perception scale: study on the validity and reliability]. *Electronic Turkish Studies*, *14*(4). http://dx.doi.org/10.29228/TurkishStudies.23378

Tabachnick, B.G., & Fidel, L.S. (2007). *Using multivariate statistics* (5th ed.). Pearson.

Tavakoli, H., & Koosha, M. (2016). The effect of explicit metacognitive strategy instruction on reading comprehension and self-efficacy beliefs: The case of Iranian university EFL students. *Porta Linguarum*, 25, 119-133. https://dialnet.unirioja.es/descarga/articulo/5412447.pdf

Taylor, A.M. (2014). L1 glossing and strategy training for improving L2 reading comprehension: A meta-analysis. *International Journal of Quantitative Research in Education*, *2*(1), 39–68. https://doi.org/10.1504/IJQRE.2014.060973

Timmerman, M.E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods, 16*(2), 209–220. https://doi.org/10.1037/a0023353

Tremblay, P.F., & Gardner, R.C. (1995). Expanding the motivation construct in language learning. *The Modern Language Journal, 79*(4), 505-518. https://doi.org/10.1111/j.1540-4781.1995.tb05451.x

Unrau, N.J., Rueda, R., Son, E., Polanin, J.R., Lundeen, R.J., & Muraszewski, A.K. (2018). Can reading self-efficacy be modified? A meta-analysis of the impact of interventions on reading self-efficacy. *Review of Educational Research*, *88*(2), 167-204. https://doi.org/10.3102/0034654317743199

Venegas, E.M. (2018). Strengthening the reader self-efficacies of reluctant and struggling readers through literature circles. *Reading & Writing Quarterly, 34*(5), 419–435. https://doi.org/10.1080/10573569.2018.1483788

Zare, M., & Davoudi Mobarakeh, S. (2011). The relationship between self-efficacy and use of reading strategies: The case of Iranian senior high school students. *Studies in Literature and Language, 3*(3), 98-105. http://dx.doi.org/10.3968/n

Zimmerman, B.J. (2013). From cognitive modeling to self-regulation: A social cognitive career path. *Educational Psychologist*, *48*(3), 135-147. https://doi.org/10.1080/00461520.2013.794676

APPENDIX

Reading Self-Efficacy Questionnaire

Reading self-efficacy questionnaire has been designed in order to measure your judgement of your reading self-efficacy. Please read the items in the questionnaire carefully and make an accurate evaluation of your reading self-efficacy by choosing the number that accurately represents your ability in each item. Remember to provide information about your age, department and gender.

Gender : ❑ Female ❑ Male

Age : ________

Department : ________

Year of study : ________

1 **Not at all**	**2** **Slightly**	**3** **Somewhat**	**4** **Fairly well**	**5** **Very well**

Item	**1**	**2**	**3**	**4**	**5**
1- I can identify the topic of a reading passage.					
2-I can identify the purpose of the author.					
3-I can use my background knowledge about the topic of the reading passage to improve my reading comprehension.					
4-I can find the explicit main idea of a reading passage.					
5-I can find the implied main idea of a reading passage.					
6-I can determine topic sentences in a reading passage.					
7-I can find supporting detail/s in a paragraph.					
8-I can use context clues to guess the meanings of unknown words in a passage.					
9-I can judge whether supporting details are relevant to the topic of the reading passage.					
10-I can distinguish facts from opinions in a reading passage.					
11-I can answer questions on the passage after reading it.					
12-I can use reading strategies like skimming and scanning to enhance my reading comprehension.					
13-I can draw logical conclusions from a reading passage.					
14-I can make logical inferences based on what is given in the reading.					
15- I can take notes of key points as reading a passage.					
16-I can summarize a reading passage after reading it.					

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 1, 220–232

https://doi.org/10.21449/ijate.960565

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

Scaling the criteria to be considered in determining exam anxiety by pairwise comparison method

Hasibe Yahsi Sari[1,*], Duygu Anil[2]

[1]Hacettepe University, Faculty of Education, Department of Educational Sciences, Ankara, Turkiye.

ARTICLE HISTORY

Received: July 01, 2021
Revised: Dec. 12, 2021
Accepted: Jan. 31, 2022

Keywords:
Exam anxiety,
Pairwise comparison,
Thurstone,
Equation of case III,
Scaling.

Abstract: In this study, it was aimed to scale the importance level of the criteria that can be taken into consideration in determining the exam anxiety of 8th grade students by means of pairwise comparison. Descriptive survey model was used in the research. The study group of the research consists of 100 8th grade students studying at a randomly selected secondary school in Kilis. The data collection tool was a questionnaire in which the students in the study group were compared in pairs, that consisted of six criteria that affect exam anxiety, including the thought of failing in lessons, the effect of the social environment (family, friends and relatives), students's self-perceptions, teacher attitudes, social and physical characteristics of the school, and the thought of not being prepared for exams adequately. Data analysis was performed on a full data matrix by using equation of case III. As a result of the research, the first two criteria that most affect the exam anxiety of 8th grade students are respectively; social and physical characteristics of the school and how students see themselves. These stimulators are respectively followed by the idea of failing in lessons and the effect of the social environment. The two criteria, which they think are the least effective on exam anxiety, are the attitudes of teachers and not being prepared for exams adequately.

1. INTRODUCTION

Anxiety is an important emotion that affects the processes that can be called milestones in people's lives. Anxiety comes to the fore especially in our academic life. Unfortunately, today, this feeling is seen frequently even at a very young age. The reason for this is that children take exams at an early age and both parents and students attribute more importance to these exams than necessary. Mandler and Sarason (1952) defined exam anxiety as a feeling of inadequacy and helplessness, abnormal somatic reactions, fear of punishment, loss of dignity, and the desire to leave the anxiety-provoking environment.

Exam anxiety is a set of multidimensional negative phenomena that may occur before or during the exam (Basol, 2017). There is a need for research to determine how and how much the academic success and quality of life of students are affected due to exam anxiety, and which factors affect exam anxiety more. Basol & Zabun (2014), in their study, found to what extent participation in private exam preparation courses, multidimensional perfectionism, parental

[*]CONTACT: Hasibe Yahşi Sarı ✉ hsbyahsi@gmail.com 🏢 Hacettepe University, Faculty of Education, Department of Educational Sciences, Ankara, Turkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

attitudes and exam anxiety explained 8th grade students' SBS success levels. According to their research results, the most important predictor in determining the success level of a student is attending private preparation courses, while other predictors are a perfectionist attitude by the parents and the level of exam anxiety of the students.

Students push their cognitive capacities in order to enter qualified high schools and maximize their talents and skills. This situation creates exam anxiety for students. Hembree (1988) analyzed 562 studies in his meta-analysis on exam anxiety. As a result of the research, he stated that exam anxiety is directly related to the fear of negative evaluation, alertness and other types of anxiety.

Exam anxiety is a frequently seen feeling especially in 8th and 12th grades (Bulut, 2010; Duman, 2008; Kayapinar, 2006; Kaya & Savrun, 2015; Kesici & Asilioglu, 2017). In his study, Duman (2008) found that there was a significant relationship between state trait anxiety and exam anxiety levels and parental attitudes of 8th grade students. Güler and Cakir (2013) analyzed the role of gender, irrational beliefs and parental attitudes on 12th grade students' exam anxiety levels. According to the research results, it was seen that only the strict control attitude perceived from the mother significantly predicted the exam anxiety total delusion and effectiveness scores. Social environment is defined in the literature as; family, friends, school and place of residence (Coban, 2009), society, peer groups, school, professional environment (Yelken, 2011). Therefore, in this study, family, friends and relatives are considered as the effect of the social environment. Zeidner (1998) revealed that negative feelings about exam anxiety could manifest as an individual having negative thoughts about himself and feeling panic. Cakmak, Sahin, and Akinci-Demirbas (2017) analyzed the relationship between exam anxiety and self-esteem of 7th and 8th grade students and searched whether some variables made a difference in exam anxiety and self-esteem scores. It was found that there was a negative significant relationship between exam anxiety and self-esteem.

The sub-dimensions in the scales also reveal the factors that affect the exam anxiety of individuals. For example, the sub-dimensions of the Exam Anxiety Scale for Children adapted to our language by Aydin and Bulgan (2017) are thoughts, off-task behaviors and autonomous reactions. Various measurement tools were used to measure the level of exam anxiety in studies. Some of these scales are: Westside Exam Anxiety Scale (Driscoll, 2007), Spielberger's Exam Anxiety Scale (Spielberger, 1980), Children's Exam Anxiety Scale (Wren & Benson, 2004). These measurement tools, which were adapted into Turkish, were used in many studies (Bacanlı &Driver, 2006; Basol & Zabun, 2014; Delioglu, 2017; Duman, 2008; Kavakci, Güler & Cetinkaya, 2011; Totan & Yavuz, 2009; Tugan, 2015). While studies about exam anxiety scale development, adaptation and application are frequently encountered, the lack of scaling studies necessitated this study. Based on the studies in the literature, six criteria affecting exam anxiety were determined: the thought of failing in the lessons, the effect of the social environment (family, friends, relatives), students's self-perceptions, teacher attitudes, the social and physical characteristics of the school, and the thought of not being prepared for exams adequately.

1.1. Scaling

Scaling is needed to measure all emotions, situations and phenomena. In order to measure affective features, we need a measurement tool. We use scales to determine the importance of the qualities we will use in a measurement tool, to develop a measurement tool that serves the purpose of measurement and whose measurement results do not change over time. We develop these scales using various scaling techniques. According to Turgut and Baykul (1992), "The mathematical properties of the measurements obtained after the measurement based on the formal structure of the measurement attributes are called *scales*". Scaling aims to obtain more sensitive and better measurable standard scales by using different statistical methods under certain assumptions with the data obtained from observer judgments or subject reactions, about psychological variables whose physical properties are unknown and whose physical dimensions cannot be defined, and which do not have standard measurement tools (Anil & Inal, 2017). With scaling, measurement and comparison are made between real life and the world of perception. In scaling, first of all, the scale is developed by assigning numbers to psychological objects, and then individuals are placed on the developed scale (Anil & Inal, 2017). There are many scaling techniques. Some of these are: scaling with pairwise comparisons, scaling with ranking judgments, scaling with classification judgments, scaling with absolute judgments, scaling with rank sums, and multidimensional scaling. In the study, the scaling technique with pairwise comparisons was used.

1.1.1. *Scaling with Pairwise Comparisons*

Scaling with pairwise comparisons, one of the scaling techniques, was developed by Thurstone. The scaling method with pairwise comparisons is used when stimuli can be given to the respondents in pairs. In this method, observers make a response at the end of the process of distinguishing the stimulus. When an observer observes a $U_j$stimulus at different times and situations, he may perceive it differently. The reason for this is the instantaneous changes of the observer. These changes cause errors in discrimination judgments. This error is expressed by this equation:

$$s_j' = s_{ij} + e_{ij} \qquad \text{Equation 1}$$

(i=1, 2, …,N)

(j,k=1, 2,…,K)

s_j': The value of the Uj stimulus in the psychological dimension.

s_{ij}: The perceived value of the Uj stimulus by the observer i.

e_{ij}: Error of observer i in discrimination judgment of sij

It is assumed that the e_{ij} error here is random and $\sum e_{ij} = 0$ (Turgut and Baykul, 1992).

1.2. Comparative Judgement Law

Thurstone (1927) showed the difference between the scale values of two stimuli, j and k, according to comparative judgement law in the psychological dimension, with equation 2:

$$S_j - S_k = z_{jk}\sqrt{\sigma_j^2 + \sigma_k^2 - 2.r_{jk}.\sigma_j.\sigma_k} \qquad \text{Equation 2}$$

Five equations of case are used for the application of comparative law. The equation of case I has no solution. Since the number of unknowns is more than the number of equations, this case is known as the ideal case and is unsolvable. In the equation of case III, scaling is possible if N discrimination judgments are given for each of the K stimuli. This is the ideal case and has no solution. In the equation of case III, the correlations in equation 2 are regarded to be zero and the following equation is obtained:

$$S_j - S_k = z_{jk}\sqrt{\sigma_j^2 + \sigma_k^2} \qquad \text{Equation 3}$$

If the standard deviations of the discrimination judgments of the observers in the equation of case IV are very close to each other, the following equation is obtained.

$$S_j - S_k = z_{jk}.0{,}707.(\sigma_j^2 + \sigma_k^2 \qquad \text{Equation 4}$$

Equation of case V is the most used one. In this case, the correlation is assumed to be zero. It is also assumed that the stimuli discrimination distributions of all observers will be equal to each other and at a constant value in a scaling trial.

$$S_j - S_k = z_{jk} \qquad \text{Equation 5}$$

1.3. Purpose of the Study

In this study, it was aimed to scale the criteria that can be taken into consideration in determining the exam anxiety of 8th grade students with the pairwise comparison method. In the study, the question "How is the ranking of the criteria that can be taken into consideration in determining the exam anxiety of 8th grade students according to the scale values?" was tried to be answered.

As it is known, 8th grade students take the high school entrance exam every year. This exam, which has the importance to affect the future of students, creates anxiety (Bacanli & Drive, 2016; Duman, 2008). With this research, it is thought that revealing the psychometric importance levels of the factors that increase students' exam anxiety is important both in terms of literature and educational practices. Thus, students are given the necessary preventive guidance and psychological counseling in terms of exam anxiety, and exam anxiety is brought to the optimum level.

There are limited studies in the literature to determine the factors affecting test anxiety (Genc, 2016). Although there are studies using scaling methods in the literature (Anil, Taymur & Oztemur, 2017), there isn’t any study on test anxiety. It is a unique research in terms of the method used and in terms of the absence of scaling studies with the pairwise comparison method in determining the criteria affecting test anxiety.

2. METHOD

In this section, the research model, study group, data collection tools and data analysis are explained.

2.1. Research Model

In the study, it was aimed to scale the criteria affecting the exam anxiety of 8th grade students by using pairwise comparisons. The model of the research is descriptive survey. This model is a quantitative research technique that aims to determine certain characteristics of the universe or sample (Büyüköztürk et al., 2008).

2.2. Study Group

The study group of the research consists of 100 8th grade students from a randomly selected secondary school in Kilis.

2.3. Data Collection Tools

While preparing the data collection tool, the literature was searched and the studies were analyzed. According to studies in the literature, 10 factors affecting students' test anxiety were determined. In order to regulate the number of factors, a pilot application was made to the students. In addition to this pilot application, in order to examine the place of test anxiety in the literature and the psychometric properties of the measurement tool, expert opinion was obtained from a total of 6 academicians (a professor of Measurement and Evaluation in Education, two doctoral faculty members of Psychological Counseling and Guidance, and three research assistants, one of whom is Measurement and Evaluation in Education, and two of which are Psychological Counseling and Guidance). After the relevant eliminations, it was decided that these 10 factors would be too much for the students. 6 out of 10 items were selected. Finally, six factors (the thought of failing in the lessons, the effect of the social environment, students’s self-perceptions, teacher attitudes, the social and physical characteristics of the school, the thought of not being prepared for the exams adequately) that affect the exam anxiety of the 8th grade students were arranged so that a pairwise comparison could be made. All factors were explained during the data collection phase. For example, it has been stated that the influence of the social environment refers to family, friends and relatives. During the data collection process, the measurement tool was applied to the students personally.

2.4. Data Analysis

Scaling approach with pairwise comparison method was used in data analysis. The basis of this technique is the comparative judgement law. In the pairwise comparison method, observers are asked to choose which of the pairs of stimuli has priority. Non-discrimination judgments are not allowed (Turgut & Baykul, 1992). In the research, Thurstone's comparative judgment law was applied with the full data matrix in the equation of case V. First, the frequency matrix (F) is calculated for $U_k>U_j$. In this matrix, the column with the smallest frequency values is the most preferred factor among the compared factors. After the frequency matrix is created, the values in this matrix are divided by the number of observers. The new matrix obtained is the ratio matrix (P). Then the z value of each ratio in the ratio matrix is calculated and converted to the unit normal deviations matrix (Z). The mean of all columns in the unit normal deviations matrix is found. After taking the average of each column, the scale values (S_j) of the criteria are calculated. In order to obtain a nominal zero point in the scale, standardized scale values (S_c) are obtained by adding the absolute value of the smallest z value among the calculated averages to the mean in each column. At the end of all these processes, the standardized scale values are displayed on the numerical axis and ranked (Turgut & Baykul, 1992).

After the scale values were calculated and ranked on the numerical axis, the internal consistency of the scaling needs to be analyzed in order to check whether the assumptions valid for Thurstone's comparative judgment law equation of case V are correct and whether the observers are careful in their judgments. Internal consistency of scaling is determined by analyzing how similar the observed p_{jk} and obtained p_{jk} ratios are (Guilford, 1954). For this purpose, the internal consistency of the scale values was analyzed. Internal consistency value was calculated with the chi-square (χ^2) test. The Chi-Square value at 10 degrees of freedom was found to be 31.48. If the chi-square table value is for *sd* =10, the Chi-Square value is 18.307. Since 31.48>18.307, it was concluded that this value was significant at the level of 0.05 with 10 degrees of freedom. In such cases, it is recommended to use the case III, which does not require the assumption that the discrimination variances are equal for all stimuli (Turgut & Baykul, 1992).

After the results, the equation of case III was applied. First, the variances of the discrimination distributions are estimated for each stimulus (Z matrix). With the values obtained, the matrix of variance sums is reached. By taking the square roots of the elements of this matrix, the square root matrix of the variance sums is formed. The S matrix is obtained by multiplying the z_{jk} values obtained from the Z matrix with the square root matrix of the variance sums. The mean values are obtained by taking the column averages of the S matrix. Finally, the scale values are calculated by shifting the smallest mean value to zero. The order of the calculated scale values is shown on the numerical axis.

3. FINDINGS

At this stage of the research, the factors affecting the exam anxiety of 8th grade students were first scaled with the equation of case V. The Chi-square value calculated for the internal consistency of the scale values using the Transformed Difference Squares Matrix was obtained as 31.48. This value is greater than the Chi-square table value (Chi-Square=18.307 for *sd*= 10). In other words, the Chi-square value obtained is statistically significant at the level of a=0.05. According to this result, the data do not correspond to the assumptions of case V and pairwise comparison. In this case, Guilford (1954, p.154) recommends the use of the case III. Therefore, the calculations were continued with the equation of case III.

While reporting the results of the analysis, the factors affecting the exam anxiety of the students were coded as follows:

A: The thought of failing in lessons,

B: The effect of the social environment,

C: Students's self-perceptions,

D: Teacher attitudes,

E: Social and physical characteristics of the school,

F: The thought of not being prepared for exams adequately.

3.1. Scaling by Equation of Case V

In the study, 100 students were asked to compare 6 factors (stimulus) given to them in pairs. The frequency values of the stimuli after the pairwise comparison using case V are given in Table 1.

Table 1. *Frequency matrix.*

			U_j				
U_k	A	B	C	D	E	F	Total
A		31	31	61	34	74	231
B	69		33	52	33	74	261
C	69	67		56	40	78	310
D	39	48	44		26	66	223
E	66	67	60	74		72	339
F	26	26	22	34	28		136
Total	269	239	190	277	161	364	1500

The values in Table 1 show the number of participants who preferred column stimuli to row stimuli. For example, the number of participants who preferred stimulus B to stimulus A is 31. In this matrix, the total number of the opposite diagonals gives the number of observers, and this is 100. By proportioning the values in the frequency matrix to the number of observers, the ratio matrix in Table 2 is obtained.

Table 2. *Ratio matrix.*

			U_j				
U_k	A	B	C	D	E	F	Total
A	0	0.31	0.31	0.61	0.34	0.74	2.31
B	0.69	0	0.33	0.52	0.33	0.74	2.61
C	0.69	0.67	0	0.56	0.4	0.78	3.1
D	0.39	0.48	0.44	0	0.26	0.66	2.23
E	0.66	0.67	0.6	0.74	0	0.72	3.39
F	0.26	0.26	0.22	0.34	0.28	0	1.36
Total	2.69	2.39	1.9	2.77	1.61	3.64	15

Then, to find the z values of the values in the P ratios matrix, the NORMTERS function, which gives the inverse of the standard normal cumulative distribution in Excel, is used, and then it is passed to the unit normal deviations matrix. The unit normal deviations matrix is shown in Table 3.

Table 3. *Unit normal deviations matrix.*

			U_j			
U_k	A	B	C	D	E	F
A		-0.496	-0.496	0.279	-0.412	0.643
B	0.496		-0.44	0.05	-0.44	0.643
C	0.496	0.44		0.151	-0.253	0.772
D	-0.279	-0.05	-0.151		-0.643	0.412
E	0.412	0.44	0.253	0.643		0.583
F	-0.643	-0.643	-0.772	-0.412	-0.583	
Total	0.482	-0.309	-1.606	0.711	-2.331	3.053
Mean	0.08	-0.052	-0.268	0.119	-0.389	0.509
Sj	0.469	0.337	0.121	0.508	0	0.898

In the unit normal deviation matrix, the mean of the distribution is zero and the standard deviation is one. In this matrix, elements that are symmetric about the main diagonal are opposite in sign but equal in absolute value. The column totals, column averages and scale values (S_j) of the Z unit normal deviations matrix are calculated, respectively. While calculating the scale values (S_j), the starting point is shifted to be zero. The S_j values obtained from this matrix give the scale values of the stimuli for the equation of case V. The found scale values are shown in Figure 1. Factors scale values and stimulus orders affecting 8th grade students' exam anxiety according to equation of case V results are shown in Table 4.

Figure 1. *Scale values of the factors affecting the exam anxiety of 8th grade students by equation of case V.*

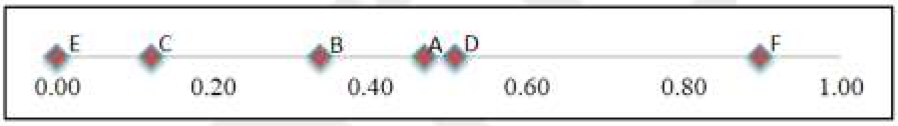


Table 4. *Factors scale values and stimulus orders affecting 8th grade students' exam anxiety according to equation of case V results.*

Factors	Scale Values	Stimulus Orders
A- Thought of failing in lesson	0.469	4
B- The effect of the social environment	0.337	3
C-Students's self-perceptions	0.121	2
D-Teacher attitudes	0.508	5
E- Social and physical characteristics of the school	0	1
F- Thought of not being prepared for exams adequately	0.898	6

Table 5. *Transformed difference squares matrix.*

	A	B	C	D	E	F
A	2025	66.007	10.727	29.568	1.485	21.6
B	66.007	2025	24.864	7.629	5.22	3.013
C	10.727	24.864	2025	28.546	9.084	0.005
D	29.568	7.629	28.546	2025	8.571	0.233
E	1.485	5.22	9.084	8.571	2025	41.901
F	21.6	3.013	0.005	0.233	41.901	2025
Total	258.453					
Chi-square	31.48026797					
sd	10					

In Table 5, the internal consistency of the scale values calculated with the equation of case V was tested with the chi-square statistics and the chi-square was found to be 31.48. This value is significant at the 0.05 level with 10 degrees of freedom. According to this result, the data does not satisfy the assumptions of the equation of case V or pairwise comparisons method. For this reason, calculations were continued by using the equation of case III.

3.2. Scaling by Equation of Case III

When the data does not satisfy the assumptions of the equation of case V or the pairwise comparison method, the calculations should be continued using the equation of case III. According to the results obtained in Table 5, the assumptions do not satisfy. For this reason, scaling was continued with the equation of case III. In Table 6, observer variances were estimated with the help of Z matrix values.

Table 6. *Z Matrix and estimation of observer variances.*

			U_j			
U_k	A	B	C	D	E	F
A		-0.496	-0.496	0.279	-0.412	0.643
B	0.496		-0.44	0.05	-0.44	0.643
C	0.496	0.44		0.151	-0.253	0.772
D	-0.279	-0.05	-0.151		-0.643	0.412
E	0.412	0.44	0.253	0.643		0.583
F	-0.643	-0.643	-0.772	-0.412	-0.583	
Vj	0.47	0.454	0.348	0.342	0.137	0.117
KVj	2.82	2.724	2.088	2.052	0.822	0.702
1/KVj	0.355	0.367	0.479	0.487	1.217	1.425
Σ 1/KVj	4.33					
K.C	2.771					
σj	-0.017	0.017	0.327	0.35	2.371	2.947
σj2	0	0	0.107	0.123	5.622	8.685

The variance values obtained for each stimulus in the last row of Table 6 were calculated by squaring the standard shifts in the last row. Then, these variances are summed in pairs to form the variance sum matrix in Table 7. By taking the square root of the values in Table 7, the square root matrix of the variance sums in Table 8 is reached.

Table 7. *Variance sum matrix.*

				U_j			
		A	B	C	D	E	F
U_k		0	0	0.107	0.123	5.622	8.685
A	0	0	0	0.107	0.123	5.622	8.685
B	0	0	0	0.107	0.123	5.622	8.685
C	0.107	0.107	0.107	0.214	0.23	5.729	8.792
D	0.123	0.123	0.123	0.23	0.246	5.745	8.808
E	5.622	5.622	5.622	5.729	5.745	11.244	14.307
F	8.685	8.685	8.685	8.792	8.808	14.307	17.37

Table 8. *Square roots of variance sums.*

U_j					
A	B	C	D	E	F
0	0	0.327	0.351	2.371	2.947
0	0	0.327	0.351	2.371	2.947
0.327	0.327	0.463	0.48	2.394	2.965
0.351	0.351	0.48	0.496	2.397	2.968
2.371	2.371	2.394	2.397	3.353	3.782
2.947	2.947	2.965	2.968	3.782	4.168

For the S matrix, the elements of the Z matrix and the square root of the variance sums are multiplied. The values above the diagonal of the S matrix in Table 9 are obtained. Elements below the diagonal and elements above the main diagonal are opposite in sign and equal in absolute value. In the S matrix, scaling is completed, as in the Z matrix of the case V, by averaging each column and shifting the smallest average value to zero.

Table 9. *S Matrix.*

	U_j					
U_k	A	B	C	D	E	F
A	0	0	-0.162	0.098	-0.977	1.895
B	0	0	-0.144	0.018	-1.043	1.895
C	0.162	0.144	0	0.072	-0.606	2.289
D	-0.098	-0.018	-0.072	0	-1.541	1.223
E	0.977	1.043	0.606	1.541	0	2.205
F	-1.895	-1.895	-2.289	-1.223	-2.205	0
Total	-0.854	-0.726	-2.061	0.506	-6.372	9.507
Mean	-0.142	-0.121	-0.344	0.084	-1.062	1.585
S_j	0.92	0.941	0.718	1.146	0	2.647

In the S matrix, the principal diagonal elements are 0. The S matrix elements are opposite in sign with respect to the main diagonal, but equal in absolute value, as in the Z unit normal deviations matrix. In Figure 2, the scale values calculated with the equation of case III are shown on the numerical axis.

Figure 2. *Scale values of factors affecting 8th grade students' exam anxiety by equation of case* III.

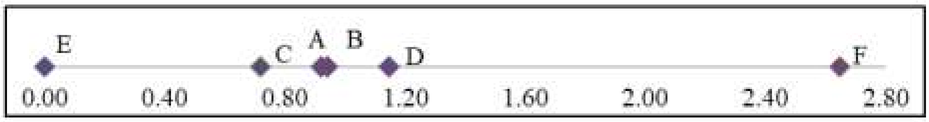


Table 10. *Factors scale values and stimulus orders affecting 8th grade students' exam anxiety according to equation of case* III *results.*

Factors	Scale Values	Stimulus Orders
A- Thought of failing in lesson	0.92	3
B- The effect of the social environment	0.941	4
C- Students's self-perceptions	0.718	2
D-Teacher attitudes	1.146	5
E- Social and physical characteristics of the school	0	1
F- Thought of not being prepared for exams adequately	2.647	6

When Table 10 is analyzed, it is seen that the first two criteria that most affect the exam anxiety of 8th grade students are the social and physical characteristics of the school and students's self-perceptions, respectively. These stimulants are followed by the thought of failing in lessons and the effect of the social environment. The two criteria that they think affect exam anxiety the least are teacher attitudes and the thought of not being prepared for exams adequately. As a result of the research, it has been seen that two factors -the thought of failing in the lessons and the effect of the social environment- have a similar effect on test anxiety.

4. DISCUSSION and CONCLUSION

In this study, it was aimed to analyze the factors affecting the exam anxiety of 8th grade students. In the scaling made for this purpose, first of all, the equation of case V of the pairwise comparison method was applied. Since the internal consistency of the results obtained was significant, the equation of case III was used.

According to the results of the research, the first two factors that most affect the exam anxiety of 8th grade students are the social and physical characteristics of the school and how students see themselves, respectively. These stimulants are followed by the thought of failing in the lessons and the effect of the social environment. The two criteria that they think affect exam anxiety the least are teacher attitudes and the thought of not being prepared for exams adequately.

The second most important factor affecting students' exam anxiety is students's self-perceptions academically. Bozanoglu (2005) in his group guidance study based on cognitive-behavioral approach stated that there are significant differences in the relationships between academic motivation, academic self-esteem, academic achievement and exam anxiety. In this study, the fact that the factor's "students's self-perceptions" being in the second place supports the results academically. The effect of the social environment (family, friends and relatives) is in the third place. The positive effect of group guidance based on the cognitive-behavioral approach on exam anxiety levels explains the high level of effect of the social environment. Students tend to get support from their social environments in order to reduce their exam anxiety.

Turan-Basoglu (2007), in his study analyzing the relationship between exam anxiety and self-confidence in adolescence, found that there is a negative correlation between exam anxiety and self-confidence of successful students. How students see themselves is related to their self-confidence. As a result of the current study, its second place shows parallelism with the literature. The family factor in the effect of the social environment is undeniable. Although it varies according to age groups, the attitude of the family towards students is another factor affecting exam anxiety. There is a relationship between exam anxiety levels and parental attitudes (Duman, 2008). The attitude of the family will positively or negatively affect students's self-perceptions. Social and physical characteristics of the school include the physical equipment of the school, academic, sports and artistic achievements, and activities for parents and students (Nartgün & Kaya, 2016). The more social and physical characteristics of the school are, the better and the more effectively the students will be able to prepare for the exams. As students become successful, their perceptions of themselves will tend to improve. All these factors are like a cycle. In future studies, the relations of these listed factors with each other or with other factors that are thought to be related should be searched. To sum up, as a result of our study, of the factors which affect exam anxiety the most, the social and physical characteristics of the school is in the first, how students see themselves is in the second, the thought of failing in the lessons is in the third, and the effect of the social environment is in the fourth place, which shows a parallelism with the literature.

The two factors that students think affect exam anxiety the least are teacher attitudes and the thought of not being prepared for exams adequately. Undesirable study habits, high expectations in exams, failing in responsibilities, family expectations and the fear that the level of intelligence will be evaluated by exam success are the reasons for exam anxiety (McDonald, 2001). When the literature is analyzed, it is seen that these two factors are among the variables that predict exam anxiety, but similar to the study, they are not important predictive variables.

This study includes a randomly selected study group consisting of 100 students. The scaling study is limited to six factors: the thought of failing in the lessons, the effect of the social environment, students's self-perceptions, teacher attitudes, the social and physical characteristics of the school, and the thought of not being prepared for the exams adequately. In future studies, the number of factors can be increased or different results can be obtained by using different factors. As a result of the research, it was seen that the two factors, the thought of failing in the lessons and the effect of the social environment, were interdependent. It is thought that it would be more beneficial to exclude one of these factors from the study. Considering these opinions, one of these two factors can be eliminated in future studies and different factors can be taken into consideration instead. The study group of this study consists of 8th grade students. Scaling study can be done at different education levels. Within the scope of the results of the study, it is possible to help students control their exam anxiety by providing group guidance on the factors that stand out and by making programs to cope with exam anxiety. In the study, scaling was analyzed in terms of exam anxiety. The importance of the results should be analyzed and practical measures should be taken in the education process. Both classroom guidance teachers and psychological counselors should carry out preventive studies in schools. In future studies, according to the results obtained in this study, scale development studies can be conducted on test anxiety and the factors affecting academic success. In addition, studies can be carried out with other scaling methods in different psychological or sociological subject areas.

Declaration of Conflicting Interests and Ethics

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

Authorship Contribution Statement

Hasibe Yahsi Sari: Investigation, literature review, visualization, data analysis, and writing-original draft. **Duygu Anil**: Research design, supervision, and validation.

Orcid

Hasibe Yahsi Sari https://orcid.org/0000-0002-0451-6034
Duygu Anil https://orcid.org/0000-0002-1745-4071

REFERENCES

Anil, D., & Inal, H. (2017). *Scaling Applications in Psychophysics.* Pegem Academy.

Anil, D., Taymur, M. O., & Oztemur, B. (2017). Scaling of the factors that are thought to be effective on the university preferences of the last grade students of high school by means of paired comparison method. *Trakya Üniversitesi Eğitim Fakültesi Dergisi*, *7*(1), 75-85.

Aydın, U., & Bulgan, G. (2017). Adaptation of children's test anxiety scale into Turkish: Validity and reliability study. *Elementary Education Online*, *16*(2), 887-899. https://doi.org/10.17051/ilkonline.2017.304742

Bacanlı, F., & Sürücü, A.G.M. (2006). An examination of the relationship between test anxiety and decision making styles of elementary school 8th grades students. *Educational Administration: Theory and Practice, 45*(45), 7-35.

Basol, G., & Zabun, E. (2014). The predictors of success in turkish high school placement exams: exam prep courses, perfectionism, parental attitudes and test anxiety. *Educational*

Sciences: Theory and Practice, 14(1), 78-87. https://doi.org/10.12738/estp.2014.1.1980

Basol, G. (2017). IDA test anxiety scale: Validity and reliability study. *The Journal of International Education Science, 4*(13), 173-193. https://doi.org/10.16991/INESJOURNAL.1506

Bozanoglu, I. (2005). The effect of a group guidance program based on cognitive-behavioral approach on motivation, selfesteem, achievement and test anxiety levels. *Ankara University Journal of Faculty of Educational Sciences, 38*(1), 17-42. https://doi.org/10.1501/Egifak_0000000110

Bulut, S.S. (2010). The effects of solution-focused brief group therapy on the treatment of exam anxiety, aggression tendencies and inadequacy in problem solving skills of secondary school students. *Gazi University Journal of Gazi Educational Faculty, 30*(2), 325-356.

Büyüköztürk, S., Kılıc Cakmak, E., Akgün, Ö.E., Karadeniz, S., & Demirel F. (2008). *Scientific research methods.* Pegem Academy.

Cakmak, A., Sahin, H., & Akinci-Demirbas, E. (2017). The analysis of relationship between test anxiety and self-esteem in the case of 7th and 8th grade students. *Kafkas University, e-Kafkas Journal of Educational Research, 4*(2). https://doi.org/10.30900/kafkasegt.315182

Coban, S. (2009). *The relationship of social environment's effects with juvenile delinquency and problem behaviours* (Publication No. 315049) [Doctoral dissertation, Hacettepe University]. National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Delioglu, H.N. (2017). *8th grade students of success of mathhematics and test and mathematics anxiety, mathematics self efficacy of investigation* (Publication No. 454808) [Master's thesis, Adnan Menderes University]. National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezDetay.jsp?id=ujzt86YZJ_iUP-fpWDH_MQ&no=AE31luBlryE0pZNS-qPPOQ

Driscoll, R. (2007). Westside test anxiety scale validation. *Education Research Information Center*. Retrieved from https://eric.ed.gov/?id=ED495968

Duman, G.K. (2008). *An Examination of the Relationship Between State – Trait Anxiety Levels, Test Anxiety Levels and Parental Attitudes in the 8th Grade Primary School Students* (Publication No. 220337) [Master's thesis, Dokuz Eylül University]. National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Genc, Y. (2016, August 24-26). *Factors affecting exam anxiety of students preparing for university exam* [Conference presentation]. In ICPESS, İstanbul,Turkey.

Guilford, J.P. (1954). *Psychometric Methods.* McGraw Hill Inc.

Güler, D., & Cakir, G. (2013). Examining predictors of test anxiety levels among 12th grade high school students. *Turkish Psychological Counseling and Guidance Journal, 4*(39), 82-94.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58*(1), 47-77. https://doi.org/10.3102/00346543058001047

Kavakci, Ö., Güler, A.S., & Cetinkaya, S. (2011). Test anxiety and related psychiatric symptoms. *Journal of Clinical Psychiatry*, *14*(1), 7-16.

Kayapinar, E. (2006). *Research into anxıety level of the 8th grades students at primary schools preparing for secondary school student selection and placement examination (The sample of Afyonkarahisar city)* [Master's thesis]. Afyon Kocatepe University. https://tez.yok.gov.tr/UlusalTezMerkezi/tezDetay.jsp?id=mD__s2fD2RrL4dDOaXEyqA&no=hv0DeEk7ECSiHlhNYWqsFA

Kaya, M., & Savrun, B.M. (2015). Relationship between attachment styles and test anxiety of students who will take the common exam for transition to secondary education system. *Journal of New Symposium*, *53*(3), 32-42. https://doi.org/10.5455/NYS.20151215070858

Kesici, A., & Asilioglu, B. (2017). Developing stress scale for secondary school students: reliability and validity study. *Kastamonu Education Journal, 25*(6), 2413-2426.

Mandler, G., & Sarason, S. B. (1952). A study of anxiety and learning. *The Journal of Abnormal and Social Psychology, 47*(2), 166–173. https://doi.org/10.1037/h0062855

McDonald, A.S. (2001). The prevalence and effects of test anxiety in school children. *Educational Psychology*, *21*(1), 89-101. https://doi.org/10.1080/01443410020019867

Nartgün, Ş., & Kaya, A. (2016). Creating school image in acccordance with private school parents' expectatons. *Journal of Research in Education and Teaching, 5*(2), 153-167.

Speilberger, C.D. (1980). *Manual For-State Trait Anxiety Inventory*. California Consulting Psychologisis Press.

Thurstone, L.L. (1927). A law of comparative judgment. *Psychological Review, 34(*4), 273–286. https://doi.org/10.1037/h0070288

Totan, T., & Yavuz, Y. (2009). The validity and reliability study of the turkish version of westside test anxiety scale. *Mehmet Akif Ersoy University Journal of Education Faculty*, *9*(17), 95-109.

Tugan, S.E. (2015). Relationship between test anxiety and academic achievement. *Karaelmas Journal of Educational Sciences*, *3*(2). 98-106.

Turan Başoğlu, S. (2007). The relationship between the pre-exam anxiety and the self-confidence and to search the effects of those oncepts on the neolagnium (Publication No. 217683) [Master's thesis, Maltepe University]. National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Turgut, M.F., & Baykul, Y. (1992). *Scaling techniques.* ÖSYM Press.

Wren, D.G., & Benson, J. (2004). Measuring test anxiety in children: Scale development and internal construct validation. *Anxiety, Stress & Coping*, *17*(3), 227-240. https://doi.org/10.1080/10615800412331292606

Yelken, R. (2011). The central institution in the prevention of juvenile delinquency is the family: Analysis of international conventions on children. Hancerli, Sevinc, Gürer & Oner (Ed.), *Driven to Crime and Victims Children* (1st ed., pp. 32-41). SABEV Press.

Zeidner, M. (1998). *Test anxiety: The state of the art.* Plenum Press.

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 1, 233–253

https://doi.org/10.21449/ijate.884855

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

The thematic content analysis of the scales used in citizenship education

Melehat Gezer[1,*]

[1]Dicle University, Faculty of Education, Department of Social Studies Teaching, Diyarbakır, Turkiye

ARTICLE HISTORY

Received: Feb. 22, 2021
Revised: Jan. 17, 2022
Accepted: Feb. 02, 2022

Keywords:
Citizenship,
Citizenship education,
Scale development,
Scale adaptation,
Thematic content analysis.

Abstract: This study aimed to scrutinize the scales used in citizenship education in Turkey through thematic content analysis. In the study, all of the scales developed/adapted within the scope of citizenship education without a year limitation were reviewed and 56 scales found in these studies were evaluated. The document analysis was used as the method of data collection. It was determined that the scales examined in the study were mostly published within the scope of doctoral dissertations and articles. Most of the scales were developed/adapted in 2016, a great majority of which were developed by researchers themselves whereas a small number of which were adapted from other cultures into Turkish. The most frequently used key words in the studies where the scales were available were "citizenship", "social studies" and "citizenship education". The sample mostly used in the scales were composed of university students and the most frequently used sample size included 201-300 participants. It was concluded that the relevant scales considered multi-factor structures in relation to citizenship. In addition, a number of deficiencies were found in analysing the psychometric properties and recommendations were made accordingly.

1. INTRODUCTION

Citizenship, which originated from the word citizen in antique Greek city states, today represents individuals' membership/loyalty to the state or political community they belong to. In addition to having a number of rights as a result of such adherence, citizens also take on certain duties and responsibilities. Citizenship as a concept which has political, legal and social bases and a dynamic phenomenon has changed throughout history. This concept has broadened more throughout history on the axis of social and economic changes and has taken on new dimensions. The concept has gone beyond expressing loyalty to a state or a community and acquired transnational properties especially along with globalisation and with the advance of technology. The reasons for the transformation in the conception of citizenship include several factors such as global environmental problems (the unproportioned use of nuclear energy-based products, global warming, climate change, environmental pollution, extinction of species of animals, etc.), the need for digital literacy (problems related to confidentiality and protection of personal information, cyber-attacks, cyber loafing, cyber bullying, etc.) and the protection of minority rights (not paying attention to the law of immigration and refuge, not giving the right of self-management, refusing cultural diversity, increase in racist movements, etc.). The above-mentioned problems concern not only a community but also the whole universe and pose a

[*]CONTACT: Melehat Gezer ✉ melehatgezer@gmail.com Dicle University, Faculty of Education, Department of Educational Sciences, Social Studies Teaching, Diyarbakır, Turkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

threat to all humanity. It is a reality that those problems cannot be eliminated with traditional mentality of citizenship (Özel, 2007) because it is no longer considered adequate to be aware of responsibilities for one's country and to perform the duties. Citizens who have responsibilities for the whole world, who choose to play a part in solving the problems, and who take action accordingly are needed today (Şahin, et al., 2016). This situation has necessitated the changes in the types of citizens that many countries wish to raise or have (Eurydice, 2017; Gezer, 2020).

Strengthening citizenship competencies through education has recently become an important theme in politics, the public, and the scientific world (Eurydice 2012, 2017). Basically, it is aimed, with citizenship education, to ensure the active participation of individuals in political and social life as free individuals, to raise awareness of individuals about the protection and support of the democratic system with common democratic values (Loobuyck, 2021), and to prepare individuals for citizenship with the awareness of their citizenship duties and responsibilities (Kerr, 1999). However, there is no common view on the content of citizenship education. Therefore, the citizenship education includes diverse content and objectives. Some countries may place more emphasis on ensuring that students have the knowledge, skills and attitudes necessary to become active and socially responsible citizens. Others may prioritize effective and constructive interaction within and between communities or pay more attention to the development of personal traits such as critical thinking (Eurydice, 2017). Thus, citizenship and citizen competencies in several areas have been re-defined and new dimensions to citizenship such as environmental/ecological citizenship, cultural citizenship, multi-cultural citizenship, minority citizenship, digital citizenship, active citizenship, economic citizenship, democratic citizenship, and constitutional citizenship have been put forward.

The expansion of perspective and meaning in the concept of citizenship has also necessitated taking those dimensions into consideration in measuring citizenship. Therefore, new scales for measuring the multi-dimensional structure of citizenship were developed (Erdem & Koçyiğit, 2019; Beseler, et al., 2021; Çermik & Akçay, 2020; Hadjichambis & Paraskeva-Hadjichambi, 2020; Homer, 2020; İşman & Güngören, 2014; Karatekin & Uysal, 2018; Kim, & Choi, 2018; Lo, et al., 2019; Şahin & Çermik, 2014; Yazıcı et al., 2017; Yıldırım, 2018), all of which were included in the literature. Depicting the scales in the literature in general terms is important in understanding the attempts better at measuring citizenship. Hence, the present study aims to put the citizenship scales used in the Turkish literature in the area of social studies education to thematic content analysis. This research functions as a scientific resource where researchers studying in the field of citizenship education can see the scales developed/adapted on the subject. The results of this specific research also present a general picture of which dimensions and at which education level citizenship education is mostly studied. In other words, what dimensions of citizenship education are considered in more detail and what dimensions are considered in a limited manner can be seen by means of this study. In this regard, the study is thought to guide researchers in terms of including in the literature the required dimensions related to scales. In addition, the scales developed/adapted within the scope of citizenship education in the research are evaluated in terms of validity and reliability processes, so the present research also provides information about the compliance of the relevant scales with the scale development/adaptation standards.

Studies of thematic content analysis are capable of contributing to making knowledge widespread and shaping the future research studies in that they consider UpToDate studies in a holistic perspective and that they demonstrate the similarities and differences between studies (Braun & Clarke, 2006; Çalık & Sözbilir, 2014). Accordingly, analysing the scales developed or adapted in citizenship education through thematic content analysis contributes to the literature. The scales are evaluated in a critical perspective and efforts made to give a general picture of the weaknesses and strengths of the studies which used the scales. Therefore, it is expected to function as a scientific resource for researchers who plan to perform studies on citizenship education since it presents the current measurement instruments. In this way, it also provides researchers in the area of citizenship education with a resource in which the researchers can see the contemporary scales put together.

Review of relevant literature demonstrates that several studies have been conducted to examine the scale development/adaptation research. Some of these studies evaluate the compatibility of the stages followed with scale development/adaptation processes independently of the subjects and disciplines in which the scales are developed or adapted (Acar Güvendir & Özer Özkan, 2015; Çüm & Koç, 2013; Kaya Uyanık et al., 2017; Tavşancıl et al., 2014; Yurdabakan & Çüm, 2017). Some other studies analyse the scale development/adaptation activities in certain disciplines such as mathematics and science (Delice & Ergene, 2015; Ergene, 2020; Tosun & Taşkesenligil, 2014), management and organisational behaviour (Kanten & Arda, 2020) and music education (Çelik & Yüksel, 2020). Some others, on the other hand, act more specifically and make a content analysis of the scales in a specific subject only. For instance, Chandu et al. (2020) analyse the measurement instruments about corona virus in more recent studies. Studies which consider the scales related to citizenship education with thematic content analysis, however, are not available in the literature. The study is therefore thought to be original in this sense.

1.1. Research Questions

This study aims to analyse the scales available in citizenship education in Turkey through thematic content analysis. In this sense, the problem sentence of the research is "How is the current situation regarding the scales developed/adapted within the scope of citizenship education in Turkey?" Based on this main problem, answers to the following sub-problems were sought in the study:

1. What is the distribution of the scales according to types of studies in which the scales are available?
2. What is the distribution of the scales according to years?
3. What is the distribution of the scales according to whether they are developed or adapted?
4. What are the key words used in the studies where the scales are available?
5. What is the distribution of the scales according to the stage of education for which validity and reliability tests were done?
6. What is the sample size used in the scales and is it enough when the number of items in the scale is considered?
7. What are the reliability estimating methods used in the scales?
8. What are the proofs of validity used in the scales?
9. What is the distribution of the scales according to the number of factors they have?
10. What is the distribution of the scales according to subjects?

2. METHOD

2.1. The Research Model

This research, which aims to examine the trends in scale development and adaptation studies in the field of citizenship education, is suitable for thematic content analysis. Several researchers emphasise that they do not consider thematic content analysis as a separate research method because it is a procedure employed in qualitative studies and that it should be considered as a technique which provides researchers with convenience (Nowell et al., 2017). Thematic content analysis is a qualitative research technique which involves describing, analysing and reporting the patterns in the data (themes) (Braun & Clarke, 2006). The technique is quite useful in summarising or analysing the basic properties of large qualitative data sets (Nowell et al., 2017). Studies using thematic content analysis are important in that they provide researchers who study in relevant areas, who cannot reach all the studies in the area, and who cannot analyse them systematically with rich resources (Ültay & Çalık, 2012).

2.2. Data Collection

The developed/adapted scales were reached by reviewing the national thesis data centre of the Council of Higher Education and by searching via Google scholar. No year limitation was applied during the browsing. The pages in Turkish were scanned by writing the Turkish key words "citizenship and scale*" on Google search engine in determining the scales to be analysed. All the probable results can be found by using the mark "*" at the end of the words while searching on Google. On searching by writing "scale*", for instance as in this study, both the word "scale in singular form" and its derivations "scales", "of the scale(s)", and "scale(s) in object position in sentences" can be accessed. Browsing was terminated on 17th January 2021. Therefore, the studies published after that date was excluded from the scope of this study.

After browsing the scales on Google scholar, the key words "scale" and "citizenship" were entered in the detailed search section on the national thesis centre database, the area of "social" was chosen and thus the scales were searched. The criteria for selecting the theses to be analysed were entered on the detailed search page of the thesis centre of the Council of Higher Education of Turkey, and thus the theses with scales developed or adapted in relation to citizenship education were included in the scope of this study. For the scale development/adaptation articles produced from the theses, only the thesis study in which the scale was published was taken into account during the scanning process. As a result, 56 scales in total which were developed or adapted were reached. All the studies mentioned are listed in Appendix-1.

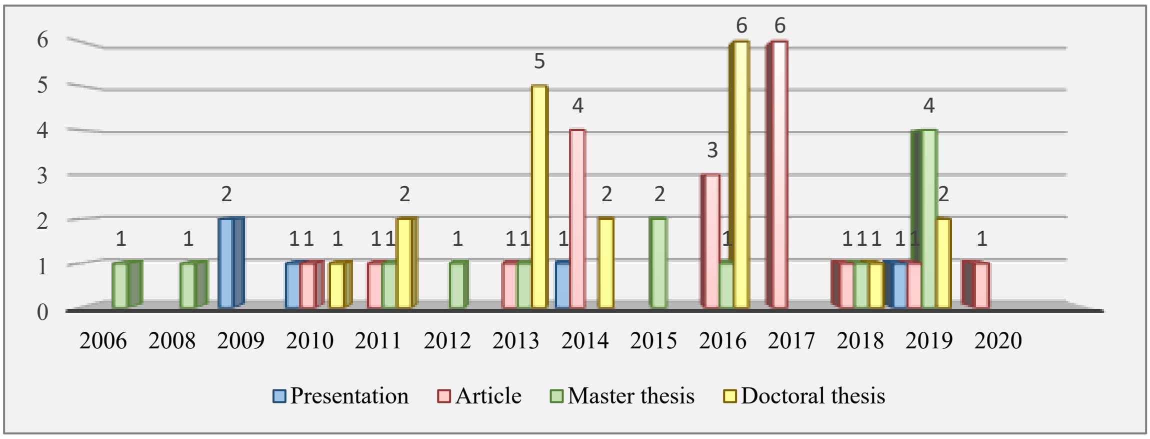
2.2. Data Analysis

Content analysis technique was used in the evaluation of the data obtained in this study. The purpose of content analysis is to reach concepts and relationships that can explain the collected data. The data summarized and interpreted in descriptive analysis are subjected to a deeper processing in content analysis, and concepts and themes that cannot be noticed with a descriptive approach can be discovered as a result of this analysis. The basic process in content analysis is to gather similar data within the framework of certain concepts and themes and to organize and interpret them in a way that the reader can understand (Yıldırım & Şimşek, 2011). A checklist to help to analyse the scales used in citizenship education was created prior to the content analysis. The checklist aimed to set standard criteria for content analysis of the scales and consisted of two sections called "study tag" and "theoretical information". When we look at the content analysis studies in the literature (Kaya Uyanık et al., 2017; Taşdelen Teker & Güler, 2019), it is seen that in the case of using a checklist, expert opinion is sought to determine the suitability of the checklist in terms of scope and content. From the point of this view, two experts of measurement and evaluation who studied scale development and scale adaptation were consulted for their opinions of the checklist. The experts recommended that measurement invariance, convergent validity, and divergent validity also be included in the heading of "validity and reliability evidence" as a label used in describing the studies analysed. Thus, the checklist was modified to include the suggestions (see Appendix-2). It has no criteria which can be interpreted differently by different individuals in the checklist. Therefore, coding by one expert was considered sufficient. While analysing the data it was found that the sample groups were described as university students in some of the scales whereas they were described as prospective teachers in some others. For this reason, in the study, the mentioned distinction was followed in the coding of the sample group. In addition, the statistical processes in testing the psychometric properties of the scales were categorised separately for scale development and scale adaptation studies because evidence provided for the validity and reliability can differ in scale development studies from the ones in scale adaptation studies. Frequency and percentage analyses of codes were calculated for each theme.

3. FINDINGS

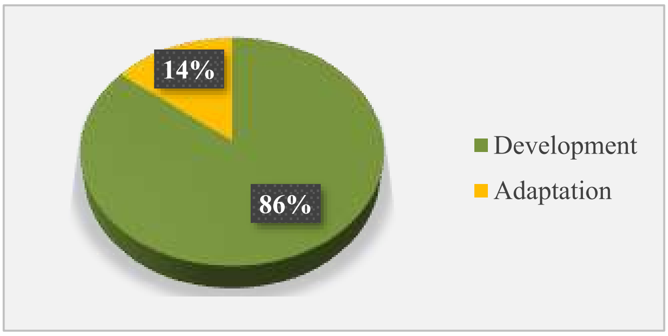
The type of studies in which the scales were published and their distributions according to years were checked in this study. The findings obtained are shown in Figure 1.

Figure 1. *The distribution of the scales according to years and types of study.*



According to Figure 1, the studies in which citizenship scales were developed or adapted were mostly in the year 2016. There was only one study in 2006, 2008, 2012 and 2020 each. The majority of the scales of citizenship education were in doctoral theses and in articles. The number of M.A theses was smaller than the number of doctoral theses or of articles, and the number of conference presentations was much smaller. Having analysed the scales according to years and types of studies, they were analysed according to whether they were developed by researchers themselves or they were adapted from another culture. The findings in this respect are shown in Figure 2.

Figure 2. *The distribution of the scales according to whether they were developed or adapted.*



It is apparent from Figure 2 that the majority of the scales analysed (86%) were developed by researchers themselves while the minority of them (14%) were adapted into Turkish culture from other cultures. After that, the answer was sought to the question of "what key words were used in the studies analysed?" The frequency of the key words used in the studies in which the scales were available is shown in Table 1. The frequencies for the key words were visualized through frequency questioning instead of tabulating them since the number of the key words was great. The size of the shapes obtained through word frequency questioning in Table 1 is directly proportionate to the frequency weight of the words.

Table 1. *The frequency questioning for the key words.*

Key Words	Frequency	Key Words	Frequency
Citizenship	21	Secondary School	3
Social Studies	16	Prospective Teachers	3
Citizenship Education	13	Democracy	3
Digital Citizenship	7	Democracy Education	3
Social Studies Education	6	Citizen	2
Peception of Citizenship	4	Democratic Citizenchip	2
Global Citizenship	4	Active Citizenchip	2
Effective Citizenchip	4	Socio-Scientific İssues	2
Validity	4	Factor Analysis	2
Reliability	4	Global Citizenship Education	2
Scale Development	4	Social Media Citizenship Perception	2
Teacher	4	Internet	2
Effective Citizenship Self-Efficacy Scale	3	Citizenship Knowledge	2
Globalization	3	Human Rights	2
Global Citizen	3	Consciousness of Citizenship	2
Values Education	3	Responsibility	2
Character Education	3	The Other	44

According to Table 1, the most frequently used key words in the studies were citizenship, social studies, and citizenship education, followed by digital citizenship, social studies education, perception of citizenship, global citizenship, effective citizenchip, validity, reliability, scale development, and teacher. There are also 44 keywords with a frequency of 1. After analysing the key words, the sampling stages and sample size were checked. It was determined that two of the scales published in the form of conference presentations were applied to secondary school students, two to prospective teachers and one to university students. Six of the scales published as article were applied to teacher candidates, five to secondary school students, five to university students, one to social studies teachers, and one to high school students. It was determined that 11 of the scales published within the scope of the thesis were applied to secondary school students, 10 to prospective teachers, four to university students, four to secondary school teachers, two to primary school students, one to high school students, and one to university students. The findings for the distribution of the studies according to sample size are shown in Table 2.

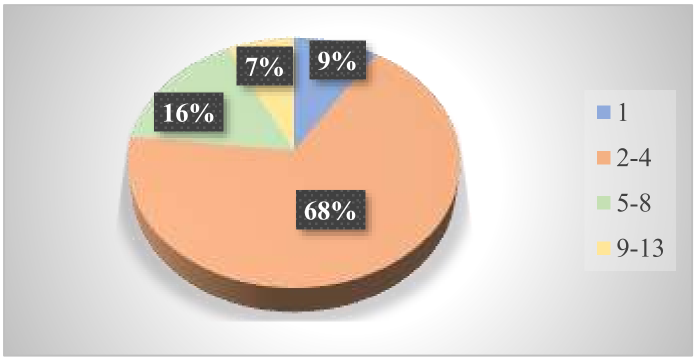
Table 2. *Findings for sample size and number of item in the scales.*

Research Id	Number of Item		Sample Size	Research Id	Number of Item		Sample Size
	Initial form*	Last form*			Initial form*	Last form*	
1	43	18	392	29		55	710
2		73	414	30		43	786
3		18	635	31	29	29	291
4	45	27	480	32	29	29	291
5	41	28	150	33	22	14	432
6	45	29	374	34		35	100
7	59	29	272	35	24	13	623
8	28	18	229	36		5	116
9		34	1063	37	51	16	635
10	120	67	625	38		20	494
11	80	32	480	39		28	238
12	13	11	241	40	64	45	500
13	31	27	180	41	45	38	672
14	44	24	532	42	41	20	1028
15	74	48	311	43	46	25	503
16	24	23	317	44	53	38	2190
17		21	400	45	10	10	250
18	145	87	2144	46		28	1099
19	42	25	544	47	35	35	185
20	63	49	438	48	23	21	295
21	65	33	670	49	40	15	200
22		25	297	50	40	23	100
23		30	241	51		84	150
24	42	18	288	52		20	1467
25		30	429	53	36	29	323
26	56	33	183	54	11	10	323
27	51	20	249	55	9	7	323
28	84	57	352	56		25	552

*While the initial form includes the number of items included in factor analyses and the last form includes the number of items in the scale as a result of factor analyses.

According to Table 2, the most frequently used sample size in the studies analysed were between 201 and 300 participants. In addition, the sample size of eight out of 56 studies was less than five times the number of items in the initial form. The number of factors was focused on after analysing the sample size. Figure 3 shows the findings.

Figure 3. *The distribution of the number of factors in the scales.*



It is apparent from Figure 3 that the number of scales with one factor is limited (only 9%). Most of the scales (68%) have 2-4 factors. The scales with 5-8 factors and those with 8-13 factors are also few. Table 3 below shows the distribution of the scales according to their subject matter.

Table 3. *The distribution of the scales of citizenship education according to their subject matter.*

	Primary School	*f*	Secondary School	*f*	High School	*f*	University Students	*f*	Prospective Teachers	*f*	Teachers of Social Studies	*f*
Conference presentations			Knowledge of citizenship	1			Global citizenship	1	Active citizenship	1		
			Perceptions of citizenship	1					Citizenship competencies	1		
Articles			Citizenship skills	1	Feelings of citizenship	1	Digital citizenship	2	Global citizenship	3		
			Attitudes towards citizenship education	2	Perceptions of good citizenship	1	Global citizenship	1	Ecological citizenship	1	Ideology of citizenship education	1
			Digital citizenship	1			Active citizenship	2	Effective citizenship	1		
			Effective citizenship	1					Digital citizenship	1		
Theses	Active citizenship	1	Attitudes towards the citixenship and Democracy education course	4	Democratic citizenship	1	Perceptions of citizenship on social media	1	Good citizenship	1	Digital citizenship	1
	Global citizenship	1	Perceptions of citizenship	2			Perceptions of citizenship	2	Perceptions of citizenship	1	Perceptions of the goals of citizenship course	1
			Citizenship consciousness	2			Citizenship competencies	1	Attitudes towards patriotism	1	Perceptions of the activities in citizenship course	1
			Ecological citizenship	1			Digital citizenship	1	Types of citizenship	1	Perceptions of the teaching quality in citizenship course	1
			Global citizenship	1					Global citizenship	1		
			Digital citizenship	1					Active citizenship	2		
									Effective citizenship	1		
									Digital citizenship	1		

According to Table 3, the scales used in citizenship education centre around global citizenship, digital citizenship, effective citizenship, and active citizenship. There are limited number of instruments to measure ecological citizenship. Besides, it was also found that the scales for determining the attitudes towards the citizenship education course, perceptions of good citizenship and of the concept of patriotism, citizenship consciousness, knowledge of citizenship, and citizenship competencies were developed or adapted. Finally, the reported validity and reliability evidence was checked. The findings are shown in Table 4 by taking scale development and scale adaptation into consideration.

Table 4. *The number of validity/reliability evidence reported in the scales of citizenship education.*

			Developed	Adapted
Validity evidence	Construct	Confirmatory factor analysis (CFA) only	0	4
		Exploratory factor analysis (EFA) only	27	2
		Both CFA and EFA	18	2
		Convergent-divergent validity	0	0
		Known groups validity	0	0
		Criterion related validity	1	1
		Measurement invariance	0	0
		Unreported	3	0
	Content	Expert opinion	40	–
		Content validity rate index	0	–
		Unreported	8	–
	No validity evidence was reported		1	0
Item analysis	Only lower-upper groups were compared		6	0
	Only item-test correlations		8	5
	Both lower-upper groups compared and item-test correlations		8	0
	Unreported		26	3
Reliability evidence	Only Cronbach's Alpha		48	8
	Test-retest method		5	1
	Parallel forms method		2	0
	Split half reliability (Spearman Brown and Guttman)		7	1
	No reliability evidence was reported		0	0

As evident from Table 4, only CFA was used in four out of eight scale adaptation studies whereas both EFA and CFA were used in two studies and only EFA was used in two studies. It was found on analysing the scales adapted in terms of item analysis that five out of eight scales did item analysis but that the remaining three scales did not do item analysis. Item-test correlation was calculated in all of the five scales in which item analysis was done. Besides, it was also found that Cronbach's Alpha internal consistency coefficient was used so as to estimate the reliability of the measurements in all the adaptation studies. Test-retest and split half reliability was also calculated in addition to internal consistency reliability in one of the scales.

According to Table 4, only EFA was done for construct validity in scale development studies mostly EFA and CFA were used in combination in a considerable number of studies. There were no scale development studies in which only CFA was used In one of the studies, however, criterion related validity was calculated in addition to EFA and CFA. The number of studies that offered no statistical evidence for construct validity was three.

Expert opinion was consulted for content validity in 40 out of 48 scale development studies. Content validity rate index was not calculated in any of them. Thus, no evidence was provided for content validity in eight of the studies. No evidence was provided for validity in one of the studies. Convergent validity, divergent validity, known groups validity, or measurement invariance were not tested in any of the scale development studies as in the case in adaptation studies. On revising the scale development studies in terms of item analysis, only lower group-upper group comparison was looked at in six studies, only item-test correlations were looked at in eight studies and both the lower group-upper group comparison and item-test correlations were looked at in eight studies. On the other hand, no statistical findings were found in 26 studies.

Cronbach's Alpha internal consistency reliability was calculated in all of the scale development studies as in scale adaptation studies. Split half reliability beside Cronbach's Alpha internal consistency coefficient was reported in seven of the studies. It was found that test-retest reliability was calculated in one of the seven studies and that test-retest reliability and parallel forms reliability were calculated in one of the seven studies. Besides, it was also found that the only reliability evidence provided apart from Cronbach's Alpha was test-retest coefficient. In one of the studies, on the other hand, test-retest reliability and parallel forms reliability in addition to Cronbach's Alpha were checked.

4 DISCUSSION and CONCLUSION

In this study, the scales developed or adapted for citizenship education in Turkey were put to thematic content analysis with no limitation on the year when they were developed or adapted. Thus, 56 scale development/adaptation studies in total were analysed in the research. The analyses demonstrated that the studies were conducted mostly in the year 2016. In terms of years, it can be said that there has been an increase in the number of studies towards the present. The majority of the scales were developed/adapted in doctoral theses and articles, while only a few of them were published in MA theses and in conference presentations. It is necessary to work with relatively large samples and to follow a multi-stage process in scale development/adaptation studies. This situation can be considered as the reason why researchers do not engage in such a process much in their master's theses and conference presentations. That is, when they consider the effort they put into collecting and analyzing the data, many researchers tend to publish their studies as an article instead of concluding their efforts with a conference presentation. In addition, the time needed for the scale development and adaptation processes may cause time anxiety for researchers writing master's thesis. Since doctoral theses are spread over a longer period compared to master's theses, researchers may experience less temporal anxiety about undertaking the scale development/adaptation process. In addition, researchers writing a master's thesis may be reluctant to engage in the scale development/adaptation process due to their lack of knowledge/experience and being at the beginning of the road. These listed factors can be considered as the reasons why scale development/adaptation processes are heavily used in doctoral theses, but not so often in master's theses.

Secondly, it was concluded that the majority of scales were developed and only a few of them were adapted to Turkish culture. Citizenship is a dynamic structure and the meaning attributed to the concept of citizenship can also differ from society to society (Schugurensky, 2005). For example, military service in Iceland is not a civic duty, whereas in Israel, military service is one of the basic duties for all citizens, regardless of whether they are men or women. In Turkey, military service is defined as a duty for every male citizen who does not have a health problem. This structure of citizenship, which changes from society to society and which is open to being affected by cultural elements, may have led researchers to scale development studies instead of adapting measurement tools developed in different cultures into Turkish.

When the keywords used in the studies in which the scales were examined, it was determined that the words "citizenship", "social studies", and "citizenship education" were mostly used. Such a result is not surprising, considering the fact that a search was made with the expression of citizenship while determining the studies to be examined in the research. In this sense, the result that is more striking about the keywords and that needs to be interpreted more is that the subject area in which the scales are developed is frequently included in the keywords. For example, the keywords of digital citizenship, global citizenship, and citizenship perception were used more frequently.

Considering the preferred sample level in the scales, it was concluded that the most frequently studied group is university students (pre-service teachers and students studying in other faculties). University students are generally a more accessible sample for researchers. Therefore, researchers may prefer to study on university students rather than studying at other educational levels. It is thought that individuals' understanding of citizenship is shaped in early adulthood because the first thing that comes to mind when the citizen is mentioned is usually the adult person. Being young, on the other hand, corresponds to a transition/becoming rather than a state and is somewhere between a child and an adult. Therefore, in this early adulthood, the individual learns citizenship through certain experiences (Kalaycıoğlu & Çelik, 2008). The fact that individuals' understanding of citizenship is only fully shaped in their first adulthood may be another reason for the scales being developed/adapted mostly on university students. When the studies were examined according to the sample sizes, it was determined that the scale development and adaptation studies were mostly carried out with research groups containing 201-300 participants. In other words, researchers may have preferred sample sizes between 200 and 300, with the thought of being sufficient for validity and reliability analyses on the one hand, and being economical on the other. Furthermore, the sample size of eight out of 56 studies was less than five times the number of items in the initial form. Researchers have made different suggestions about the number of participants that should be included in factor analysis studies. Cattell (1978) recommends that three to six times the number of items in the scale be included in the study group in factor analysis studies and states that 200 participants are acceptable for factor analysis and 500 participants is a very good number. Gorsuch (1983) recommends that there be at least five participants in the study group for each item in the scale in factor analysis studies, however, he states that the number of participants should not be less than 100 (Cramer, 2003). These suggested criteria for sample size may be a reference for researchers. As a result, it can be said that in most of the studies, the sample size selection was determined to be five times the number of items.

Another result of the research is related to the number of factors in the examined scales. While a single factor structure was observed in a very small part of the scales, a multidimensional structure consisting of at least two factors was revealed in most of the scales. This result can be interpreted as a reflection of the transformation of citizenship into a comprehensive concept that includes many dimensions.

When the scales subject to the research were examined in terms of validity evidence, it was determined that EFA was applied in almost all of the scale development studies. EFA is an exploratory analysis to reveal the structure observed in the scale. Since there is no empirical dimensioning in scale development studies, it is expected that EFA will be applied in almost all of these studies. In a substantial part of the scale development studies, CFA was performed together with EFA Since it is not possible to reach definite results in social sciences as in science, it is important to support the results obtained with more than one evidence. It is estimated that the use of CFA together with EFA in studies stems from this idea.

When we look at the scale adaptation studies, it can be seen that the rate of research that does not include EFA and only applied CFA is higher than that of scale development. In scale adaptation studies, there is a measurement tool that was previously developed in another culture, that is, the factor structure was empirically revealed. In other words, in adaptation studies, it is questioned whether the factor structure in the culture in which the scale was developed is also valid for a certain target culture. In this respect, it is a reasonable result that only CFA was applied in most of the scale adaptation studies. Convergent-divergent validity and measurement invariance were not tested in any of the scale development/adaptation studies. This result can be attributed to the competence levels of the researchers who developed/adapted the scales in the field of measurement and evaluation. Indeed, when the related literature is reviewed, it is seen that almost all of the studies on measurement invariance in Turkey belong to researchers in the field of measurement and evaluation. This situation observed in measurement invariance also shows itself in convergent-divergent validity studies. The fact that criterion-related validity was included in only one study can be explained by the difficulty in the data collection process because this type of validity requires the application of another measurement tool related to the subject of the scale, along with the scale that is aimed to be developed/adapted to the participants. This can complicate the data collection process and make it difficult for researchers to test criterion-related validity.

In most of the scale development studies examined in the research, expert opinion was sought in order to make a judgment about the content validity. The assessment of content validity relies on using a panel of experts to evaluate instrument items and rate them based on their relevance and representativeness to the content domain. It is recommended to use the statistics such as percent agreement and modified Kappa in order to obtain a Content Validity Index (CVI) based on expert judgements. Content validity indices are essential factors in the instrument development process and should be treated and reported as important as other types of construct validation (Almanasreh, et al, 2019). However, no study has reported the content validity index based on expert opinions. More clearly, the evidence presented for content validity remained at the qualitative level, but was not converted into a quantitative value. In some of the studies, no information was given about the processes to ensure content validity. However, no matter how well the researcher has reviewed the literature and prepared the items carefully, he/she should seek the opinions of experts in order to ensure the content validity of the measurement tool he/she has created and preferably make these views more concrete by calculating the content validity index.

Considering the reliability evidences, it was seen that the Cronbach alpha internal consistency coefficient was calculated in all of the scale development and adaptation studies. This result is in line with the conclusions of the study in which Acar Güvendir and Özer Özkan (2015) examined the articles on scale development and adaptation in the field of educational sciences and the studies in which Şahin and Boztunç Öztürk (2018) subjected the scale development studies in the field of education to content analysis. The number of studies in which parallel form and test-retest reliability were calculated is very limited. This result can be associated with Cronbach's alpha being a more useful reliability determination method for researchers. More clearrly, Cronbach's alpha is a reliability estimation method based on a single application. On the other hand, test-retest reliability requires applying the scale to the same group twice with a certain time interval. Again, in the parallel form method the reliability coeffeicient is estimated by administering another measurement tool that measures the same construct as the scale developed/adapted by the researcher to the same sample group (Boztunç Öztürk & Şahin, 2021). It can be stated that the necessity of two different applications leads to the use of test-retest and parallel form reliability less frequently.

Another remarkable result regarding reliability is that in addition to Cronbach's alpha, split-half reliability is reported in some measurement tools. Like Cronbach's alpha, split-half reliability is a useful method of determining internal consistency. The average of all possible split-half reliability values that can be calculated for a measurement tool is identical to the Cronbach's alpha internal consistency coefficient if two halves are equal (Warrens, 2015). In this sense, while Cronbach's alpha has already been reported, the calculation of the split-half reliability does not provide more information about the internal consistency of the measurements. Reporting Cronbach's alpha and split-half reliability together shows that this situation can be ignored by the researchers.

According to the examinations on item analysis, it was understood that in about half of the scale development and adaptation studies, no analysis for item discrimination was included. The fact that item discrimination indices are generally parallel to factor loads obtained from factor analysis can be considered as the reason for this situation. However, it is important not to be satisfied with factor analysis and to calculate item discrimination in order to provide more evidence about the validity of the items. In the study, coding was done by a single researcher. Although the researcher performed the coding twice in different time intervals for reliability, the fact that the coding was done by a single researcher can be seen as a limitation of this research. In this sense, the codings should be done at least by two researchers for supporting the reliability of the results in such studies with two or more authors.

Declaration of Conflicting Interests and Ethics

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

Authorship Contribution Statement

Melehat Gezer: Investigation, Resources, Visualization, Formal Analysis, Writing original draft, Methodology, and Validation.

Orcid

Melehat Gezer https://orcid.org/0000-0001-7701-3203

REFERENCES

Acar Güvendir, M., & Özer Özkan, Y. (2015). Türkiye'deki eğitim alanında yayımlanan bilimsel dergilerde ölçek geliştirme ve uyarlama konulu makalelerin incelenmesi [The examination of scale development and scale adaptation articles published in Turkish academic journals on education]. *Electronic Journal of Social Sciences, 14*(52), 23–33. https://doiorg/1017755/esosder54872

Almanasreh, E., Moles, R., & Chen, T.F. (2019). Evaluation of methods used for estimating content validity. *Research in Social and Administrative Pharmacy, 15*(2), 214–221. https://doiorg/101016/jsapharm201803066

Beseler, C.L., Jones, L.M., & Mitchell, K.J. (2021). Measuring online prosocial behaviors in primary school children: Psychometric properties of the online civility scale. *Contemporary School Psychology.* https://doiorg/101007/s40688-021-00401-5

Boztunç Öztürk, N., & Şahin, M.G. (2021). Bilimsel araştırmaya temel oluşturan ölçme kavramları [Measurement concepts that form the basis of scientific research]. In B. Çetin, M. İlhan, & M. G. Şahin, (Eds.), *Eğitimde araştırma yöntemleri [Research methods in education],* (pp 1–34). Pegem.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research Psychology, 3*(2), 77-101. Retrieved from https://wwwtandfonlinecom/doi/abs/101191/1478088706qp063oa

Cattell, R.B. (1978). *The scientific use of factor analysis in behavioral and life sciences.* Plenum

Chandu V.C., Marella Y., Panga, G.S., Pachava, S., & Vadapalli, V. (2020). Measuring the impact of COVID-19 on mental health: A scoping review of the existing scales. *Indian J Psychol Med, 42*(5), 421–427. https://doiorg/101177%2F0253717620946439

Cramer, D. (2003). *Advanced Quantitative Data Analysis.* McGraw Hill Education.

Çalık, M., Ayas, A., & Ebenezer, J. V. (2005). A review of solution chemistry studies: Insights into students' conceptions. *Journal of Science Education and Technology, 14*(1), 29–50. Retrieved from: https://linkspringercom/article/101007/s10956-005-2732-3

Çalık, M., & Sözbilir, M. (2014). Parameters of content analysis. *Education and Science, 39*(174), 33–38. https://doiorg/1015390/EB20143412

Çelik, D., & Yüksel, G. (2020). Müzik eğitimi kapsamında yapılan ölçek geliştirme çalışmalarının çok yönlü incelenmesi [Versatile analysis of scale development studies conducted within the scope of music education]. *Journal of the Human and Social Science Researches, 9*(5), 4059–4087. https://doiorg/1015869/itobiad793488

Çermik, E., & Akçay, B. (2020). Çevresel vatandaşlık bilgi testinin geliştirilmesi ve ortaokul öğrencilerinin bilgi düzeylerinin belirlenmesi [Developing environmental citizenship knowledge test and determining the knowledge levels of secondary school students]. *Turkish Studies - Education, 15*(2), 731-750. https://dxdoiorg/1029228/TurkishStudies42112

Çüm, S., & Koç, N. (2013). Türkiye'de psikoloji ve eğitim bilimleri dergilerinde yayımlanan ölçek geliştirme ve uyarlama çalışmalarının incelenmesi [The review of scale development and adaptation studies which have been published in psychology and educatıon journals in Turkey]. *Education Science and Practice, 12*(24), 115–135.

Delice, A., & Ergene, Ö. (2015). Investigation of scale development and adaptation studies: an example of mathematics education articles. *Karaelmas Journal of Educational Sciences, 3*(1), 60–75. Retrieved from https://dergiparkorgtr/en/download/article-file/2160898

Erdem, C., & Koçyiğit, M. (2019, April 25–28). *Adapting the digital citizenship scale to Turkish: validity and reliability study [Conference presentation abstract].* 28th International Conference on Educational Sciences (ICES 2019), Ankara, Turkey.

Ergene, Ö. (2020). Matematik eğitimi alanında ölçek geliştirme ve ölçek uyarlama makaleleri: betimsel içerik analizi [Scale development and adaptation articles in the field of mathematics education: descriptive content analysis]. *Journal of Education for Life, 34*(2), 360–383. https://doiorg/1033308/266748742020342207

Eurydice (European Commission/EACEA/) (2012). *Citizenship education in Europe.* Author.

Eurydice (European Commission/EACEA/) (2017). *Citizenship education at school in Europe – 2017 Eurydice Report*. Publications Office of the European Union.

Gezer, M. (2020). Ortaokul öğrencilerinin perspektifinden iyi insan iyi vatandaş [Good human good citizen from the perspective of secondary school students]. *Çukurova University Faculty of Education Journal*, *49*(2), 995–1024. https://doiorg/1014812/cufej673422

Hadjichambis, A.C., & Paraskeva-Hadjichambi, D. (2021). Environmental citizenship questionnaire (ECQ): The development and validation of an evaluation instrument for secondary school students. *Sustainability*, *12*(3), 821-833. https://doiorg/103390/su12030821

Homer, S. T. (2020, Jul 4-5). *Perceived corporate citizenship: A scale development and validation study adopting a bottom-up approach [Conference full text].* 2 International European Conference on Interdisciplinary Scientific Researches Full Text Book, pp 431–440.

İşman, A., & Güngören, Ö. C. (2014). Dijital vatandaşlık [Digital citizenship]. *TOJET: The Turkish Online Journal of Educational Technology, 13*(1), 73–77. Retrieved from http://wwwtojetnet/articles/v13i1/1317pdf

Kalaycıoğlu, S., & Çelik, K. (2008). Genç insanın vatandaş olma ve tanınma hakkı [Young Person's Right to Become Citizen and to be Recognized]. *İnsan Hakları Yıllığı*, 26, 41–57. Retrieved from https://dergiparkorgtr/tr/pub/ihy/issue/61998/928117

Kanten, P., & Arda, B. (2020). Yönetim ve örgütsel davranış yazınındaki ölçek geliştirme çalışmalarının metodolojik açıdan analizi [The methodological analysis of scale development studies in management and organizational behavior fields]. *Business and Economics Research Journal*, *11*(2), 581–590. https://doiorg/1020409/berj2020269

Karatekin, K., & Uysal, C. (2018). Ekolojik vatandaşlık ölçeği geliştirme çalışması [Ecological citizenship scale development study]. *International Electronic Journal of Environmental Education, 8*(2), 82–104.

Kaya Uyanık, G., Güler, N., Taşdelen Teker, G., & Demir, S. (2017). Türkiye'de eğitim alanında yayımlanan ölçek geliştirme çalışmalarının uygunluğunun çok yüzeyli rasch modeli ile incelenmesi [Investigation of scale development studies conducted in educational sciences published in Turkey by many-faceted rasch model]. *Journal of Measurement and Evaluation in Education and Psychology, 8*(2), 183–199. https://doiorg/1021031/epod291367

Kerr, D. (1999). Citizenship education in the curriculum: An international review. *The School Field: International Journal of Theory and Research in Education, 10*(3/4), 5–31. Retrieved from http://citeseerxistpsuedu/viewdoc/summary?doi=10115852377

Kim, M., & Choi, D. (2018). Development of youth digital citizenship scale and ımplication for educational setting. *Educational Technology & Society, 21*(1), 155–171. Retrieved from https://wwwjstororg/stable/26273877?seq=1#metadata_info_tab_contents

Lo, K.W.K., Kwan, K.P., Ngai, G., & Chan, S.C.F. (2019, January 10–12). *Cross-cultural validation of the global citizenship scale for measuring impacts of international service-learning in Hong Kong setting [Paper presentation].* The 3rd International Conference on Service-Learning, Hong Kong.

Loobuyck, P. (2021). The policy shift towards citizenship education in Flanders How can it be explained?. *Journal of Currıculum Studies*, *53*(1), 65-82. https://doiorg/101080/0022027220201820081

Nowell, L.S., Norris, J.M., White, D.M., & Moules, N.J. (2017). Thematic Analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods, 16*, 1–13. https://doiorg/101177%2F1609406917733847

Özel, S. (2007). *Küreselleşme döneminde vatandaşlık.* Retrieved from http://wwwanayasagovtr/files/pdf/anayasa_yargisi/soli_ozelpdf

Schugurensky, D. (2005). *Citizenship and citizenship education: Canada in an international context*. Text prepared as input for discussion for the course AEC3131 ON.

Şahin, İ.F., & Çermik, F. (2014). Küresel vatandaşlık ölçeğinin Türkçeye uyarlanması: Güvenirlik ve geçerlik çalışması [Turkish adaptation of global citizenship scale: Reliability and validity]. *Eastern Geographical Review,* 31, 207–218. https://doiorg/1017295/dcd30443

Şahin, M.G., & Boztunç Öztürk, N. (2018). Eğitim alanında ölçek geliştirme süreci: Bir içerik analizi çalışması [Scale development process in educational field: a content analysis research]. *Kastamonu Education Journal, 26*(1), 191-199. https://doi:1024106/kefdergi375863

Şahin, M., Şahin, S., & Göğebakan Yıldız, D. (2016). Sosyal bilgiler eğitimi programı ve dünya vatandaşlığı: Öğretmen adaylarının perspektifinden [The curriculum of social studies education and World citizenship: From perspective of prospective teachers]. *Hacettepe University Journal of Education, 31*(2), 369-390. http://dxdoiorg/1016986/HUJE2016015386

Taşdelen Teker, G., & Güler, N. (2019). Thematic content analysis of studies using generalizability theory. *International Journal of Assessment Tools in Education, 6*(2), 279–299. http://dxdoiorg/1021449/ijate569996

Tavşancıl, E., Güler, G., & Ayan, C. (2014, Jun 9–13). *Review of attitude scales developed in Turkey between 2002 and 2012 regarding scale developing process [Conference presentation abstract]*. 4th Congress on Measurement and Evaluation in Education and Psychology (EPOD 2014), Ankara, Turkey.

Tosun, C., & Taşkesenligil, Y. (2014, Sep 11–14). *Document analysis of scales and achievement tests developed/adapted in the field of science education in Turkey [Conference presentation abstract]*. XI National Science and Mathematics Education Congress, Adana, Turkey.

Warrens, M.J. (2015). On Cronbach's alpha as the mean of all split-half reliabilities. *Quantitative Psychology Research*, *89*, 293-300. https://doiorg/101007/978-3-319-07503-7_18

Ültay, N., & Çalık, M. (2012). A thematic review of studies into the effectiveness of context-based chemistry curricula. *Journal of Science Education and Technology, 26*(6), 686–701. https://doiorg/101007/s10956-011-9357-5

Yazıcı, S., Arslan, H., Çetin, E., & Dil, K. (2017). Aktif yurttaşlık ölçme aracının geliştirilmesi üzerine bir çalışma [A study on the development of active citizen questionnaire]. *International Periodical for the Languages, Literature and History of Turkish or Turkic, 12*(13), 1–22. http://dxdoiorg/107827/TurkishStudies11645

Yurdabakan, İ., & Çüm, S. (2017). Davranış bilimlerinde ölçek geliştirme (Açıklayıcı Faktör Analizine Dayalı) [Scale development in behavioral sciences (Based on exploratory factor analysis)]. *Turkish Journal of Family Medicine and Primary Care*, *11*(2), 108–126. https://doiorg/1021763/tjfmpc317880

Yıldırım, C. (2018). *Ortaöğretim öğrencilerinin demokratik vatandaşlık tutumlarının resmi ve örtük program açısından incelenmesi [The examination of the secondary education students' democratic citizenship attitudes towards the formal and hidden curriculum]* [Doctoral thesis]. Adnan Menderes University.

Yıldırım, A., & Şimşek, H. (2011). *Sosyal bilimlerde nitel araştırma yöntemleri [Qualitative research methods in the social sciences]*. Seçkin.

APPENDIX

Appendix-1: *List of studies included in the research*

1. Acun, İ, Demir, M., & Göz, N.L. (2010). Öğretmen adaylarının vatandaşlık yeterlilikleri ile eleştirel düşünme becerileri arasındaki ilişki [The relationship between student teachers' citizenship skills and critical thinking skills]. *Journal of Social Studies Education Research*, *1*(1), 107–123. Retrieved from https://jsserorg/indexphp/jsser/article/view/119
2. Akın, A., Sarıçam, H., Akın, Ü., Yıldız, B., Demir, T., & Kaya, M. (2014, April 28-30). *The validity and reliability of the Turkish version of the global citizenship scale [Conference Presentation Abstract]*. III International Social Studies Education Symposium (Isses III 2014), Ankara, Turkey.
3. Altıntaş, İ.N. (2016). *Sosyal bilgiler öğretmen adaylarının aktif vatandaşlık kazanımları: eylem araştırması [Active citizenship goals of social studies teacher candidates: Action research]* [Doctoral thesis]. Gazi University.
4. Altıok, A. (2019). *Sosyal bilgilerde hizmet ederek öğrenmenin öğrencilerin iyi vatandaşlık algı, bilgi ve tutumlarına etkisi ve öğretmen görüşleri [The effect of the service-learning in social studies to the students' perception of good citizenship, their knowledge and attitudes and teachers' views]* [Doctoral thesis]. Gazi University.
5. Arslan, H., Dil, K., Çetin, E., & Yazıcı, S. (2017). Aktif yurttaşlık öz-yeterlik ölçeği: Bir geçerlik ve güvenirlik çalışması [Active citizenship self-efficacy scale: A reliability and validity study]. *Journal of Human Sciences*, *14*(3), 2797-2809. http://dxdoiorg/1014687/jhsv14i34771
6. Arslan, S. (2014). *Çokkültürlü toplumlarda vatandaşlık eğitimine yönelik öğretmen ve öğrenci düşüncelerinin incelenmesi [Examination of teacher and student opinions on citizenship education in multicultural societies]* [Doctoral thesis]. Marmara University.
7. Balbağ, N.L. (2016). *İlkokul sosyal bilgiler dersi bağlamında öğrenci ve öğretmenlerin küresel vatandaşlık algıları [Elementary teachers' and students' perceptions of global citizenship in the social studies course]* [Doctoral thesis]. Anadolu University.
8. Baştürk, N. (2011). *İlköğretim 8 sınıf vatandaşlık ve demokrasi eğitimi dersi öğretim programı kazanımlarının öğrenci görüşlerine göre değerlendirilmesi (Konya ili örneği) [Evaluation of the elementary 8th grade citizenship and democracy education course learning outcomes according to student views (Case of Konya)]* [Doctoral thesis]. Atatürk University.
9. Bozbek, M., & Demir, S.B. (2014). Vatandaşlık ve demokrasi eğitimi dersine yönelik tutum ölçeği; Geçerlik ve güvenirlik çalışması [Attitude scale towards citizenship and democracy education lesson: Development, validity and reliability study]. *Dicle University Journal of Ziya Gökalp Education Faculty, 23*, 323–351.
10. Çevik Kansu, C. (2014). *İlkokul 4 sınıf öğrencilerinde etkin vatandaşlık eğitiminin etkililiği [Efficiency of Active Citizenship Education on 4th Grade of Primary Students]* [Doctoral thesis]. Ondokuz Mayıs University.
11. Çiçek, S. (2018). *Sosyal bilgiler öğretmen adaylarının iyi vatandaşlık algılarının incelenmesi [Examining the perceptions of the social studies teacher candidates towards good citizenship]* [Master's thesis]. Akdeniz University.
12. Demirbaş, İ. (2016). *Üniversite öğrencilerinin vatandaşlık algısının belirlenmesi [The determination of the university students' citizenship perception]* [Master's thesis]. Kastamonu University.
13. Doğanay, A. (2009, May 28–30). *Evaluation of pre-service teachers' perception of citizenship and their actions in the context of political socialization [Conference Presentation Abstract]*. 1st International European Union, Democracy, Citizenship and Citizenship Education Symposium, Uşak, Turkey.

14. Durualp, E. (2016). Ortaokul öğrencilerinin vatandaşlık algılarının bazı sosyolojik değişkenler açısından incelenmesi *[Investigation of citizenship perception of middle school students from the point of some sociological variables]* [Doctoral thesis]. Ankara University.
15. Elçi, A.C., & Sarı, M. (2016). Bilişim Teknolojileri ve Yazılım dersinde dijital vatandaşlık Bir ölçek geliştirme çalışması [Digital citizenship in the Information Technology and Software course: A scale development study]. *Journal of Human Sciences, 13*(2), 3602–3613. http://dxdoiorg/1014687/jhsv13i23838
16. Erdem, C., & Koçyiğit, M. (2019, April 25–28). *Adapting the digital citizenship scale to Turkish: validity and reliability study [Conference presentation abstract]*. 28th International Conference on Educational Sciences (ICES 2019), Ankara, Turkey.
17. Göl, E. (2013). *Sosyal bilgiler öğretmen adaylarının küresel vatandaşlık tutum düzeylerinin farklı değişkenler açısından incelenmesi [The examination of global citizenship attitude levels of social studies nominee instructors according to different variants]* [Master's thesis]. Ahi Evran University.
18. Gürbüz, G. (2006). *İlköğretim 7 ve 8 sınıflarda vatandaşlık bilgisi dersinde demokrasi eğitimi [Democracy education of elementary school at 7th and 8th classes at citizenship lesson]* [Doctoral thesis]. Abant İzzet Baysal University.
19. İçen, M., Öztürk, C., & Yılmaz, A. (2017). Vatandaşlık duygusu ölçeği güvenirlik ve geçerlik çalışması [Validity and reliability of the sense of citizenship scale]. *International Journal of Field Education, 3*(2), 26–36. https://doiorg/1032570/ijofe370382
20. İkinci, İ. (2016). *Sosyal bilgiler öğretmenlerinin vatandaşlık algıları ve vatandaşlık eğitimi ile ilgili düşüncelerinin incelenmesi [Investigation of social studies teachers' perceptions of citizenship and citizenship education]* [Doctoral thesis]. Dumlupınar University.
21. İşman, A., & Güngören, Ö.C. (2014). Dijital Vatandaşlık [*Digital citizenship*]. *TOJET: The Turkish Online Journal of Educational Technology*, *13*(1), 73–77.
22. Karaduman, H. (2011). *6 sınıf sosyal bilgiler dersinde dijital vatandaşlığa dayalı etkinliklerin öğrencilerin dijital ortamdaki tutumlarına etkisi ve öğrenme öğretme sürecine yansımaları [The effects of digital citizenship based activities on students' attitudes in digital environments and reflections to learning teaching process in the 6th grade social studies course]* [Doctoral thesis]. Marmara University.
23. Karatekin, K., & Uysal, C. (2018). Ekolojik vatandaşlık ölçeği geliştirme çalışması [Ecological citizenship scale development study]. *International Electronic Journal of Environmental Education, 8*(2), 82–104.
24. Karışan, D., & Yılmaz Tüzün, Ö. (2017). Dünya vatandaşlığı için karakter ve değerler ölçeğinin Türkçe'ye uyarlanması: Geçerlik ve güvenirlik çalışması [Adaptation of character and values as global citizen assessment questionnaire into Turkish: Validity and reliability study]. *Pamukkale University Journal of Education*, *42*, 74-85. http://dxdoiorg/109779/PUJE823
25. Kaya, B. (2013). *Sosyal bilgiler öğretmen adaylarının vatandaşlık algıları ile politik ilgi ve katılımları arasındaki ilişkinin incelenmesi [Examining the relationship between citizenship perceptions of social studies teacher candidates and their political interests and participation]* [Doctoral thesis]. Marmara University.
26. Kaya, B., & Ersoy, A. F. (2014). Vatandaşlık ve demokrasi eğitimi dersinin sekizinci sınıf öğrencilerinde vatandaşlık algısının oluşmasıyla ilişkisi [The effect of citizenship and democracy on eight class students growing as a conscious citizen]. *Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi Dergisi, 23*, 252–303.
27. Kılınç, E., & Dere, İ. (2013). Lise öğrencilerinin 'iyi vatandaş' kavramı hakkındaki görüşleri [High school students' perception of the concept of 'good citizen]. *Journal of Social Studies Education Research*, *4*(2), 103–124.

28. Kılınç, H.H. (2013). *8 sınıf vatandaşlık ve demokrasi eğitimi dersi kazanımlarının gerçekleşme düzeyine ilişkin öğretmen-öğrenci görüşleri ile derse yönelik öğrenci tutumları [Teachers' and students' perceptions of realization level of 8th grade citizenship and democracy education course learning attainments and students' attitudes towards the course]* [Doctoral thesis]. Fırat University.
29. Kocadağ, T. (2012). *Öğretmen adaylarının dijital vatandaşlık düzeylerinin belirlenmesi [Determining the digital citizenship levels of prospective teachers]* [Master's thesis]. Karadeniz Teknik University.
30. Kuş, Z., Güneş, E., Başarmak, U. & Yakar, H. (2017). Gençlere yönelik dijital vatandaşlık ölçeğinin geliştirilmesi: Geçerlik ve güvenirlik çalışması [Development of a digital citizenship scale for youth: A validity and reliability study]. *Journal of Computer and Education Research, 5(*10), 298–316. https://doiorg/1018009/jcer335806
31. Malkoç, S. (2020). *Sosyal bilgiler öğretmenlerinin vatandaşlık tiplerinin belirlenmesi [Determination of the citizenship types of social studies teachers]* [Doctoral thesis]. Gazi University.
32. Öntaş, T., Çoban, O., & Atmaca, T. (2020). Adaptation of civic education ideologies scale to the Turk culture: It's reliability and validity [Vatandaşlık eğitimi ideolojileri ölçeğinin Türk kültürüne uyarlanması: Güvenirlik ve geçerlik çalışması]. *International Journal of Social Science Research, 9*(1), 1-20.
33. Özdemir Özden, D. (2011). *İlköğretim okullarında çevresel vatandaşlık eğitimi [Environmental citizenship education in primary schools]* [Doctoral thesis]. Marmara University.
34. Sabancı, O. (2008). *İlköğretim 7 sınıf öğrencilerinin sosyal bilgiler dersinde yer alan vatandaşlık konularıyla ilgili kavramsal anlamaları [Elementary school 7th grade students' conceptual understandings related to the citizenship subjects taking place in social studies course]* [Doctoral thesis]. Gazi University.
35. Sağlam, H. (2011). Öğretmen adaylarının etkili vatandaşlık yeterlik düzeyleri [Proficiency levels of student teachers effective citizenship]. *Kastamonu Education Journal, 19*(1), 39–50.
36. Som Vural, S. (2016). *Üniversite öğrencilerinin bakış açısıyla dijital vatandaşlık göstergelerinin incelenmesi [Investigation of digital citizenship indicators through university students' perceptions]* [Doctoral thesis]. Anadolu University.
37. Şahin, İ.F., & Çermik, F. (2014). Küresel vatandaşlık ölçeğinin Türkçeye uyarlanması: Güvenirlik ve geçerlik çalışması [Turkish adaptation of global citizenship scale: Reliability and validity]. *Eastern Geographical Review,* 31, 207-218.
38. Şahin, M., Şahin, S., & Göğebakan Yıldız, D. (2016). Sosyal bilgiler eğitimi programı ve dünya vatandaşlığı: Öğretmen adaylarının perspektifinden [The curriculum of social studies education and World citizenship: From perspective of prospective teachers]. *Hacettepe University Journal of Education, 31*(2), 369-390. http://dxdoiorg/1016986/HUJE2016015386
39. Tarhan, E. (2019). *Vatandaşlık bağlamında sosyal bilgiler öğretmen adaylarının vatanseverlik tutumları ve görüşleri [Patriotism attitudes and views of preservice social studies teachers within the context of citizenship]* [Master's thesis]. Bolu Abant İzzet Baysal University.
40. Tonga, D. (2013). *8 sınıf öğrencilerinin vatandaşlık bilinci düzeylerinin çeşitli değişkenler açısından değerlendirilmesi [Evaluation of the levels of citizenship consciousness of students grade 8 in terms of several variables]* [Doctoral thesis]. Gazi University.
41. Türküresin, K. (2019). *Ortaokul öğretmenlerinin dijital vatandaşlık davranışlarının incelenmesi [Investigation of digital citizenship behaviors of middle school teachers]* [Master's thesis]. Dumlupınar University.

42. Utku, M. (2015). *Üniversite öğrencilerinin vatandaşlık ve sosyal medya (sosyal ağ) vatandaşlık algısının çeşitli değişkenlere göre incelenmesi [Examining perceptions of citizenship and perceptions of citizenship to social media on students of university according to various variables]* [Doctoral thesis]. Erzincan University.
43. Ünal, F. (2019). *Developing the citizenship skills scale: a validity and reliability study [Conference Presentation Abstract].* Ejer Congress Conference Proceedings, pp 442–446, Anı.
44. Ünal, F. (2019). *Ortaokul 8 sınıf öğrencilerinin hak, sorumluluk ve katılımcılık bağlamında vatandaşlık bilincine ilişkin görüşlerinin hammond modeliyle değerlendirilmesi [The evaluation of the views of secondary school 8th year student on citizenship consciousness in the context of right, responsibility and participation using hammond's model]* [Doctoral thesis]. Bartın University.
45. Ünal, F. (2019, April 25–28). *Developing citizenship knowledge scale: Validity and reliability study [Conference Presentation Abstract]* XII International Educational Research Congress, Rize, Turkey.
46. Yazgan, A.D. (2013). *Öğretmen adaylarının medya okuryazarlık düzeyleri ile aktif vatandaşlığa ilişkin demokratik değer düzeyleri arasındaki ilişki [The relationship between preservice teachers' levels of media literacy and their democratic value for active citizenship]* [Doctoral thesis]. Çanakkale Onsekiz Mart University.
47. Yazıcı, S., Arslan, H., Çetin, E., & Dil, K. (2017). Aktif yurttaşlık ölçme aracının geliştirilmesi üzerine bir çalışma [A study on the development of active citizen questionnaire]. *International Periodical for the Languages, Literature and History of Turkish or Turkic, 12*(13), 1–22. http://dxdoiorg/107827/TurkishStudies11645
48. Yıldırım, C. (2018). *Ortaöğretim öğrencilerinin demokratik vatandaşlık tutumlarının resmi ve örtük program açısından incelenmesi [The examınatıon of the secondary educatıon students' democratıc cıtızenshıp attıtudes towards the formal and hıdden currıculum]* [Doctoral thesis]. Adnan Menderes University.
49. Yıldırım, Y., & Çalışkan, H. (2020). Ortaokul öğrencileri için etkin vatandaşlık değerleri ölçeğinin (evdö) geliştirilmesi [Developing effective citizenship values scale (ecvs) for secondary school students]. *Millî Eğitim, 49*(228), 335–364. Retrieved from https://doiorg/1037669/milliegitim742091
50. Yiğen, V. (2019). *Etkili vatandaş yetiştirmede sosyal bilgiler dersinin rolü [The role of social studies in raising effective citizenship]* [Master's thesis]. İnönü University.

Appendix-2: *The checklist used in the study*

Marking tag	No	....
	Title	....
	Type	1) Conference presentation 2) Article 3) Thesis (Master thesis) 4) Thesis (Doctoral thesis)
	Authors	....
	Journal/University	....
	Year of publication	....
	Key words	....
Theoretical knowledge	The subject matter of the scale	....
	Scale development/adaptation	1) Scale development 2) Scale adaptation
	Types of sampling	1) Primary school 2) Secondary school 3)High school 4) University 5) Prospective teachers 6) Teachers of social studies
	Sample size	1) 100–200 2) 201–300 3) 301–400 4) 401–500 5) 501–600 6) 601-700 7) 701-800 8) 801–900 9) 901–1000 10) 1001–2000 11) 2000 and later
	Validity evidence	1) Construct validity a Factor analysis a1 Only confirmatory factor analysis (CFA) a2 Only exploratory factor analysis (EFA) a3 Both CFA and EFA b Convergent-divergent validity c Known groups validity d Criterion-related validity e Measurement invariance (multi-group CFA) 2) Content validity a Expert opinion b Content validity rate index 3) Unreported
	Item analysis	1) Lower-upper group -comparison 2) Item-test correlations
	Reliability evidence	1) Cronbach's Alpha 2) Test-retest method 3) Parallel forms method 4) Split half reliability (Spearman Brown and Guttman) 5) Unreported
	Number of factors	....

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 1, 254–266

https://doi.org/10.21449/ijate.948182

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate **_Research Article_**

Item parameter recovery via traditional 2PL, Testlet and Bi-factor models for Testlet-Based tests

Sumeyra Soysal[1,*], Esin Yilmaz Kogar[2]

[1]Necmettin Erbakan University, Ahmet Keleşoğlu Faculty of Education, Department of Educational Sciences, Educational Measurement and Evaluation, Konya, Turkiye
[2]Niğde Ömer Halisdemir University, Faculty of Education, Department of Educational Sciences, Educational, Measurement and Evaluation, Niğde, Turkiye

ARTICLE HISTORY

Received: June 04, 2021

Revised: Jan. 22, 2022

Accepted: Feb. 03, 2022

Keywords:
Local dependence,
Testlet model,
Bi-factor model,
Random effect parameter,
Parameter recovery.

Abstract: The testlet comprises a set of items based on a common stimulus. When the testlet is used in the tests, there may violate the local independence assumption, and in this case, it would not be appropriate to use traditional item response theory models in the tests in which the testlet is included. When the testlet is discussed, one of the most frequently used models is the testlet response theory (TRT) model. In addition, the bi-factor model and traditional 2PL models are also used for testlet-based tests. This study aims to examine the item parameters estimated by these three calibration models of the data properties produced under different conditions and to compare the performances of the models. For this purpose, data were generated under three conditions: sample size (500, 1000, and 2000), testlet variance (.25, .50, and 1), and testlet size (4 and 10). For each simulation condition, the number of items in the test was fixed at $i = 40$ and 100 replications were made under each condition. Among these models, it was concluded that the TRT model gave less biased results than the other two models, but the results of the bi-factor model and the TRT were more similar as the sample size increased. Among the examined conditions, it was determined that the most effective variable in parameter recovery was the sample size.

1. INTRODUCTION

Item response theory (IRT, Lord & Novick, 1968) is a model that is widely used for test development and test scoring, because of its strong mathematical modeling. One of the important assumptions of this theory is local independence (LI). This assumption is generally expressed as “the examinee's trait (ability or proficiency) value, denoted provides all the necessary information about the examinee's performance, and once trait level is considered, all other factors affecting examinee performance are random” (Wainer et al., 2000, p.248). However, this assumption can be violated when the items share a common stimulus. In the literature, such items are generally referred to as testlets.

The concept of testlet, first expressed by Wainer and Kiely (1987), is the name given to a group of items associated with a single comprehensive stimulus. The testlet shows a set of items that

*CONTACT: Sumeyra Soysal ✉ sumeyrasoysal@hotmail.com Necmettin Erbakan University, Ahmet Keleşoğlu Faculty of Education, Department of Educational Sciences, Educational Measurement and Evaluation, Konya, Turkiye

share a single common stimulus, such as a reading passage or an information graph, and where performance on each item depends on both a general ability and a specific ability related to a specific content or situation (Li, 2017). Such items are widely used in many national and international large-scale tests because of their various advantages. For example, testlets allow over one item to be asked based on the same stimulus, allowing over one information to be collected from a stimulus, thus improving the efficiency of the test (information per unit time) (Wainer et al., 2000). Another advantage of testlets is that they help develop a more realistic and context-based test (Li, 2017). Through these context-dependent items, measuring higher-level skills may become more workable (DeMars, 2006). It is known that testlets provide a significant advantage in computer adaptive test (CAT) applications. In CAT applications, there is a specific item selection algorithm for each person. Here, there is a context-effect caused by the content of the items. However, this effect is reduced, as individuals will encounter the same context when they take the same testlet. In short, the use of testlets in CAT applications provides greater control of the negative effects of single items, allowing as much fairness as possible among test takers (Pak, 2017). However, in such items, some students have a special interest or better prior background knowledge in a passage than other students, in this situation they are likely to perform better on the items related to this passage than on other items of the same difficulty level, or they tend to perform better than other students with the same general ability level (Li, 2017, p.1). Therefore, testlets lead to the emergence of additional sources of variance, such as content knowledge (Chen & Thissen, 1997). DeMars (2006) states that responses to items in a testlet may be related to testlet-specific background knowledge or skills, or to a secondary characteristic, such as testlet-specific interest or other motivational factors. This situation has revealed the necessity of a special examination of testlet items.

Another disadvantage of testlets is that testlets violate the LI assumption of the unidimensional IRT. Although this assumption is violated, the use of unidimensional IRT models for such items leads to inaccurate in parameter estimations (Sireci et al., 1991; Wainer & Wang, 2000; Yen, 1993). Therefore, different models have been developed to handle testlets. The psychometric framework that deals with testlets is known as testlet response theory (TRT) models (Bradlow et al., 1999; Wainer et al., 2000; Wainer et al., 2007). This model includes one more parameter explaining the interaction between each item and each examinee within a testlet, besides the parameters in the traditional IRT model for dichotomous items. Another solution to model the dependency among test items in testlets is the bi-factor model (Rijmen, 2010). Recently, multilevel models have also been used to address local item dependence among items (Jiao et al., 2005; Jiao et al., 2012). These models consider local item dependence because of item clustering. In addition, although there are testlets, it is very common to apply the traditional IRT model to items that are scored dichotomously. Because when the testlet effect is determined to be low, the negligibility of this effect or the usability of traditional IRT estimations, which are more familiar to researchers, are discussed in the literature (Glas et al., 2000; Eckes, 2014; Eckes & Baghaei, 2015; Min & He, 2014). It has also been examined with polytohomus IRT models that treat testlets as a single item, and it has been stated that there is a need for models that give more information about testlets (Wainer, 1995).

Since unidimensional, testlet and bi-factor models are widely used in testlet examinations, it is important to evaluate whether the parameters estimated from these models can be accurately estimated. Since all conditions cannot be tested on the real data set, this study was carried out on simulation data. The advantage of knowing the real parameter values in simulation studies makes the accuracy of the estimation method is measurable. This study, it is aimed to examine the data produced under different testlet conditions with traditional two-parameter logistic (2PL), the TRT, and the bi-factor models by varying the sample size, the size of the testlet variance, and the number of testlets. Koizol (2016) stated that the bi-factor model did not receive enough attention in testlet reviews. Liu and Liu (2012) stated that it is not clear to practitioners in which cases traditional IRT models should be used instead of a newly proposed testlet model. In this study, it is aimed to provide more helpful information to practitioners by considering many possible conditions. It is expected that this study, which also includes the bi-factor model in testlet examinations, will contribute to filling the gaps in this subject. Because of this study, determining the conditions under which local item dependency has serious effects on parameter estimations with the help of many conditions tested can guide the researchers in choosing the right model.

1.1. Calibration Models

There are strategies developed over different models to deal with the local independence assumption violation caused by testlets. Traditionally, the items in the testlets have been treated as independent items like other items in the test, and traditional IRT models have been used as the calibration model. The traditional 3PL model for dichotomous data is specified as

$$P_{ni}\,(1) = c_i + \,(1 - \,c_i)\,\frac{\exp[a_i\,(\theta_n - \,b_i\,)]}{1 + \exp[a_i\,(\theta_n - \,b_i\,)]}, \tag{1}$$

where $P_{ni}(1)$ is the probability of response 1 (correct) for person n on the ith item; a_i, b_i, and c_i are the discrimination, difficulty, and guessing parameters, respectively; and θ_n is person's ability. However, this approach has been found to cause biased parameter estimation and overestimation of test reliability (Sireci et al., 1991; Thissen et al., 1989; Tuerlinckx & De Boeck, 2001; Wainer & Thissen, 1996; Wainer & Wang, 2000; Yen, 1993).

In another approach, the testlets were evaluated as a single item and scored in polytomous. Although this approach has been found to yield partially good results (Wainer, 1995), it has several shortcomings. For example, this approach is insufficient for situations that require more information about the items in the tests. Because of this approach, the testlet score is represented by the sum of the correct number (Wainer et al., 2000). These total scores lose answer pattern knowledge for each individual test taker. This loss of information can lead to an increase in measurement errors, which directly reduces overall test reliability (Keller et al., 2003; Sireci et al., 1991; Yen, 1993; Zenisky et al., 2002). Since polytomous models were not used in this study, the details of the model were not included.

Although it is stated that the violation of local independence does not cause serious problems when the length of the testlets is moderate (4-6 items/testlet), it is stated that as the testlets get longer, a special psychometric model is needed that can control local dependence (Wainer et al., 2007). In addition, in these models, attention should be paid to maintaining the item level as the unit of analysis. Bradlow et al. (1999) proposed a TRT model by adding a parameter (a testlet effect parameter) to the traditional 2PL model for items nested in the same testlet. This parameter represents the dependence between items within the same testlet, and the variances of the random testlet effects were assumed to be constant across testlets (Wang & Wilson, 2005). Later, this model was developed for 3PL (Wainer et al., 2000). The model is

$$P_{ni}\,(1) = c_i + \,(1 - \,c_i)\,\frac{\exp\left[a_i\left(\theta_n - \,b_i\, - \,\gamma_{nd(i)}\right)\right]}{1 + \exp\left[a_i\left(\theta_n - \,b_i\, - \,\gamma_{nd(i)}\right)\right]}, \tag{2}$$

As seen, in this model, unlike Equation 1, there is a $\gamma_{nd(i)}$ parameter. $\gamma_{nd(i)}$ is the random effect for person n on testlet $d_{(i)}$, which describes the interaction between persons and items within the testlet. The testlet effect is a random effect variance caused by local item dependency (LID), and the greater the variance, the greater the effect in the testlet (Wainer & Wang, 2000). The size of the variance shows the size of the dependence between the items. When the testlet random effect is zero, this model is equal to the traditional 3PL model.

The advantages of this model compared to polytomous models are expressed as follows (Wang et al., 2000; Wang & Wilson, 2005). The unit of analysis is the test items, not the testlets, so that the information in the response patterns within the testlets is not lost. The second advantage is that familiar item parameter concepts, such as item discrimination and item difficulty, are still valid and functional. Another advantage is that the standard item scoring scales (1 for a correct answer and 0 for an incorrect answer) remain unchanged. Thus, an easy transfer from the traditional IRT model to using the TRT model is provided.

This model uses the same item discrimination for both the theta and testlet traits, and this has been discussed as a limitation of the model (Li et al., 2004). When the data do not fit this constraint, the model is misspecified (Koziol, 2016). To handle discrimination parameters for these traits separately, the responses to the testlet items can be handled within the bi-factor model, which is a multidimensional model. It is already known that the testlet model is a special case of the bi-factor model (Rijmen, 2010). The answers given to the items in the bi-factor model are a function of both the primary trait and one of the secondary traits, and when this model is considered in the context of the testlet, secondary traits are testlet traits (DeMars, 2006). In this model, unlike the 3PL testlet model, the discrimination of an item on theta is not constrained proportionally to the discrimination on the corresponding testlet trait. The bi-factor model for dichotomous data is

$$P_{ni}\ (1) = c_i + \ (1 - \ c_i)\ \frac{\exp(a_{ip}\Theta_{np} + a_{is}\Theta_{ns} + d_i\ )}{1 + \exp(a_{ip}\Theta_{np} + a_{is}\Theta_{ns} + d_i\ )}, \tag{3}$$

where a_{ip} is the ith item slope parameter for the primary trait, a_{is} is the ith item slope for the sth secondary trait, θ_{np} is the nth person latent trait score for the primary dimension, θ_{ns} is the nth person latent score for the sth secondary trait, d_i is the ith item intercept parameter (d_i = -$a_i b_\mathrm{i}$), and c_i is the ith item guessing parameter. So, the TRT model in equation 2 can be viewed as a special case of the more general the bi-factor model in equation 3.

The testlet effect was investigated by simulation studies under different conditions. In these studies, different estimation methods improved item estimations (Luo & Wolf, 2012), equating methods were examined (Tao & Cao, 2016), evaluation of model comparison criteria (DeMars, 2012), ability parameter estimations were improved in CAT applications (Pak, 2017). The focus is on cases such as examining the effects when there are the different number of response categories (Wang et al., 2002). There are also studies in the literature evaluating parameter estimations got from different models with a similar purpose to the current study (Bradlow et al., 1999; DeMars, 2006; Koziol, 2016). The difference of this research from the mentioned studies is that it deals with more simulation conditions together.

2. METHOD

2.1. Simulation Design

Three independent variables were manipulated: a) sample size: 500, 1000, and 2000; b) testlet number: 40 dichotomous items in 4 or 10 testlets (10 items per each of 4 testlets and 4 items per each of 10 testlets); c) variance of the testlet effect: .25, .50, and 1, representing small to large effects. Wang and Wilson (2005, p.133) stated that the variances of the testlets in the real tests can be very diverse (from as small as almost zero to as large as the variance of the latent trait). In this study, the latent trait was generated with a standard normal distribution [θ ~ N(0, 1)]. Therefore, the largest variance of the testlet was chosen as 1.00. In this study, total 18 simulation conditions are considered, since a fully crossed design is used. Number of items was fixed to 40 to mimic a test of relatively medium length.

2.2. Data Generation

Similar to DeMars (2012), item discrimination and difficulty parameters were generated from a log-normal distribution N(0,1) ranging from .5 to 2.0 and a standard normal distribution N(0,1), respectively. Ability parameter and testlet variance were also generated from N(0,1) for the three possible testlet variance values determined by the specific simulation condition (same as Luo and Wolf, 2019, p.71). Based on these specifications, 40 dichotomously scored item response data were randomly generated. 100 replications were implemented for 18 conditions. Data generation was carried out through the R program.

2.3. Data Analysis

Each simulated data set for the traditional IRT, the TRT, and the bi-factor and model was calibrated using the *mirt* package (Chalmers, 2020) in R programme with the full information with the maximum likelihood estimation method with expectation-maximization (EM) algorithm. The stopping rule of the EM algorithm was set to the number of iterations = 500 or when maximum change = .00010. The models mentioned in the calibration model title in the previous section are as 3PL models. However, the guessing parameter was not considered in this study and the 2PL versions of the models were used. Because the three-parameter TRT model may encounter the problem of model convergence in practice (Wainer et al., 2007).

The performance of the three models is assessed using four criteria: the root-mean-square-error (RMSE) (i.e., total error), the bias (i.e., systematic error), mean absolute error (MAE), and the correlation between the estimated parameters and the true parameters. They are defined as;

$$RMSE\ (\hat{\pi}) = \sqrt{\frac{\sum_1^N \sum_1^R (\hat{\pi}_r - \pi)^2}{R\ X\ N}}\ , \tag{4}$$

$$Bias\ (\hat{\pi}) = \frac{\sum_1^N \sum_1^R (\hat{\pi}_r - \pi)}{R\ X\ N}, \tag{5}$$

$$MAE\ (\hat{\pi}) = \frac{\sum_1^N \sum_1^R |(\hat{\pi}_r - \pi)|}{R\ X\ N}, \tag{6}$$

where $\hat{\pi}_r$ is the estimated model parameter for the rth replication, π is the true model parameter, R is the number of replications, and N is the number of items.

3. FINDINGS

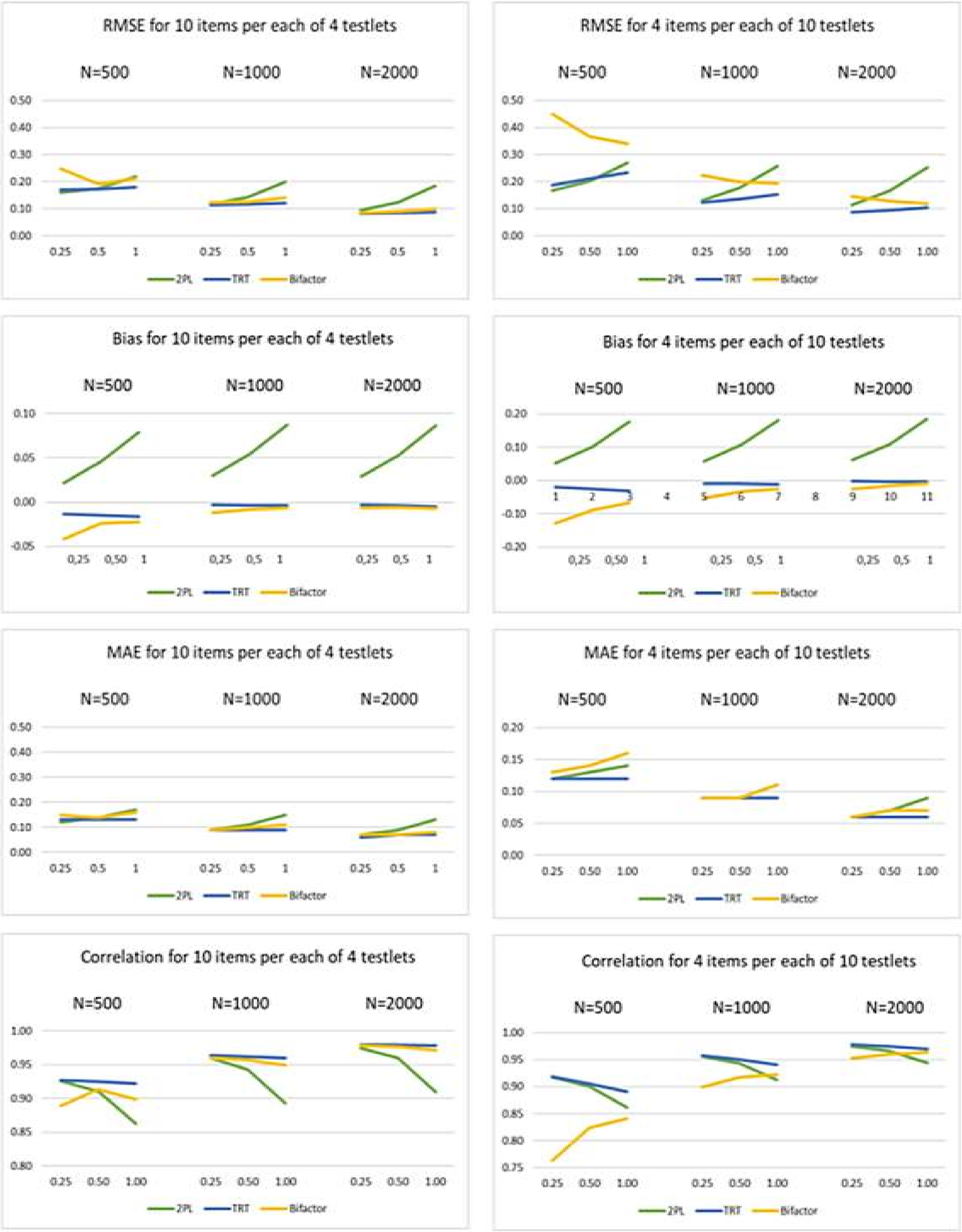
The recovery of item discrimination and item difficulty parameters across calibration models and testlet size conditions are presented in Figure 1 and Figure 2, respectively. Also, the complete set of results are summarized in the appendix as Table A1 and Table A2, respectively.

3.1. Recovery of Item Discrimination Parameters

As seen in Figure 1, under all conditions, the TRT model outperformed the traditional 2PL and bi-factor models concerning to the RMSE, the bias, the MAE of the estimated item discriminations, and correlations between the estimated and true item discriminations. For 10 items per each of 4 testlets, the performance of the TRT model outperformed with increased sample size but nearly remained stable across testlet variance (which RMSES were .17, .12 and .08 across sample size 500, 1000, and 2000, respectively). With same pattern, MAEs were .13, .9, and .7 across sample size 500, 1000, and 2000, respectively. The bi-factor model showed better recovery with increased sample size, but its performance slightly decreased with increased testlet variance.

The bi-factor model performed equivalently to the TRT model when the sample size was especially 1000 and 2000, which differences of RMSE and of correlation between the estimated and true item discriminations never exceeded .01 and MAE never exceed .02. This model showed the worst recovery under N = 500 condition when the criterion was correlation. Overall, the traditional 2PL model was the worst, showing large number of non-convergence conditions with increased testlet variance compared to both the TRT model and the bi-factor model. This means that EM cycles terminated after 500 iterations, not when the maximum change = .00010.

Figure 1. *Recovery of item discrimination.*

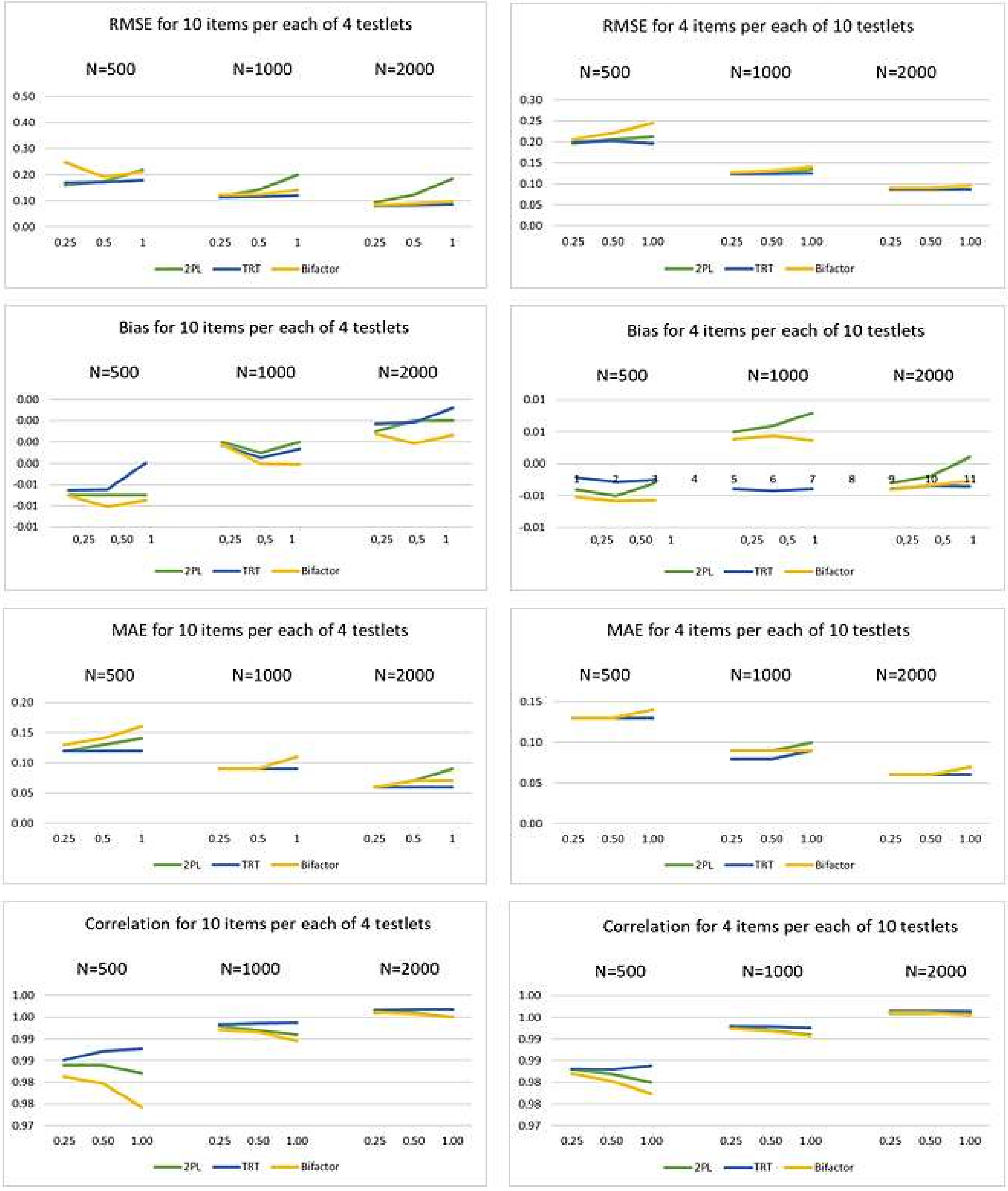


The bias showed that under all conditions, the traditional 2PL model underestimated the discrimination parameters than the true ones but the opposite tendency for the TRT and bi-factor models. For 4 items per each of 10 testlets, a similar pattern was held for the three calibration models with worse recovery, and the TRT model outperformed the other calibration models. The largest difference in RMSE between 10 items per each of 4 testlets and 4 items per each of 10 testlets was .07, .05, and .20 for the traditional 2PL, the TRT, and the bi-factor models, respectively. The bi-factor model showed more non-convergence conditions, especially when the sample size was 500. According to the correlation, the performance of the bi-factor model decreased with compared to the 10 items per each of 4 testlets, although the performance of the other two models nearly remained stable (which the differences never exceed .01).

3.2. Recovery of Item Difficulty Parameters

As seen in Figure 2, under all conditions, the TRT model slightly outperformed the traditional 2PL and bi-factor models with respect to the RMSE, the bias, the MAE of the estimated item difficulties, and correlations between the estimated and true item difficulties.

Figure 2. *Recovery of item difficulty.*



When the outcome criteria was RMSE for 10 items per each of 4 testlets, the performance of the TRT model performed better with increased sample size and slightly better with increased testlet variance under N = 500 but remained stable across testlet variance when the sample size was 1000 and 2000. Besides, for the MAE under all sample sizes, the TRT model performed stable across testlet variance. The bi-factor model outperformed recovery with increased sample size, but its performance slightly decreased with increased testlet variance. Considered RMSE, this model showed the worst performance under the N = 500 condition. The traditional 2PL model had the same pattern as the bi-factor model when the criteria were MAE under both 10 items per each of 4 testlets and 4 items per each of 10 testlets conditions. Again, the magnitude of bias ranged from .00 to .01, and correlations between the estimated and true item difficulty ranged from .98 to 1.00 and were the same across three calibration models and testlet size conditions. The differences in RMSE and in MAE were quite small, the largest difference between 10 items per each of 4 testlets and 4 items per each of 10 testlets was .02, .02, and .03 for the traditional 2PL, the TRT, and the bi-factor models, respectively.

4. DISCUSSION and CONCLUSION

Using testlets in tests violates the LI assumption. The TRT model and the bi-factor model have been widely used by researchers and practitioners to address local item dependency among the items in the same testlet. Besides these models, traditional 2PL models continue to be used for tests with testlets. In this study, dichotomous data simulated under different conditions (sample size, testlet size, and testlet variance size) were handled with three calibration models, the traditional 2PL, the TRT, and the bi-factor models, and the performances of the item parameters got from these three models were compared.

The TRT model outperformed the traditional 2PL and the bi-factor models regarding testlet size conditions, types of parameters, and outcome criteria. When the sample size was small, the performance of the bi-factor model was the worst under all other conditions and showed an irregular pattern. The reason is why a few item parameters in several replications were estimated quite differently from the true values, insomuch that RMSE was even .80 within in the replication itself. Besides, such a situation was not encountered in small samples for MAE, which produced more regular results. In this study, the stopping rule of the EM algorithm was set to the number of iterations = 500 or when maximum change = .00010. In all conditions where N = 500 and in some conditions for N = 1000, the EM cycles in the bi-factor model estimations stopped when they reached the maximum iteration. This had been an attempt to increase errors of the model estimation a little more than the normal. For the TRT model, a similar situation was observed in far less replication for N = 500. The time of the TRT and the bi-factor model estimations got longer under conditions of the large number of testlet, but the estimation time for the traditional 2PL model was barely or never impacted.

For both the traditional 2PL model and especially the bi-factor model, discrimination parameter recovery accuracy was negatively affected by increased testlet variance and the number of testlet but almost remained stable in the TRT model. The three calibration models themselves performed similar difficulty parameter recovery under conditions of the small number of items per testlet and the large number of items per testlet. Increased number of sample size was positively affected by both two types of parameter recovery for the three calibration models, especially the traditional 2PL and the bi-factor models. These findings are in line with the findings of DeMars (2006), Liu and Liu (2012), and Koziol (2016), who generated the data according to the TRT model (as was done in our study), that the performance of the TRT model was the best to the traditional 2PL model and the bi-factor model. Koziol (2016) examined the recovery of the parameters under only sample size was 1000 and used MAE to compare the efficacy of the three calibration models for recovery of item and person parameters. Our findings on recovery of the item discrimination and parameters with MAE (in Appendix, Table

A1 and Table A2) under N = 1000 were highly consistent with Koziol (2016). In contrast, Koziol (2016) reported that recovery of the item difficulty parameter only suffered under the largest testlet dependency condition (i.e., the large testlet variance and the large number of items per testlet condition). The difference between the current study and Koziol's findings (2016) could arise out of the estimation methods used within these two studies.

Sample size had a bigger impact on item parameter estimates than the other testlet conditions. Because the data followed the TRT model in this study, item parameters recovered the best with this model, as expected. In case of fully crossing the data generation according to calibration models in additional research, recovery and accuracy of parameters can be examined. However, under a large sample size and a small number of testlet, the performance of the bi-factor model could be as good as the TRT model. Also, under small testlet variance for any sample size, performing the traditional 2PL model could be as good as the TRT model. It should not be forgotten that even minor differences can have significant consequences in high-stakes contexts. Therefore, it is considered that more studies are needed on the parameter recovery and accuracy of modeling approaches. As with all studies, the results based on this study are limited to the conditions (i.e., testlet variance, the number of items per each of testlet, sample size, calibration models) given by the method. In this study, only the recovery of the item parameters was examined, the recovery of the ability parameter could be examined to vary outcome criteria and testlet conditions for future research. Also, another limitation of the present study is that we only used a medium-length test. The size of the number of items in the test can also be considered as a condition of the study.

Although testlet item structures are used in large-scale testing applications or classroom assessment, testlet dependency is generally ignored when calculating the test scores of individuals. As in this study, the effect of testlet dependency may be small or insignificant, but we do not know the exact magnitude of this effect in real-world testing situations. Therefore, as Koziol (2016) pointed out, it needs to be investigated whether test results will be biased if the testlet dependency is neglected or modeled incorrectly. To conclude, the findings of the current study show that the testlet and the bi-factor models provide to handle with LID and these two models give similar results in large samples.

Declaration of Conflicting Interests and Ethics

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

Authorship Contribution Statement

Sumeyra Soysal: Investigation, Methodology, Simulation Study, Formal Analysis, and Writing-original draft. **Esin Yilmaz Kogar**: Investigation, Resources, Methodology, Formal Analysis, and Writing-original draft.

Orcid

Sumeyra Soysal https://orcid.org/0000-0002-7304-1722
Esin Yilmaz Kogar https://orcid.org/0000-0001-6755-9018

REFERENCES

Bradlow, E.T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*(2), 153-168. https://doi.org/10.1007/bf02294533

Chalmers, R.P. (2020). *mirt: Multidimensional item response theory*. R package version 1.33.2. [Computer software manual]. http://www.R-project.org/

Chen, W.H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265–289. https://doi.org/10.3102/10769986022003265

DeMars, C.E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43*(2), 145-168. https://doi.org/10.1111/j.1745-3984.2006.00010.x

DeMars, C.E. (2012). Confirming testlet effects. *Applied Psychological Measurement, 36*, 104-121. https://doi.org/10.1177/0146621612437403

Eckes, T. (2014). Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach. *Language Testing, 31*(1), 39-61. https://doi.org/10.1177/0265532213492969

Eckes, T., & Baghaei, P. (2015). Using testlet response theory to examine local dependence in C-tests. *Applied Measurement in Education, 28*(2), 85-98. https://doi.org/10.1080/08957347.2014.1002919

Glas, C.A.W., Wainer, H., & Bradlow, E.T. (2000). MML and EAP estimation in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 271–288). Kluwer-Nijhoff.

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement, 49*(1), 82-100. https://doi.org/10.1111/j.1745-3984.2011.00161.x

Jiao, H., Wang, S., & Kamata, A. (2005). Modeling local item dependence with the hierarchical generalized linear model. *Journal of Applied Measurement, 6*(3), 311-321.

Keller, L., Swaminathan, H., & Sireci, S.G. (2003). Evaluating scoring procedures for context-dependent item sets. *Applied Measurement in Education, 16*, 207-222. https://doi.org/10.1207/s15324818ame1603_3

Koziol, N.A. (2016). Parameter recovery and classification accuracy under conditions of testlet dependency: a comparison of the traditional 2PL, testlet, and bi-factor models. *Applied Measurement in Education, 29*(3), 184-195. https://doi.org/10.1080/08957347.2016.1171767

Li, F. (2017). *An information-correction method for testlet-based test analysis: From the perspectives of item response theory and generalizability theory* (Report No. ETS RR-17-27). ETS Research Report Series. https://doi.org/10.1002/ets2.12151

Liu Y, & Liu H.Y. (2012). When should we use testlet model? A comparison study of Bayesian testlet random-effects model and standard 2-pl bayesian model. *Acta Psychologica Sinica, 44*(2), 263-275. https://doi.org/10.3724/sp.j.1041.2012.00263

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Luo, Y., & Wolf, M.G. (2019). Item parameter recovery for the two-parameter testlet model with different estimation methods. *Psychological Test and Assessment Modeling, 61*(1), 65-89.

Min, S., & He, L. (2014). Applying unidimensional and multidimensional item response theory models in testlet-based reading assessment. *Language Testing, 31*(4), 453-477. https://doi.org/10.1177/0265532214527277

Paek, I., & Cole, K. (2019). *Using R for item response theory model application*s. Routledge.

Pak, S. (2017). *Ability parameter recovery of a computerized adaptive test based on rasch testlet models* [Doctoral dissertation, University of Iowa]. Iowa University Libraries https://doi.org/10.17077/etd.5akqn3gy

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47*, 361-372. https://doi.org/10.1111/j.1745-3984.2010.00118.x

Sireci, S.G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*(3), 237-247. https://doi.org/10.1002/j.2333-8504.1991.tb01389.x

Tao, W., & Cao, Y. (2016). An extension of IRT-based equating to the dichotomous testlet response theory model. *Applied Measurement in Education, 29*(2), 108-121. https://doi.org/10.1080/08957347.2016.1138956

Thissen, D., Steinberg, L., & Mooney, J.A. (1989). Trace lines for testlets: A use of multiplecategorical-response models. *Journal of Educational Measurement, 26*, 247-260. https://doi.org/10.1111/j.1745-3984.1989.tb00331.x

Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item iterations on the estimated discrimination parameters in item response theory. *Psychological Methods, 6*(2), 181-195. https://doi.org/10.1037/1082-989x.6.2.181

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education, 8*, 157–186. https://doi.org/10.1207/s15324818ame0802_4

Wainer, H., Bradlow, E.T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245–270). Kluwer-Nijhoff.

Wainer, H., Bradlow, E.T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge University Press.

Wainer, H., & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185-201. https://doi.org/10.1111/j.1745-3984.1987.tb00274.x

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*(1), 22-29. https://doi.org/10.1002/j.2333-8504.1998.tb01749.x

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*(3), 203-220. https://doi.org/10.1002/j.2333-8504.2001.tb01851.x

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187-213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Zenisky, A.L., Hambleton, R.K., & Sired, S.G (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. *Journal of Educational Measurement, 39*(4), 291-309. https://doi.org/10.1111/j.1745-3984.2002.tb01144.x

APPENDIX

Table A1. *Recovery of item discrimination parameters.*

Calibration Model	Conditions		Testlet Size							
			10 items per each of 4 testlets				4 items per each of 10 testlets			
	SS	TV	RMSE	Bias	MAE	Corr	RMSE	Bias	MAE	Corr
Traditional 2PL	500	.25	.16	.02	.12	.93	.17	.05	.13	.92
		.50	.17	.05	.14	.91	.20	.10	.15	.90
		1.00	.22	.08	.17	.86	.27	.18	.21	.86
	1000	.25	.12	.03	.09	.96	.13	.06	.10	.96
		.50	.14	.05	.11	.94	.18	.11	.14	.94
		1.00	.20	.09	.15	.89	.26	.18	.20	.91
	2000	.25	.09	.03	.07	.98	.11	.06	.09	.98
		.50	.12	.05	.09	.96	.17	.11	.13	.97
		1.00	.18	.09	.13	.91	.25	.18	.19	.94
Testlet Response Theory	500	.25	.17	-.01	.13	.93	.19	-.02	.14	.92
		.50	.17	-.01	.13	.92	.21	-.03	.15	.91
		1.00	.18	-.02	.13	.92	.23	-.03	.17	.89
	1000	.25	.11	.00	.09	.96	.12	-.01	.09	.96
		.50	.12	.00	.09	.96	.14	-.01	.10	.95
		1.00	.12	.00	.09	.96	.15	-.01	.11	.94
	2000	.25	.08	.00	.06	.98	.09	.00	.07	.98
		.50	.08	.00	.07	.98	.09	.00	.07	.97
		1.00	.09	.00	.07	.98	.10	.00	.08	.97
Bi-factor	500	.25	.25	-.04	.15	.89	.45	-.13	.24	.76
		.50	.19	-.02	.14	.91	.37	-.09	.21	.82
		1.00	.21	-.02	.16	.90	.34	-.07	.20	.84
	1000	.25	.12	-.01	.09	.96	.22	-.05	.13	.90
		.50	.13	-.01	.10	.96	.20	-.03	.13	.92
		1.00	.14	-.01	.11	.95	.19	-.03	.12	.92
	2000	.25	.08	-.01	.07	.98	.14	-.02	.09	.95
		.50	.09	-.01	.07	.98	.13	-.02	.08	.96
		1.00	.10	-.01	.08	.97	.12	-.01	.08	.96

Note. RMSE: Root mean square error, MAE: Mean absolute error, Corr: Pearson correlation coefficient.

Table A2. *Recovery of item difficulty parameters.*

Calibration Model	Conditions		Testlet Size							
			10 items per each of 4 testlets				4 items per each of 10 testlets			
	SS	TV	RMSE	Bias	MAE	Corr	RMSE	Bias	MAE	Corr
Traditional 2PL	500	.25	.19	-.01	.12	.98	.20	.00	.13	.98
		.50	.19	-.01	.13	.98	.21	-.01	.13	.98
		1.00	.20	-.01	.14	.98	.21	.00	.14	.98
	1000	.25	.13	.00	.09	.99	.13	.01	.09	.99
		.50	.13	.00	.09	.99	.13	.01	.09	.99
		1.00	.15	.00	.11	.99	.14	.01	.10	.99
	2000	.25	.09	.00	.06	1.00	.09	.00	.06	1.00
		.50	.10	.00	.07	1.00	.09	.00	.06	1.00
		1.00	.12	.00	.09	1.00	.10	.00	.07	1.00
Testlet Response Theory	500	.25	.19	-.01	.12	.99	.20	.00	.13	.98
		.50	.18	-.01	.12	.99	.20	.00	.13	.98
		1.00	.17	.00	.12	.99	.20	.00	.13	.98
	1000	.25	.12	.00	.09	.99	.12	.00	.08	.99
		.50	.12	.00	.09	.99	.12	.00	.08	.99
		1.00	.12	.00	.09	.99	.13	.00	.09	.99
	2000	.25	.08	.00	.06	1.00	.09	.00	.06	1.00
		.50	.08	.00	.06	1.00	.09	.00	.06	1.00
		1.00	.08	.00	.06	1.00	.09	.00	.06	1.00
Bi-factor	500	.25	.22	-.01	.13	.98	.21	-.01	.13	.98
		.50	.23	-.01	.14	.98	.22	-.01	.13	.98
		1.00	.27	-.01	.16	.97	.24	-.01	.14	.98
	1000	.25	.14	.00	.09	.99	.13	.00	.09	.99
		.50	.14	.00	.09	.99	.13	.00	.09	.99
		1.00	.16	.00	.11	.99	.14	.00	.09	.99
	2000	.25	.09	.00	.06	1.00	.09	.00	.06	1.00
		.50	.10	.00	.07	1.00	.09	.00	.06	1.00
		1.00	.11	.00	.07	1.00	.10	.00	.07	1.00

Note. RMSE: Root mean square error, MAE: Mean absolute error, Corr: Pearson correlation coefficient.