e-ISSN: 2148-7456

a peer-reviewed
online journal

hosted by DergiPark

International Journal of Assessment Tools in Education

Volume: 12 Issue: 1 March 2025

https://dergipark.org.tr/en/pub/ijate
https://ijate.net/index.php/ijate

e-ISSN: 2148-7456

Volume 12 Issue 1 March 2025

Editor Dr. Safiye BILICAN DEMIR
Address Department of Educational Sciences, Kırıkkale University, Ankara yolu Km 7. Yahşihan 71450, Kirikkale, Türkiye

E-mail ijate.editor@gmail.com
safiyebilican@gmail.com

Publisher Dr. Izzet KARA
Address Pamukkale University, Education Faculty, Kinikli Campus, 20070, Denizli, Türkiye
Phone +90 258 296 1036
Fax +90 258 296 1200
E-mail ikara@pau.edu.tr
ijate.info@gmail.com

Journal Contact Dr. Eren Can AYBEK
Address Department of Educational Sciences, Pamukkale University, Faculty of Education, Kinikli Yerleskesi, Denizli, 20070, Türkiye
Phone +90 258 296 31050
Fax +90 258 296 1200
E-mail erencanaybek@gmail.com

Address Dr. Anil KANDEMIR
Department of Educational Sciences, Agri Ibrahim Cecen University, Faculty of Education, Agri, Türkiye
akandemir@agri.edu.tr

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

Cover Design IJATE

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 comprehension 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 Türkiye)].

In IJATE, there are no charges under any procedure for submitting or publishing an article.

Indexes and Platforms:

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

CONTENTS

Research Articles

International Journal of Assessment Tools in Education
2025, Vol. 12, No. 1, 1–19

https://doi.org/10.21449/ijate.1468431

journal homepage: https://dergipark.org.tr/en/pub/ijate *Research Article*

Self-efficacy of teachers in special education schools for teaching through play

Nejmi Yıldırım[1], Şerife Şenay İlik[2*]

[1]Milli Eğitim Bakanlığı, Konya, Türkiye
[2]Necmettin Erbakan University, Ahmet Keleşoğlu Education Faculty, Konya, Türkiye

ARTICLE HISTORY

Received: Apr. 15, 2024
Accepted: July 21, 2024

Keywords:
Special education schools,
Self-efficacy,
Teaching through play,
Teachers working in special education,
Scale development.

Abstract: This study aims to ascertain the degree to which special education teachers believe they can effectively teaching through play. A descriptive survey model was used in the research. Our study sample consisted of 241 teachers working in special education in a state institution under the Konya Provincial Directorate of National Education. The data collection tool of the research is the 18-item "Self-Efficacy Scale for Teachers Working in Special Education in the Process of Teaching Through Play" developed by the researchers. As a result of the research, it was concluded that the teachers participating in the study had very high self-efficacy perceptions in terms of planning instruction according to program stages and developmental characteristics, as well as in the application process and evaluation related to developmentalcharacteristics. However, their perceptions of self-efficacy regarding the application method were high. Regarding gender, it was shown that female teachers have stronger self-efficacy, and preschool instructors have higher average scores than class teachers and special education teachers based on the undergraduate graduation variable. In addition, it was determined in the research that the self-efficacy perceptions of teachers in terms of the process of teaching through play are similar according to the professional seniority and the disability type of students they work with.

1. INTRODUCTION

Play is an activity that makes the child happy while engaging them actively in the process and sustaining their attention. It is mostly an activity in which the child participates willingly. Play, which is of great importance in the child's development, is important for the development of the student. Clues obtained from the child's behaviors during play provide important information about his/her development. Play is a way for children to explore themselves and the world, as well as to express themselves (Pehlivan, 2014). Through play, children develop language, personality, and behavior, and thus prepare for situations they may encounter later in life (Manwaring, 2011). In other words, through play, students prepare for their future lives. Students can simulate dangerous situations in real life through play. In this way, they can learn what they need to learn about life through play. Additionally, when used in an academic environment, play provides a natural and enjoyable teaching environment for the child, enriching and diversifying the child's surroundings (Tuğrul, 2014). A child who experiences

*CONTACT: Şerife Şenay İLİK ✉ senayilik@gmail.com Necmettin Erbakan University, Ahmet Keleşoğlu Education Faculty, Konya, Türkiye

e-ISSN: 2148-7456

the feeling of success through play will show an increased interest in learning experiences that will affect other areas of life (Tuzcuoğlu & Tuzcuoğlu, 2004). Developing games suitable for the level of communication initiation and maintenance necessary in the communication process for individuals with special needs, who experience limitations, should be ensured. Thus, by having an active role in the game, their communication skills develop, and they learn to understand social roles, control their reactions, and direct their feelings (Akandere, 2003; Stagnitti O'Connor, 2012).

Teachers of individuals with special needs have responsibilities to fulfill in order to provide the mentioned benefits of the teaching process through games. When using play in the education of individuals with special needs, it should be remembered that students also have limitations during play and need help in initiating and sustaining play. In teaching play skills, first of all, there should be a rich environment for the individual and this environment in which the child is involved should be designed per the needs and characteristics of the child (MEB, 2014). The process of designing this environment should start with games and toys that the child knows and feels comfortable with. The game should be played with the same person at the beginning and there should be no change in the game for a while for the child to get used to the game. The play environment should appeal to multiple senses. In addition, having a familiar person with the child makes it easier for the child to adapt to the game and continue playing (Sarı, 2017). In addition to music and fun in the game, the child should be encouraged with various reinforcement schedules, and guidance should be given to the child without letting him/her feel it. The game should be stopped if it is thought that it is moving away from the targeted goal, and the student should be left alone by not insisting on its continuation. In addition to giving the child the joy of learning, the child should also be given the chance to make mistakes in the game (Güneş, 2015). Teachers should plan the process, purpose, and content of teaching through play well. Objectives should be appropriate for students and content should be enriched within the possibilities. The process of teaching through play should be continued by taking into account the age group, disability type, and needs of individuals with special needs. Students should be active in play and control the rules themselves during the play process. Teachers should consider the sequential nature of the subject in the process of teaching through play. At the same time, the subjects in the game should be suitable for the complementary nature of the subjects through play. Evaluation criteria in the process of teaching through play should be clear, understandable, and appropriate for the student level. At the end of the evaluation, it should be checked whether the achievements have been attained or not. Play is a method for the development and learning of individuals with special needs. However, when teachers transfer the game to educational environments, they need to pay attention to some situations. Since individuals with special needs experience problems in interest and concentration, it is important to provide diversity in play tools and materials. While normal individuals can manage themselves with any toy and play, individuals with special needs may need adaptations in toys and guidance in the game (Brodin, 1999). For example, if our student with special needs has a disability that prevents him/her from holding a toy, some arrangements need to be made for the student to play with this toy.

Some of these studies are as follows: Kaya (2010) examined the effectiveness of a play intervention program (OMP) on the cognitive skills of 3-5-year-old children with special needs, and found that their performance improved positively after the intervention. Ergin (2017) investigated the effectiveness of teaching by increasing the variety of imaginary play behaviors of children with autism spectrum disorder (ASD) with increasing hints and found that all participants in the study acquired and retained the play gains included in the play theme. Kaptan (2018) studied the effectiveness of video modeling in teaching sociodramatic play to children with autism spectrum disorder. The research concluded that video modeling is effective in teaching sociodramatic play to children with autism spectrum disorder. Kaplan (2019) examined the effectiveness of teaching counting skills through play for students with intellectual disabilities and showed that play was effective and that students were able to exhibit and generalize these gains after the applications. In another study examining the effect of play on the social development of children with special needs, teacher opinions were included and it was stated that play has a positive effect on attention, and teachers play games that reinforce cooperation, sharing, and classroom activities and that they are good at creativity, enrichment, and drama (Yaman, 2019). Janson (2001) analysed the joint play interaction of visually impaired and sighted preschool children by stating that co-play involves common physical space, social thought and experience, and common symbols rather than just sharing the physical environment. The findings of the study revealed that children with disabilities could not use common symbols in joint play. Therefore, it was found that they experienced difficulties in joint play. The study concluded that individual characteristics are the points to be considered in joint play. Stanley (2003) conducted a study on the relationship between symbolic play and other developmental areas (non-verbal cognitive competence, receptive language, expressive language, and social development) and found that there is a strong relationship between the symbolic play behaviors of autistic children and their non-verbal cognitive competence and that social development is related to verbal competence and social competence is the determining feature of symbolic play. Fridenson Hayo *et al*. (2017) investigated the outcomes of "Emotiplay," a cross-cultural serious game developed to teach emotions to children with autism. The study concluded that Emotiplay is an effective and motivating psycho-educational intervention. It was found to teach the recognition of cross-cultural expressions from faces, voices, and body language, and to integrate these skills contextually for children with high-functioning autism. Cano *et al*. (2019) showed that using Game Analytics information is an effective way to evaluate both the game design and implementation, especially when other evaluation types requiring user participation are limited. The study was based on an evidence-based evaluation of a learning game for users with intellectual disabilities. Jeong *et al*. (2020) conducted a study on the development of 'ZOOCUS,' a board game with multiple experiences for intellectually disabled students. This study aimed to improve the attention and concentration of intellectually disabled students by combining board games and AR applications in "ZOOCUS," an AR board with multiple experiences developed to improve the social skills, concentration, and working memory of intellectually disabled students. This study found that the attention and concentration of intellectually disabled students were improved, and by adding an AR function to the board game, various visual-auditory elements were provided to maximize feedback according to the game behavior. As seen in many previous studies, board games can develop some basic skills necessary for intellectually disabled students. However, among the many board games used in the studies, there is no case where a board game specifically developed for intellectually disabled students is used and developed commercially.

In the literature, there are studies determining teacher competence. Some of these studies conducted on teacher competence are as follows: Kadim (2012) examined the self-efficacy of preschool teachers in teaching through play according to various variables. No significant difference was found in teachers' self-efficacy in implementing and evaluating play activities according to gender, age, education status, seniority, education age group, class size, school location, and school type variables. Significant differences were obtained only in terms of age levels for preschool teachers' professional self-efficacy in play teaching. Piştav Akmeşe and Kayhan (2017) examined the self-efficacy of teachers working in special education in teaching through play. The study used the "Preschool Period Play Teaching Self-Efficacy Questionnaire" developed by Kadim (2012) to determine the self-efficacy of teachers working in special education in play teaching. The findings of the study showed that there was a significant difference in planning, implementation, and evaluation of self-efficacy according to the graduation field, receiving education related to play, and professional seniority variables. Another result of the study is that the education level variable is effective in evaluating play-teaching activities, professional self-efficacy, and play-teaching effectiveness. In terms of the gender variable, its effect was observed in the sub-dimension of implementing play activities. Celep (2020) examined the self-efficacy levels of teachers working in preschool special education schools and their creative personality characteristics and the relationship between them. In the research, the "Personal Information Form", "Preschool Period Play Teaching Self-Efficacy Questionnaire", and "Creative Personality Traits Scale" were used as data collection tools. The findings of the research showed that there was a significant difference in the self-efficacy levels of teachers working in preschool special education schools according to variables. At the same time, it was found that the creative personality characteristics of teachers showed significant differences according to variables such as gender, age, professional seniority, class size, presence of auxiliary staff in the classroom, receiving education related to play at the university, following publications related to play, and receiving education related to play and creativity. Additionally, a significant high relationship was found between the self-efficacy levels of teachers working in preschool special education schools and their creative personality characteristics.

Special needs individuals' educational needs can vary, with each individual showing individual differences based on their needs. In consideration of these differences, teachers in special education use different teaching methods and techniques. It is important that these methods and techniques enrich the education of special needs students and be engaging. One of the most effective and engaging ways to enrich teaching is to incorporate play into education, in other words, to teach through play. Teaching through play aims to meet the educational needs of special needs students engagingly and enjoyably, unlike typically developing students. In this context, for teachers to effectively use teaching through play in their educational activities with students, they need to have certain competencies. It is necessary to determine the competencies of teachers in teaching through play, which they use to meet the educational needs of special needs of individuals. Based on the studies conducted in the field, a scale determining the competence of teachers of students with special needs has not been found utilized in the present research. Therefore, it is considered important to develop a scale called "Special Education Teachers' Self-Efficacy in the Teaching through Play Process" by the researcher to determine the self-efficacy of special education teachers in teaching through play and to determine the self-efficacy of special education teachers in teaching through play according to various variables using this scale. In this regard, this study aims to determine the self-efficacy of special education teachers in the teaching through play process. In line with this aim, the following sub-problems will be addressed in the research:

1. How confident are special education teachers in their ability to educate through play?

2. Does the gender variable have a significant impact on the self-efficacy of special education instructors in the process of teaching via play?

3. Does the professional seniority variable significantly affect the self-efficacy of special education teachers in the process of teaching via play?

4. Does a special education teacher's field of graduation affect how confident they feel about themselves when it comes to the play-based learning process?

5. Does the self-efficacy of special education instructors in the play-based learning process vary depending on the disability group they work with?

2. METHOD

In this section, information about the research model, population and sample, data collection tools, and data analysis is provided.

2.1. Research Model

This study aims to determine the self-efficacy of special education teachers in the teaching through play process. In line with this aim, a descriptive survey model was used in the study. Studies conducted using the descriptive survey model aim to reveal the characteristics of individuals or groups. According to the descriptive survey model, research aims to describe a current or past situation as it is. The case, person, or objects that constitute the subject of the research are tried to be described without any intervention under existing conditions (Karasar, 2009).

2.2. Population-Sample

The target population of this study consists of 632 special education teachers working in state institutions affiliated with the Konya Provincial Directorate of National Education in the 2021-2022 academic year. The number of teachers that need to be randomly selected to represent 632 special education teachers with a confidence interval of 95% (α= .05) is 239 (Yazıcıoğlu &Erdoğan, 2014). Within the scope of the research, the participation of 241 special education teachers randomly selected was ensured, thus meeting the required sample size.

2.2.1. *Information about the population and sample*

The distribution of the teachers working in special education who participated in the study according to the memorable characteristics is shown in Table 1.

Table 1. *Distribution of participants by diagnostic characteristics.*

		f	%
Seniority	0-10 years	140	58.1
	11-20 years	68	28.2
	20 years and above	33	13.7
Field of Undergraduate	Special Education Teaching	151	62.7
	Classroom teaching	51	21.2
	Pre-school teaching	39	16.2
Disability group	Mildly Mentally Disabled	73	30.3
	Medium-Severe Mentally Disabled	82	34.0
	Autism	75	31.1
	Hearing loss	7	2.9
	Blind	4	1.7
Gender	Woman	133	55.2
	Male	108	44.8
	Total	241	100.0

When Table 1 is examined, it is understood that 58.1% of the participants have 0-10 years, 28.2% have 11-20 years, and 13.7% have 20 years and more professional seniority. A large proportion of the participants (62.7%) graduated from the special education department. 30.3% of the participants stated that they work with mildly intellectually disabled, 34% with moderate to severe intellectually disabled, 31.1% with autism, 2.9% with hearing loss, and 1.7% with visually impaired groups. 55.2% of the participants are female, and 44.8% are male.

2.3. Data Collection Tool

2.3.1. *Development of items for the self-efficacy scale of teachers working in special education for the teaching through play process*

This study aims to determine the self-efficacy of teachers working in special education in the teaching through play process. In line with this aim, a Likert-type competency scale was developed following the steps of scale development (DeVellis, 2017; Tezbaşaran, 2008). The scale development steps followed in the study are as follows;

- Literature review
- Development of item pool by deciding on the appropriate measurement tool

- Presenting the item pool to experts
- Preparing the draft scale
- Conducting pilot studies
- Data collection
- Validity and reliability studies
- Finalizing the scale (DeVellis, 2017).

The conceptual structure and sub-dimensions of the scale were determined by conducting a literature review on the self-efficacy of teachers working in special education in the teaching through play process (Cano *et al*., 2019; Celep, 2020; Ergin, 2017; Fridenson Hayo *et al*., 2017; Janson, 2001; Jeong *et al*., 2020; Kadim, 2012; Kaplan, 2019; Kaptan, 2018; Kaya, 2010; Piştav Akmeşe & Kayhan, 2017; Stanley, 2003; Yaman, 2019). Based on this literature review, a draft item pool consisting of 54 items was created under three sub-dimensions: planning, implementation, and evaluation. The items were formulated considering the steps in planning, implementing, and evaluating the teaching process in special education. Additionally, care was taken to ensure that the items did not encompass multiple behaviors, judgments, or attitudes. Opinions on the 54 items in the item pool were gathered from academics and experts actively engaged in the field. The aim of obtaining expert opinions was to determine the content validity, which indicates the extent to which the items measure the intended aspects and their adequacy in terms of quantity and quality (Büyüköztürk, 2015). It is crucial for the researchers developing the measurement tool and the experts evaluating the scale to have a shared understanding of the scale's content (DeVellis, 2017; Tavşancıl, 2018). Opinions were obtained from 6 Special Education Teachers, 4 Preschool Teachers, 2 Physical Education and Play Teachers working in special education, 1 Play Therapist, 3 faculty members from the Special Education Department at Necmettin Erbakan University, and 1 scale development (statistics) expert. Based on the feedback, attention was paid to ensuring the items were easily comprehensible by the participants and written in clear and concise language. Following expert feedback and a literature review, the scale was reduced to 37 items without compromising content validity

Before the pilot application, our scale, which was prepared as a 37-item Likert scale, was applied to 30 third-year special education teacher candidates for pre-application. After the application, 2 items that the teacher candidates stated were not fully understood were revised. Apart from this change, no other change was needed. It was determined that teacher candidates filled the draft scale in an average of 20 minutes, so the filling time of the scale was determined as 20 minutes. After the pilot application, the Pilot Scale Form, prepared after the pre-application, was applied to 143 special education teachers working in state institutions affiliated with Ankara and Kırıkkale Provincial Directorates of National Education.

2.3.2. *Results of exploratory factor analysis*

Factor analysis, known as a multivariate analysis technique, aims to select the most correlated variables among many variables to create fewer conceptually meaningful new variables (Çokluk *et al*., 2010). Sample sufficiency and the suitability of the data for factorization should be checked before the analysis. The calculated Kaiser-Meyer-Olkin (KMO= .88>.70) coefficient indicated that the sample size was sufficient. The result of the Barlett Sphericity test ($\chi^2(153)$) = 1323.10; $p<.001$) indicated that the data were suitable for factor analysis.

The necessary assumptions were met, and factor analysis was conducted. Principal component analysis is one of the factor extraction methods. In this study, this method was used to conduct the factor analysis. The value of .32 was assigned as the cutoff point for factor loadings (Tabachnick & Fidell, 2007). As a result of applying factor analysis, it was observed that the eigenvalues of five factors were above one. Also, a plateau was formed in the eigenvalue factor graph after the fifth point. The contribution of the components after the fifth point to the variance is small. At the same time, it was observed that they were approximately the same. Based on these results, it was decided that the number of factors should be five. In the next step after determining the number of factors and the decision, the scale items were forced into five factors for analysis. The Varimax orthogonal rotation method was used. Items with factor loadings below the cutoff point (m1, m9, m18, m19, m37) and overlapping items that loaded on multiple factors (m5, m10, m11, m12, m13, m14, m15, m16, m17, m20, m21, m27, m31, m33) were each removed from the scale, and the analysis was repeated. As a result of the final analysis, it was observed that 18 items remained on the scale. The factor structure of the Self-Efficacy Scale for the Teaching through Play Process is shown in Table 2.

Table 2. *Teachers working in special education exploratory factor analysis results of the self-efficacy scale for the game teaching process.*

Item number	Factor load					MOV*	DMTK**	self-worth	Variance explained (%)
	1	2	3	4	5				
m29	.80	.18	.20	.09	.26	.79	.81	7.32	40.69
m23	.78	.22	.17	.12	.09	.71	.74		
m24	.74	.18	.22	.03	.20	.67	.71		
m30	.73	.28	.32	.06	.04	.72	.75		
m28	.71	.09	.02	.22	.27	.63	.65		
m32	.57	.31	.26	.31	-.02	.58	.61		
m3	.25	.88	.06	.12	.15	.88	.84	1.74	9.66
m4	.22	.86	.21	.09	.07	.84	.80		
m2	.25	.80	.17	.25	.10	.80	.78		
m35	.13	.10	.84	-.03	.17	.77	.61	1.51	8.37
m36	.35	.14	.74	.11	.06	.71	.65		
m34	.25	.18	.72	.23	.10	.68	.62		
m8	.20	.12	.20	.81	.08	.77	.67	1.29	7.18
m6	.00	.11	.01	.81	.09	.67	.55		
m7	.19	.13	.03	.75	.06	.62	.57		
m22	.05	.17	.03	.02	.86	.77	.54	1.14	6.34
m25	.30	.10	.25	.09	.77	.80	.70		
m26	.30	-.01	.15	.23	.58	.60	.53		

*MOV= Item common variance , **DMTK= Corrected Item – Total Correlation

Exploratory Factor Analysis (EFA) revealed that the factor loadings of the items in the first factor ranged from .57 to .80, in the second factor from .80 to .88, in the third factor from .72 to .84, in the fourth factor from .75 to .81, and in the fifth factor from .58 to .86. The commonality values of the items in Factor Analysis need to be greater than .40 (Field, 2013). The results indicated that this condition was met for all items. The five-factor scale explained 72.24% of the total variance. It is considered important for the variance explained by the factors to exceed 50%, as this means more than half of the variance of the variables is explained. The representational power of the items is at a high level (Yaşlıoğlu, 2017). The first, second, third, fourth, and fifth factors were named Self-Efficacy for Implementation Process, Self-Efficacy for Planning Teaching According to Program Stages, Self-Efficacy for Evaluation, Self-Efficacy for Planning Teaching According to Developmental Characteristics, and Self-Efficacy for Implementation Method, respectively.

2.3.3. *Results of confirmatory factor analysis*

The results of the Exploratory Factor Analysis indicated that the Self-Efficacy Scale for Teachers Working in Special Education in the Teaching through Play Process had a five-factor structure. In the next step, it was tested whether the five-factor structure of the scale was confirmed with the collected data. For this purpose, confirmatory factor analysis (CFA), which aims to test the fit of the proposed factor structure, was conducted (Yurt, 2023). The analysis was performed using the Maximum Likelihood Estimation method. The fit indices for the five-factor model are presented in Table 3, including the Root Mean Square Error of Approximation (RMSEA), Standardized Root Mean Residual (SRMR), Goodness of Fit Index (GFI), Adjusted Goodness of Fit Index (AGFI), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and Incremental Fit Index (IFI).

Table 3. *Fit values of the three-factor structure of the self-efficacy scale for the game teaching process of teachers working in special education.*

Criterion	Good Fit	Acceptance Possible Fit	Obtained Values	Source
(χ^2/df) -	≤ 3	≤ 4-5	1.37	Byrne, 1989
RMSEA	≤ .05	.06-.08	.05	Browne & Cudeck, 1993
SRMR	≤ .05	.06-.08	.06	
GFI	≥ .90	.85-.90	.88	Tanaka Huba, 1985; Jöreskog & Sörbom, 1984
AGFI	≥ .90	.80-.90	.84	
CFI	≥ .95	.90-.94	.96	Bollen, 1989
TLI	≥ .95	.90-.94	.95	
IFI	≥ .95	.90-.94	.96	

When Table 3 is examined, it is understood that the five-factor structure of the scale is in good agreement with the data obtained from the Self-Efficacy Scale for the Play Teaching Process of Teachers Working in Special Education, and the five-factor structure of the scale is confirmed. The five-factor model of the scale is shown in Figure 1. All factor loadings shown in the model are statistically significant at the p<.001 level.

Figure 1. *CFA diagram of the self-efficacy scale for the game teaching process*. χ^2=171.20; df =125; p<.001

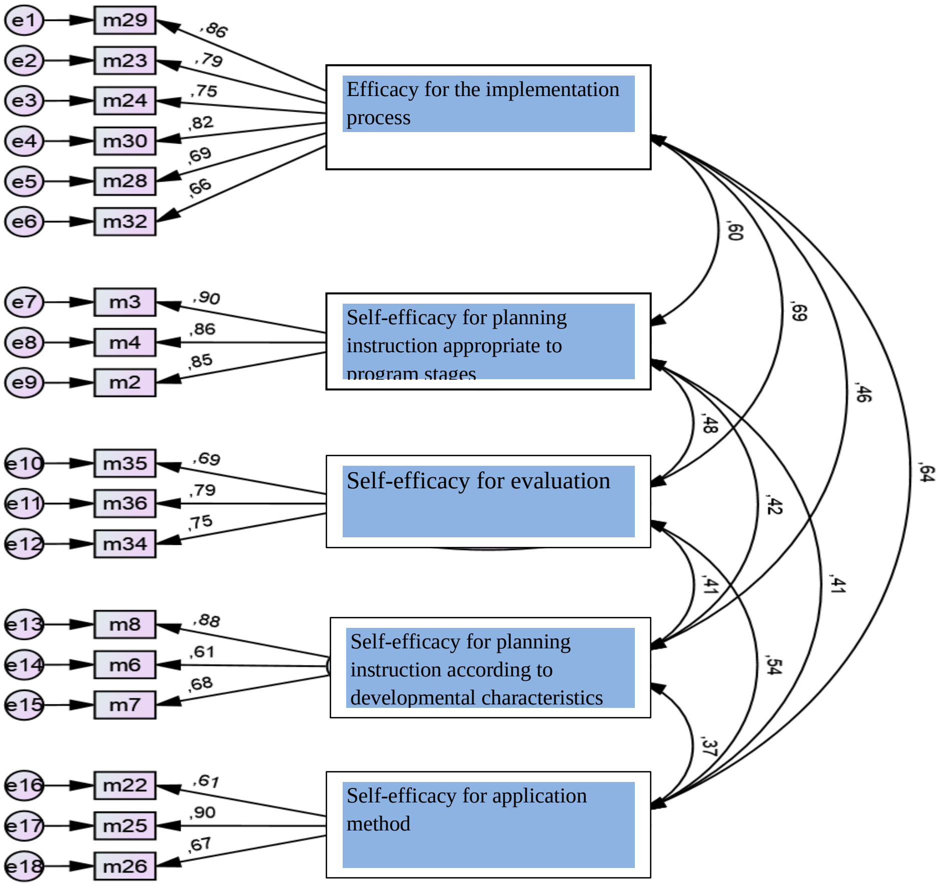

e1
m29
,86
e2
m23
,79
e3
m24
,75
e4
m30
,82
e5
m28
,69
e6
m32
,66
Efficacy for the implementation process
e7
m3
,90
e8
m4
,86
e9
m2
,85
Self-efficacy for planning instruction appropriate to program stages
e10
m35
,69
e11
m36
,79
e12
m34
,75
Self-efficacy for evaluation
e13
m8
,88
e14
m6
,61
e15
m7
,68
Self-efficacy for planning instruction according to developmental characteristics
e16
m22
,61
e17
m25
,90
e18
m26
,67
Self-efficacy for application method
,60
,69
,46
,64
,48
,42
,41
,41
,54
,37

Table 4. *Confirmatory factor analysis results of special education teachers' self-efficacy scale for teaching through play process.*

Factor	Item number	Factor load	C.R.	AVE	MSV	MaxR (H)
Efficacy for the implementation process	m29	.87***	.89	.59	.47	.91
	m23	.79***				
	m24	.75***				
	m30	.82***				
	m28	.69***				
	m32	.66***				
Self-efficacy for planning instruction appropriate to program stages	m3	.90***	.90	.76	.36	.91
	m4	.86***				
	m2	.85***				
Self-efficacy for evaluation	m35	.69***	.79	.55	.47	.80
	m36	.79***				
	m34	.75***				
Self-efficacy for planning instruction according to developmental characteristics	m8	.88***	.77	.54	.21	.83
	m6	.61***				
	m7	.68***				
Self-efficacy for application method	m22	.61***	.78	.54	.42	.85
	m25	.90***				
	m26	.67***				

*** *p*<.001, C.R. = Composite reliability, AVE = Average variance extracted, MSV = Maximum shared variance, MaxR (H) = Maximum reliability.

Upon reviewing Table 4, it can be observed that the factor loadings of the items in the scale ranged from .61 to .90 as a result of the Confirmatory Factor Analysis (CFA). It was determined that the internal reliability criterion was met, with CR>.70 and AVE>.50. The criterion for convergent validity (CR>AVE) was also entirely met, indicating that convergent validity was achieved (Malhotra & Dash, 2011; Yurt, 2023). In terms of discriminant validity, it was observed that the condition MSV<AVE was entirely met. Additionally, it was found that the MaxR(H) reliability value was greater than the CR values, supporting the conclusion that discriminant validity was achieved (Hu & Bentler, 1999).

2.3.4. *Reliability analysis results*

Cronbach's Alpha coefficients were calculated to determine the reliability of the Self-Efficacy Scale for Teachers Working in Special Education in the Teaching through Play Process. Values between .60-.80 indicate that the measurement tool is quite reliable, while values between .81-1.00 indicate that the measurement tool is highly reliable (Özdamar, 2004).

Table 5. *Cronbach alpha coefficients of self-efficacy scale factors for the game teaching process.*

Dimension	Number of items	Cronbach Alpha
Self-efficacy for the implementation process	6	.89
Self-efficacy for planning instruction appropriate to program stages	3	.90
Self-efficacy for evaluation	3	.78
Self-efficacy for planning instruction according to developmental characteristics	3	.75
Self-efficacy for application method	3	.75
overall scale	18	.90

When Table 5 is examined, the alpha coefficients calculated for the factors of Self-Efficacy for Application Process, Self-Efficacy for Planning Instruction According to Program Stages, Self-

Efficacy for Evaluation, Self-Efficacy for Planning Instruction According to Developmental Characteristics, and Self-Efficacy for Application Method are .89, .90, .78, .75, and .75, respectively. The alpha coefficient calculated for the overall scale is .90. The obtained coefficients have shown that the reliability of the measuring instrument based on internal consistency is at a sufficient level.

According to the results of the validity analysis, it is understood that the Self-Efficacy Scale for Teaching through Play for Teachers Working in Special Education has a 5 factors structure. It has been observed that the six-factor structure is consistent with the collected data. The final form of the scale consists of 18 items. It has been determined that the reliability of the measuring instrument based on internal consistency is at a satisfactory level. The results obtained have shown that the measuring instrument can be used to determine the self-efficacy perceptions of teachers working in special education regarding the teaching through play process.

2.4. Data Analysis

In the scope of the research, descriptive analyses were conducted to examine the scores obtained from the Self-Efficacy Scale for Teaching through Play for Teachers Working in Special Education. Skewness and kurtosis coefficients were used to examine the distribution of scores obtained from the scale. Skewness and kurtosis coefficients were calculated, and the normal distribution assumption was examined for the scores obtained from the scales. Skewness and kurtosis coefficients within the ±1.5 range indicate that the normal distribution assumption is met (Tabachnick & Fidell, 2013). Skewness and kurtosis coefficients calculated for the scores obtained from the measuring instrument in this study were within the specified range (see Table 6). The results obtained have shown that the scores obtained from the Self-Efficacy Scale for Teaching through Play for Teachers Working in Special Education have a normal distribution. In this regard, data were analyzed using parametric analysis techniques.

Table 6. *Skewness and kurtosis coefficients of the scores obtained from the self-efficacy scale for the game teaching process of teachers working in special education.*

Variables	Distortion		Kurtosis	
	z	*SE*	*Z*	*SE*
Self-efficacy for planning instruction appropriate to program stages	-.66	.16	-.65	.31
Self-efficacy for planning instruction according to developmental characteristics	-1.35	.16	.80	.31
Self-efficacy for application method	.36	.16	-.53	.31
Self-efficacy for the implementation process	-.70	.16	.50	.31
Self-efficacy for evaluation	-.76	.16	.09	.31
Scale total score	-.57	.16	.08	.31

SE = Standart error

The study used an independent samples t-test to compare the scores obtained from the Scale of Self-Efficacy for Teaching with Games for Special Education Teachers according to the gender variable. One-way analysis of variance (ANOVA) was applied to compare the scores obtained from the scale according to the variables of professional seniority, the field of graduation, and the disability group. The Scheffe post hoc test was used to determine which groups the differences observed in the ANOVA were dependent on. The Scheffe test is used when the number of individuals in the groups is different, and the variances are homogeneous (Kayri, 2009). Some groups with a small number of participants were combined with other groups for analysis. A significance level of p <.05 was considered significant for the analyses. IBM SPSS 26.0 statistical package program was used for the analyses.

3. FINDINGS

Firstly, the levels of self-efficacy for teaching with games for special education teachers were examined according to the participants' scores. In the next step, the levels of self-efficacy for teaching with games for special education teachers were compared and examined according to the variables of gender, professional seniority, undergraduate graduation field, and the disability group worked with.

Table 7. *Descriptive statistics for scores obtained from the self-efficacy scale for teaching through play for teachers working in special education.*

Variables	Min.	Max.	Mean	Average / number of items*	*SD*
Self-efficacy for planning instruction appropriate to program stages	8	15	13.03	4.34	1.99
Self-efficacy for planning instruction according to developmental characteristics	10	15	14.07	4.69	1.34
Self-efficacy for application method	6	15	10.50	3.50	2.14
Self-efficacy for the implementation process	13	30	25.41	4.24	3.67
Self-efficacy for evaluation	7	15	12.85	4.28	1.92
Scale total score	49	90	75.83	4.21	8.76

*1.00-1.80 very low, 1.81-2.60 low, 2.61-3.40 medium, 3.41-4.20 high, 4.21-5.00 very high

When Table 7 is examined, it is understood that the mean scores for self-efficacy in planning instruction according to program stages, self-efficacy in planning instruction according to developmental characteristics, self-efficacy in application method, self-efficacy in application process, self-efficacy in evaluation, and total scale scores are calculated as 13.03 (*SD*=1.99), 14.07 (*SD*=1.34), 10.50 (*SD*=2.14), 25.41 (*SD*=3.67), 12.85 (*SD*=1.92), and 75.83 (*SD*=8.76), respectively. According to the obtained mean scores, it is understood that the self-efficacy perceptions of the special education teachers participating in the research regarding planning instruction according to program stages, planning instruction according to developmental characteristics, application process, and evaluation are at a very high level. The self-efficacy perceptions of the participating teachers regarding the application method are at a high level.

Table 8. *Self-Efficacy score means, standard deviations and independent groups t test results for the teaching through play process by gender.*

Variables	Gender	*n*	*Mean*	*SD*	*t*	*Df*	*p*
Self-efficacy for planning instruction appropriate to program stages	Woman	133	13.22	1.81	1.60	239	.11
	Male	108	12.81	2.19			
Self-efficacy for planning instruction according to developmental characteristics	Woman	133	14.33	1.20	3.42	239	.00*
	Male	108	13.75	1.44			
Self-efficacy for application method	Woman	133	10.57	2.02	.56	239	.58
	Male	108	10.42	2.28			
Self-efficacy for the implementation process	Woman	133	25.98	3.12	2.73	239	.01*
	Male	108	24.70	4.16			
Self-efficacy for evaluation	Woman	133	13.17	1.73	2.90	239	.00*
	Male	108	12.46	2.07			
Scale total score	Woman	133	77.25	7.61	2.82	239	.01*
	Male	108	74.09	9.75			

*p<.05

When examining Table 8, it is understood that there is no significant difference in the mean scores of self-efficacy for planning instruction according to program stages ($t_{(239)}$=1.60; p>.05) and self-efficacy for instructional methods ($t_{(239)}$=.56; p>.05) based on gender. However, there is a significant difference in the mean scores of self-efficacy for planning instruction according to developmental characteristics ($t_{(239)}$=3.42; p<.05), self-efficacy for instructional processes ($t_{(239)}$=2.73; p<.05), self-efficacy for evaluation ($t_{(239)}$=2.90; p<.05), and total scale scores ($t_{(239)}$=2.82; p<.05) based on gender. Female teachers had significantly higher mean scores in self-efficacy for planning instruction according to developmental characteristics, self-efficacy for instructional processes, self-efficacy for evaluation, and total scale scores.

Table 9. *Self-Efficacy score means, standard deviations, and ANOVA results for the teaching through play process by professional seniority.*

Variables	Professional seniority	*n*	*Mean*	*SD*	*F*	*p*
Self-efficacy for planning instruction appropriate to program stages	0-10 years	140	12.95	1.96	2.86	.06
	11-20 years	68	12.84	2.09		
	20 years and above	33	13.79	1.78		
Self-efficacy for planning instruction according to developmental characteristics	0-10 years	140	14.01	1.36	.35	.71
	11-20 years	68	14,12	1.41		
	20 years and above	33	14.21	1.11		
Self-efficacy for application method	0-10 years	140	10.54	2.14	.06	.94
	11-20 years	68	10.43	2.13		
	20 years and above	33	10.52	2.18		
Self-efficacy for the implementation process	0-10 years	140	25.37	3.88	.66	.52
	11-20 years	68	25,18	3.47		
	20 years and above	33	26.06	3.13		
Self-efficacy for evaluation	0-10 years	140	12.69	1.94	1.19	.30
	11-20 years	68	13.07	1.90		
	20 years and above	33	13.09	1.83		
Scale total score	0-10 years	140	75.54	8.95	.81	.45
	11-20 years	68	75.57	9.14		
	20 years and above	33	77.64	6.96		

When examining Table 9, it is understood that there is no significant difference in the mean scores of self-efficacy for planning instruction according to program stages ($F_{2;240}$=2.86; p>.05), self-efficacy for planning instruction according to developmental characteristics ($F_{2;240}$=.35; p>.05), self-efficacy for instructional methods ($F_{2;240}$=.06; p>.05), self-efficacy for instructional processes ($F_{2;240}$=.66; p>.05), self-efficacy for evaluation ($F_{2;240}$=1.19; p>.05), and total scale scores ($F_{2;240}$=.81; p>.05) based on years of professional experience. It is understood that the perception of self-efficacy for the use of games in the teaching process is similar among teachers with 0-10 years, 11-20 years, and 20 years and above of professional experience.

When examining Table 10, it is understood that there is no significant difference in the mean scores of self-efficacy for instructional methods ($F_{2;240}$=2.60; p>.05), self-efficacy for instructional processes ($F_{2;240}$=1.40; p>.05), self-efficacy for evaluation ($F_{2;240}$=.74; p>.05), and total scale scores ($F_{2;240}$=2.42; p>.05) based on undergraduate graduation field. However, it is observed that there is a significant difference in the mean scores of self-efficacy for planning instruction according to program stages ($F_{2;240}$=3.63; p<.05) and self-efficacy for planning instruction according to developmental characteristics ($F_{2;240}$=3.58; p<.05) based on the field of graduation. According to the results, teachers who graduated from preschool teaching have significantly higher mean scores in self-efficacy for planning instruction according to program stages and self-efficacy for planning instruction according to developmental characteristics compared to teachers who graduated from special education teaching and classroom teaching.

Table 10. *Self-Efficacy score means, standard deviations, and ANOVA results for the teaching through play process by field of graduation.*

Variables	Undergraduate Graduation	*n*	*Mean*	*SD*	*F*	*p*	Post - Hoc
Self-efficacy for planning instruction appropriate to program stages	Special Education Teaching [a]	151	12.84	2.00	3.63	.03*	c >a, c >b,
	Classroom Teaching [b]	51	13.02	2.15			
	Preschool Teaching [c]	39	13.79	1.58			
Self-efficacy for planning instruction according to developmental characteristics	Special Education Teaching [a]	151	13.98	1.37	3.58	.03*	c >a, c >b,
	Classroom Teaching [b]	51	13.94	1.52			
	Preschool Teaching [c]	39	14.59	.79			
Self-efficacy for application method	Special Education Teaching	151	10.30	2.00	2.60	.08	-
	Classroom teaching	51	10.61	2.38			
	Pre-school teaching	39	11,15	2.22			
Self-efficacy for the implementation process	Special Education Teaching	151	25,26	3.75	1.40	.25	-
	Classroom teaching	51	25.18	3.58			
	Pre-school teaching	39	26.31	3.40			
Self-efficacy for evaluation	Special Education Teaching	151	12.96	1.91	.74	.48	-
	Classroom teaching	51	12.59	2.09			
	Pre-school teaching	39	12.79	1.70			
Scale total score	Special Education Teaching	151	75.30	8.83	2.42	.09	-
	Classroom teaching	51	75.27	9.73			
	Pre-school teaching	39	78.64	6.51			

[h] Scheffe Test, $*p<.05$

When examining Table 11, it is understood that there is no significant difference in the mean scores of self-efficacy for planning instruction according to program stages ($F_{3;240}$=.34; *p*>.05), self-efficacy for planning instruction according to developmental characteristics ($F_{3;240}$=.82; *p*>.05), self-efficacy for instructional methods ($F_{3;240}$=2.28; *p*>.05), self-efficacy for instructional processes ($F_{3;240}$=.88; *p*>.05), self-efficacy for evaluation ($F_{3;240}$=.96; *p*>.05), and total scale scores ($F_{3;240}$=1.03; *p*>.05) based on the disability group. It is understood that teachers working with mildly intellectually disabled, moderately to severely intellectually disabled, autism, hearing impaired, and visually impaired groups have similar perceptions of self-efficacy for the use of play-based teaching methods.

Table 11. *Self-efficacy score means, standard deviations, and ANOVA results for the teaching through play process by disability group.*

Variables	Disability group studied	*n*	*Mean*	*SD*	*F*	*p*
Self-efficacy for planning instruction appropriate to program stages	Mildly Mentally Disabled	73	12.88	2.12	.34	.80
	Medium-Severe Mentally Disabled	82	13.10	2.02		
	Autism	75	13.05	1.90		
	Hearing & Visually Impaired	11th	13.45	1.69		
Self-efficacy for planning instruction according to developmental characteristics	Mildly Mentally Disabled	73	13.92	1.48	.82	.48
	Medium-Severe Mentally Disabled	82	14.10	1.40		
	Autism	75	14.12	1.17		
	Hearing & Visually Impaired	11th	14.55	.93		
Self-efficacy for application method	Mildly Mentally Disabled	73	10.12	1.89	2.28	.08
	Medium-Severe Mentally Disabled	82	10.61	2.29		
	Autism	75	10.56	2.15		
	Hearing & Visually Impaired	11th	11.82	2.04		
Self-efficacy for the implementation process	Mildly Mentally Disabled	73	24.88	3.77	.88	.45
	Medium-Severe Mentally Disabled	82	25.79	3.52		
	Autism	75	25.44	3.67		
	Hearing & Visually Impaired	11th	25.91	4.16		
Self-efficacy for evaluation	Mildly Mentally Disabled	73	12.66	1.88	.96	.41
	Medium-Severe Mentally Disabled	82	12.90	1.97		
	Autism	75	13.08	1.82		
	Hearing & Visually Impaired	11th	12.27	2.37		
Scale total score	Mildly Mentally Disabled	73	74.42	8.66	1.03	.38
	Medium-Severe Mentally Disabled	82	76.43	9.00		
	Autism	75	76.24	8.46		
	Hearing & Visually Impaired	11th	78.00	9.59		

4. DISCUSSION and CONCLUSION

According to the results of the study determining the self-efficacy levels of teachers working in special education in the play-based teaching process, they consider themselves highly competent in planning instruction, implementing instructional processes, and evaluating instructional activities. The self-efficacy perceptions of the teachers participating in the study regarding instructional methods are also high. The study by Kadim (2012) on the self-efficacy beliefs of preschool teachers in the preschool education program supports our study as it shows that teachers' self-efficacy perceptions regarding planning are high. Guo, *et al.* (2014) found high self-efficacy perceptions among preschool special education teachers in their studies, which is similar to our study. The study by Piştav Akmeşe and Kayhan (2017) examined the self-efficacy of special education teachers in game-based teaching and found that their self-efficacy perceptions regarding planning were very high, supporting our study. In Celep's (2020) study, which examined the levels of play-teaching self-efficacy and creative personality traits of teachers working in preschool special education schools and the relationship between them, it was found that teachers had high levels of self-efficacy related to play. This finding is similar to the results of our study. According to the findings of the study on the self-efficacy levels of teachers in the game-based teaching process based on the gender variable, it is understood that there is no significant difference in the mean scores of self-efficacy for planning instruction according to program stages and self-efficacy for instructional methods based on gender. However, it was found that the responses provided by female teachers had higher average self-efficacy scores in terms of developmental characteristics, planning instruction, implementation process, and evaluation. Koç's (2015) study on the self-efficacy beliefs of preschool teachers in activities in the preschool education program found a significant difference in favor of female teachers, supporting our study. In the study by Piştav Akmeşe and Kayhan (2017), which examined the self-efficacy of special education teachers in play-based teaching, it was found that in terms of the gender variable, female teachers had a higher self-efficacy in the application dimension of game-based teaching, which supports our study. Tortop and Ocak (2010) examined the opinions of classroom teachers on educational game applications and found that contrary to our study, male teachers were found to be competent in educational game activities. They concluded that this situation was in parallel with doing sports in educational games and that male teachers do more sports than female teachers, so they are more competent in educational game activities.

In the study on the self-efficacy levels of teachers in the play-based teaching process based on professional seniority, it was found that teachers with 1-10 years, 11-20 years, and over 20 years of professional seniority have similar self-efficacy perceptions for game-based teaching. Other studies have also shown no significant difference between competence and professional seniority (Dickey, 2017; Semerci & Uyanık Balat, 2008).

Similarly, in a study determining the level of self-efficacy in the process of teaching through play among teachers working in special education, based on the variable of undergraduate degree, it was found that teachers who specified their undergraduate degree as preschool education had significantly higher mean self-efficacy scores in planning appropriate instruction and planning instruction according to developmental characteristics compared to teachers who specified their undergraduate degree as special education or elementary education. Our study is similar to the study by Piştav Akmeşe and Kayhan (2017), which examined the self-efficacy of special education teachers in teaching through play. In their study they found that teachers with a degree in preschool education had higher self-efficacy in teaching through play compared to teachers with degrees in special education or elementary education. It can be considered that the higher self-efficacy of special education teachers who graduated from preschool education in planning instruction according to program stages, planning instruction according to developmental characteristics, application method, application process, and evaluation compared to teachers who graduated in hearing impairment, visual impairment, mental disabilities, and classroom teaching is due to the content of the courses on play and play-based teaching in the preschool education undergraduate program. The education of children in the preschool period in terms of cognitive, social, physical, and language development, the implementation of these educations with on-the-spot observation, and the fact that special education teachers who graduated in preschool education have higher self-efficacy scores in teaching through play.

When looking at the teachers working in special education, it is understood that the self-efficacy perceptions of teachers working with mild intellectual disabilities, moderate-severe intellectual disabilities, autism, hearing impairment, and visual impairment are similar in the teaching through play process. Kaner *et al*. (2007) state that the presence or absence of disabilities among students does not cause differences in teachers' beliefs in professional competence, which supports our research. Other studies show no difference in self-efficacy between disability groups (Cantimer, 2015; Kaner, 2010). In light of the findings of the study, it is evident that teachers' self-efficacy in teaching through play is influenced by many factors. Considering that the majority of special education children are likely to be in the play stage, it is recommended to minimize factors that may hinder special education teachers' competence in teaching through play, increase the number of university-level courses on teaching through play, and provide in-service training.

Acknowledgments

This research is derived from the Master's Thesis titled Self-Efficacy of Special Education Teachers in Game Teaching Process (2021) at Necmettin Erbakan University, Institute of Educational Sciences, Department of Special 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**: Necmettin Erbakan University Social and Human Sciences Scientific Research Ethics Committee, 2021/164 on 19/03/2021.

Contribution of Authors

Nejmi Yıldırım: Investigation, Visualization, Software, Formal Analysis, and Writing-original draft. **Şerife Şenay İlik**: Methodology, Resources, Supervision and Validation.

Orcid

Nejmi Yıldırım https://orcid.org/0000-0002-8911-8069
Şerife Şenay İlik https://orcid.org/0000-0001-7092-379X

REFERENCES

Akandere, M. (2003). *Eğitici okul oyunları* [*Educational school games*]. Nobel Publishing.

Brodin, J. (1999). Play in children with severe multiple disabilities: play with toys a review. *International Journal of Disability, Development and Education, 46*(1), 25-34. https://doi.org/10.1080/103491299100704

Bollen, K.A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research, 17*(3), 303316. https://doi.org/10.1177/0049124189017003004

Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. *Sage Focus Editions*, 154, 136-136.

Byrne, B. (1989). *A Primer of LISREL, Basic Assumptions and Programming for Confirmatory Factor Analysis Models*. Springer.

Büyüköztürk, Ş. (2015). *Sosyal bilimler için veri analizi el kitabı (Genişletilmiş 21. Baskı)* [*Manual de análisis de datos para las ciencias sociales (21ª edición ampliada]).* Pegem Akademi.

Cano, A.R., García-Tejedor, A.J., Alonso-Fernández, C., & Fernández-Manjón, B. (2019). Game analytics evidence-based evaluation of a learning game for intellectual disabled users. *IEEE Access, 7*, 123820-123829. https://doi.org/10.1109/ACCESS.2019.2938365

Cantimer, G.G. (2015). *Özel eğitim gereksinimli çocukların öğretmenlerinin mesleki ve matematik öğretim özyeterlilik algılarının belirlenmesi* [*determination of vocational and mathematics teaching self-efficacy perceptions of teachers of children with special education needs*] [Unpublished Master's Thesis]. Marmara University. https://doi.org/10.14687/jhs.v14i1.3995

Celep, M.U. (2020). *Okul öncesi özel eğitimde çalışan öğretmenlerin oyun öğretimine ilişkin öz yeterlikleri ve yaratıcı kişilik özelliklerinin incelenmesi* [*Investigation of teachers' self-efficacy related to game teaching and creative personality characteristics of teachers working in preschool special education*] [Unpublished Master's Thesis]. Marmara

University.

DeVellis, R.F. (2017). *Scale development: theory and applications (4th ed.).* Sage.

Dickey, A. (2017). *Exploring the relationship between special education administrative support and the self-efficacy of special education teachers* [Unpublished PhD Thesis]. California State University.

Ergin, G. (2017). *Özel eğitimde oyun ve müzik* [*Play and music in special education*]. Pegem Publications.

Field, A. (2013). *Discovering Statistics Using IBM SPSS (4th ed.).* Sage Publications.

Fridenson-Hayo, S., Berggren, S., Lassalle, A., Tal, S., Pigat, D., Meir-Goren, N., … Golan, O. (2017). 'Emotiplay': a serious game for learning about emotions in children with autism: results of a cross-cultural evaluation. *Eur Child Adolesc Psychiatry, 26*, 979–992. https://doi.org/10.1007/s00787-017-0968-0

Guo, Y., Dynia, J.M., Pelatti, C.Y., & Justice, L.M. (2014). Self-efficacy of early childhood special education teachers: Links to classroom quality and children's learning for children with language impairment. *Teaching and Teacher Education*, *39*, 12-21. https://doi.org/10.1016/j.tate.2013.11.005

Güneş, F. (2015). Oyunla öğrenme yaklaşımı [Game learning approach]. *Electronic Turkish Studies*, *10*(11).

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

Janson, U. (2001). Togetherness and diversity in pre-school play international journal of early years education. *Carfax Publishing, 9*(2), 135-143. https://doi.org/10.1080/713670687

Jeong, H., Park, M., Choe, M.S., Kwon, H.J., Sung, J.H. (2020). Development of multi-experience AR board game 'ZOOCUS' for intellectual disabled students. *Journal of Korea Game Society, 20*(1), 121-132.

Jöreskog, K.G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the simplis command language*. Scientific Software International, Inc.

Kadim, M. (2012). *Okul öncesi öğretmenlerini oyun öğretimine ilişkin öz yeterliliklerinin incelenmesi* [*Investigation of self-efficacy of preschool teachers regarding play teaching*] [Unpublished Master's Thesis]. Abant İzzet Baysal University.

Kaner, S. (2010). Özel gereksinimli olan ve olmayan öğrencilerin öğretmenlerinin öz-yetkinlik inançları [Self-efficacy beliefs of teachers of students with and without special needs]. *Journal of Ankara University Faculty of Educational Sciences, 43*(1), 193-217.

Kaner, S., Şekercioğlu, G., & Yellice-Yüksel, B. (2007). Öğretmenlerin ve anababaların öz-yetkinlik inançları, tükenmişlik algıları ve çocukların problem davranışları [Teachers' and parents' self-efficacy beliefs, burnout perceptions and children's problem behaviors]. *Ankara University Scientific Research Projects.*

Kaplan, B. (2019*). Zihin yetersizliği olan öğrencilere sayma becerilerinim kazandırılmasında oyunla öğretimin etkililiği* [*The effectiveness of game instruction in teaching counting skills to students with mental disabilities*] [Unpublished Master's Thesis]. Necmettin Erbakan University.

Kaptan, S. (2018). *Otizm spektrum bozukluğu olan çocuklara video modelle öğretim yöntemiyle sosyodramatik oyun öğretimi* [*Sociodramatic game teaching to children with autism spectrum disorder by video modelle teaching method*] [Unpublished Master's Thesis]. Abant Izzet Baysal University.

Karasar, N. (2009). *Bilimsel araştırma yöntemi* [*Scientific research method*]. Nobel Yayın Dağıtım.

Kaya, A. (2010). *Oyun müdahale programının 3-5 yaş arasındaki özel gereksinimli çocukların bilişsel becerilerinin desteklenmesindeki etkililiğinin incelenmesi* [*Investigation of the effectiveness of the play intervention program in supporting cognitive skills of 3-5 year old children with special needs*] [Unpublished Master's Thesis]. Ankara University.

Kayri, M. (2009). *Araştırmalarda gruplar arası farkın belirlenmesine yönelik çoklu karşılaştırma (post-hoc) teknikleri* [*Multiple comparison (post-hoc) techniques for determining the difference between groups in research*]. *Fırat University Journal of Social Sciences, 19*(1), 51-64.

Koç, F. (2015). *Okul öncesi öğretmenlerinin okul öncesi eğitim programındaki etkinliklere yönelik öz-yeterlik inançlarının incelenmesi* [*Investigation of preschool teachers' self-efficacy beliefs towards activities in preschool education program*] [Unpublished Master's Thesis]. Van Yüzüncü Yıl University.

Manwarin, J.S. (2011). *High stakes play: "early childhood special educators" perspectives of play in pre-kindergarten classrooms* [Unpublished PhD Thesis]. Florida: University of South Florida.

MEB. (2014). *Özel eğitimde oyun etkinlikleri* [*Play activities in special education*]. Child Development and Education.

Metin, N., Şahin, S., & Şanlı, E. (1999). Okul öncesi düzeyde ve dört -dokuz yaş grubundaki zihinsel engelli çocukların tercih ettikleri oyun köşeleri ve oynadıkları oyun tiplerinin incelenmesi [Investigation of the preferred play corners and the types of games played by mentally disabled children in preschool level and four-nine age group]. *Journal of Special Education, 2*(3), 14-24.

Özdamar, K. (2004). *Paket programlar ile istatistiksel veri analizi (çok değişkenli analizler)* [*Statistical data analysis with package programs (multivariate analysis)*]. Kaan Kitabevi.

Pehlivan, H. (2014). *Oyun ve Öğrenme* [*Play and Learning*]. Anı Publishing.

Piştav-Akmeşe, P., & Kayhan, N. (2017). Özel eğitim öğretmenlerinin oyun öğretimine ilişkin öz-yeterlik düzeylerinin incelenmesi [Investigation of special education teachers' self-efficacy levels regarding play teaching]. *Ankara University Faculty of Educational Sciences Journal of Special Education, 18*(1), 1-26. https://doi.org/10.21565/ozelegitimdergisi.274303

Sarı, Y.D. (2017). *Erken çocuklukta oyun ve oyun yolu ile öğrenme* [*Play in early childhood and learning through play*]. Nobel Publishing.

Semerci, D., & Uyanık Balat, G. (2008). Okul Öncesi Öğretmenlerinin Sınıf Yönetimi Becerileri ve Öz Yeterlik Algıları Arasındaki İlişkinin İncelenmesi [Investigation of the Relationship Between Preschool Teachers' Classroom Management Skills and Self-Efficacy Perceptions]. *Journal of Inonu University Faculty of Education, 19*(3), 494-519.

Stagnitti, K., & O'Connor, C. (2012). Impact of the Learn to Play program on play, social competence and language for children aged 5–8 years who attend a specialist school. *Australian Occupational Therapy Journal*, 302-311. https://doi.org/10.1111/j.1440-1630.2012.01018.x

Stanley, P. (2003). *Security of attachment and symbolic play: a correlaction analysis of 3 to 5 year old children* [Unpublished Dissertation]. USA: California School of Professional Psychology.

Tabachnick, B.G., & Fidell, L.S. (2007). *Using multivariate statistics (5th ed.).* Boston: Allyn and Bacon.

Tanaka, J.S., & Huba, G.J. (1985). A fit index for covariance structure models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology, 38*(2), 197-201. https://doi.org/10.1111/j.2044-8317.1985.tb00834.x

Tavşancıl, E. (2018). *Tutumların ölçülmesi ve SPSS ile veri analizi* [*Measurement of attitudes and data analysis with SPSS*] *(5. Baskı).* Nobel Yayın Dağıtım.

Tezbaşaran, A. (2008). *Likert tipi ölçek geliştirme kılavuzu* [*Likert type scale development guide*]. Türk Psikologlar Derneği Yayınları.

Tortop, Y., & Ocak, Y. (2010). Sınıf öğretmenlerinin eğitsel oyun uygulamalarına yönelik görüşlerinin incelenmesi [Examining the opinions of classroom teachers on educational game applications]. *Spor ve Performans Araştırmaları Dergisi, 1*(1), 14-22. https://doi.org/10.9761/JASSS2388

Tuğrul, B. (2014). Oyunun üç kuşaktaki değişimi [The change of the game in three generations]. *International Journal of Social Science*, 1-16.

Tuzcuoğlu, N., & Tuzcuoğlu, S. (2004). *Dikkat Geliştiren Oyunlar* [*Attention Enhancing Games*]. Morpa.

Yaman, D. (2019). *Oyunun özel gereksinimli çocuklarin sosyal gelişimine etkisi hakkindaki öğretmen görüşleri* [*Teachers' views on the effect of play on the social development of children with special needs*] [Unpublished Master's Thesis]. Nicosia: Near East University Institute of Educational Sciences.

Yaşlıoğlu, M.M. (2017). Sosyal bilimlerde faktör analizi ve geçerlilik: Keşfedici ve doğrulayıcı faktör analizlerinin kullanılması [Factor analysis and validity in social sciences: Using exploratory and confirmatory factor analysis]. *Journal of Istanbul University Faculty of Business Administration, 46*, 74-85.

Yazıcıoğlu, Y., & Erdoğan, S. (2014). *SPSS Uygulamalı Bilimsel Araştırma Yöntemleri* [*SPSS Applied Scientific Research Methods*]. Detay Yayıncılık.

Yurt, E. (2023). *Sosyal bilimlerde çok değişkenli analizler için pratik bilgiler: SPSS ve AMOS uygulamaları [Practical Insights for Multivariate Analyses in Social Sciences: SPSS and AMOS Applications].* Ankara: Nobel

International Journal of Assessment Tools in Education
2025, Vol. 12, No. 1, 20–32

https://doi.org/10.21449/ijate.1517994

journal homepage: https://dergipark.org.tr/en/pub/ijate **_Research Article_**

Artificial intelligence as an automated essay scoring tool: A focus on ChatGPT

Ahmet Can Uyar[1*], Dilek Büyükahıska[2]

[1]Sivas Cumhuriyet University, School of Foreign Languages, Department of English, Sivas, Türkiye
[2]Ondokuz Mayıs University, Faculty of Education, Department of English Language Teaching, Samsun, Türkiye

ARTICLE HISTORY

Received: July 18, 2024

Accepted: Oct. 07, 2024

Keywords:
Automated essay scoring,
Artificial intelligence,
ChatGPT,
Foreign language writing,
Writing evaluation.

Abstract: This study explores the effectiveness of using ChatGPT, an Artificial Intelligence (AI) language model, as an Automated Essay Scoring (AES) tool for grading English as a Foreign Language (EFL) learners' essays. The corpus consists of 50 essays representing various types including analysis, compare and contrast, descriptive, narrative, and opinion essays written by 10 EFL learners at the B2 level. Human raters and ChatGPT (4o mini version) scored the essays using the International English Language Testing System (IELTS) TASK 2 Writing band descriptors. Adopting a quantitative approach, the Wilcoxon signed-rank tests and Spearman correlation tests were employed to compare the scores generated, revealing a significant difference between the two methods of scoring, with human raters assigning higher scores than ChatGPT. Similarly, significant differences with varying degrees were also evident for each of the various types of essays, suggesting that the genre of the essays was not a parameter affecting the agreement between human raters and ChatGPT. After all, it was discussed that while ChatGPT shows promise as an AES tool, the observed disparities suggest that it has not reached sufficient proficiency for practical use. The study emphasizes the need for improvements in AI language models to meet the nuanced nature of essay evaluation in EFL contexts.

1. INTRODUCTION

AI has become one of the indispensable parts of our everyday lives with new tools emerging each day whose functions range from advanced interaction with humans to image creation. In this context, educational settings are unsurprisingly being embellished with such tools for the purpose of enhancing the process of teaching. Not limited to the process of teaching itself, AI tools have also started to become a subject of educational assessment. In the realm of language learning and teaching, the evaluation and the assessment of written products in the target language stand as pivotal measures of linguistic prowess. Traditionally, writing evaluation has been predominantly conducted by the course instructors, drawing upon their expertise and understanding of language nuances, context, and cultural intricacies. However, the rapid advancements in AI have introduced a revolutionary shift in this landscape. AI-powered

*CONTACT: Ahmet Can UYAR ✉ ahmetcanuyar97@gmail.com Sivas Cumhuriyet University, School of Foreign Languages, Department of English, Sivas, Türkiye

e-ISSN: 2148-7456

systems, leveraging sophisticated algorithms and machine learning models, now offer an alternative or complementary method for assessing and grading EFL learners' written products.

The idea of evaluating essay writings based on a machine algorithm dates back to the 1960s, the decade when Page (1966) proclaimed that computer-based essay scoring parallel to that of human scoring was on the horizon. Page (1966) attempted to articulate what is recognized today as AES or automated writing evaluation (AWE), which can be described as the utilization of technology to assess and rate essays. It entails employing computer programs to examine and assign scores to written pieces by considering predetermined standards like language accuracy, vocabulary depth, logical flow, sentence structure, and meaningful connection. Although this technology is becoming more advanced in assessing the meaningful discourse of written works, initial AES systems faced criticism for concentrating solely on superficial aspects while overlooking content-related characteristics, leaving them susceptible to cheating tactics by learners like padding with extra words and commas (Attali, 2013). While many AES systems have been documented to focus on analyzing surface-level linguistic and structural elements, human assessment of essays prioritizes different aspects of language use and discourse (Huang, 2014). However, along with the advancements in Natural Language Processing (NLP), Large Language Model (LLM) technologies, and related AI developments, it is probable to assert that these shortcomings started to fade, and a new era for AES has started rising.

According to Huang (2014), the rise of AES stems from two primary factors. Firstly, the immense burden on educators due to teaching demands, class instructions, and the substantial time and effort required for grading students' written work is notable, accounting for nearly 30% of their workload (Mason & Grove-Stephenson, 2002). With limited time and resources, instructors struggle to effectively assess compositions and provide feedback. Introducing AES systems into classrooms could alleviate this overwhelming workload by handling and evaluating writing assignments. Secondly, as Huang (2014) noted, AES offers a distinct advantage over human evaluation by ensuring consistency in scoring, as its criteria are programmed and executed uniformly. In contrast, human assessment is susceptible to inconsistency due to cognitive fatigue, distractions, and interruptions over time. Particularly in large-scale assessments like "Test of English as a Foreign Language" (TOEFL) or "Graduate Record Examination" (GRE), human subjectivity can lead to varying and unreliable ratings.

These factors have been discussed to be overcome by developing AES technologies. Therefore, in recent years, the utilization of AES and AWE systems has garnered significant attention within educational settings for assessing written compositions, particularly among EFL learners. Numerous studies (Almusharraf & Alotaibi, 2022; Chen & Pan, 2022; Huang, 2014; Manap *et al.*, 2019; Wang & Bai, 202; Zribi & Smaoi, 20211) have explored the effectiveness of various AES and AWE systems in evaluating writing proficiency, highlighting their strengths and limitations. These studies predominantly focused on employing distinct software tools such as Criterion, Grammarly, PaperRater, and other automated assessment platforms to assess writing quality. Despite the rich literature on this subject, one notable area that remains relatively unexplored is the utilization of LLMs, such as ChatGPT, in the domain of essay grading for EFL learners. Notably, while the existing research has examined the efficacy of different software tools in automated grading, the assessment potential of ChatGPT, an advanced language generation model created by OpenAI, remains relatively uninvestigated (Bui & Barrot, 2024). Therefore, this study represents an earnest effort to bridge this gap by investigating the feasibility and effectiveness of ChatGPT in evaluating EFL learners' essays, thereby contributing to the existing literature by exploring a novel avenue in AES. In this vein, a comparison was pursued in order to reveal the (un)parallelism between the scores generated by human raters and ChatGPT for the essays written by EFL learners. The investigation was shaped by the subsequent research questions:

1. Is there a significant difference between the scores generated by human raters and ChatGPT?

2. Does the genre of the essay play a significant role in the agreement between the scores produced by human raters and ChatGPT?

1.1. Automated Essay Scoring

AES systems have gained attention in education, particularly in evaluating writing proficiency for EFL learners. Much of the research effort was built on specific AES tools which were specifically designed for this pursuit, rather than LLMs or NLP models like ChatGPT. However, many of these tools were argued to fall short of adequately corresponding to the complex nature of essay scoring. This is because these specific tools were only capable of recognizing the essays from a mechanical or rule-based perspective. In line with this, studies have revealed both the benefits and limitations of these tools. For instance, Huang (2014) examined the AES tool Criterion in an EFL context, finding a weak correlation between AES and human scoring, with Criterion often assigning higher scores. The study showed that the tool focused more on language mechanics, while human raters emphasized discourse and writing quality. Huang (2014) concluded that the AES tool Criterion, though efficient, may not capture the nuanced elements of writing as well as humans.

Similarly, Manap *et al*. (2019) compared PaperRater with human evaluation and found that this AES tool was more lenient, showing a moderate correlation with human scores. Despite its utility in providing quick feedback, PaperRater's ability to assess deeper content and relevance was limited. Zribi and Smaoi (2021) echoed this finding, noting that PaperRater assigned consistently higher scores than human raters for intermediate-level EFL learners, raising concerns about its reliability. A similar finding in that the AES tool assigned higher scores than human raters was evident in Chen and Pan (2022), who explored the effectiveness of Aim Writing in improving Chinese college students' English writing skills. While the tool helped with grammar and vocabulary, human feedback was more related to structure and organization. The study found a notable correlation between the tool and human scores, although Aim Writing consistently assigned higher scores.

On the other hand, several studies revealed that certain AES tools assigned scores lower than those of human raters, a finding that is in contrast with the body of research aforementioned. Namely, Almusharraf and Alotaibi (2022) evaluated Grammarly's effectiveness compared to human raters, focusing on writing errors in 197 EFL essays. While Grammarly detected more errors, human raters assigned higher scores overall. The study highlighted that the tool excelled at catching specific grammar issues but struggled with more complex writing aspects such as sentence flow and coherence. The authors recommended using the tool as a supplementary tool rather than a standalone solution. A similar finding in the context of high-quality essays was published by Wang and Bai (2021), who assessed the accuracy of two AES systems, Pigai and iWrite, using 486 essays from non-English majors. According to the findings, both systems agreed with each other but differed significantly from human raters, especially for high-quality essays, in which AES systems tended to score lower. This suggests that while AES systems can handle lower-quality writing better, they struggle with more advanced writing, emphasizing the need for further refinement.

While specific tools designed to serve as AES systems in the market still attract considerable attention, the shift recently geared towards the latest developments in AI. That is to say, AI-based chatbots which are becoming more and more capable of producing human-like speech day by day are the recent subjects of AES. One of these renowned tools which are in public use with its user-friendly interface is ChatGPT.

1.2. ChatGPT as an AES Tool

ChatGPT, which stands for chat generative pretrained transformer, is an AI-powered LLM tool developed by OpenAI (accessible at https://chat.openai.com). It helps computers understand and generate text that resembles human speech. Because ChatGPT was released relatively recently, studies that take it under focus in the context of AES are considerably restricted in number (Bui & Barrot, 2024).

OpenAI occasionally releases a new version of ChatGPT, which creates a panorama of different versions such as ChatGPT-3, ChatGPT-3.5, or ChatGPT-4. Besides, it is also possible for it to be used in connection to other tools or with modifications. Therefore, the existing literature presents an amalgam of studies that utilized various versions or variations of ChatGPT. For instance, Mizumoto and Eguchi (2023) examined the use of OpenAI's text-davinci-003 model, part of GPT-3.5, as an AES tool. Using the TOEFL11 corpus, which includes 12.100 essays, they assessed the model's accuracy and reliability, particularly when linguistic features like lexical diversity and syntactic complexity were added. Statistical analyses showed that including these features significantly improved essay scoring accuracy. While the ChatGPT model demonstrated a certain level of reliability, the study concluded that AI-based AES should support human raters rather than replace them. The authors also highlighted the need for further research into ethical concerns and student motivation in AI-based assessments.

Yancey *et al*. (2023) explored the use of LLMs, particularly ChatGPT-3.5 and ChatGPT-4, for AES in high-stakes English language tests for language learners. Using short essay responses from the Duolingo English Test which were scored by human raters, the study aimed to compare the performance of ChatGPT models to that of existing AWE systems and assess inter-rater agreement between the models and human raters. Results showed that ChatGPT-4, when given calibration examples, closely matched the performance of current AWE systems. However, the model's agreement with human ratings varied based on the test-takers' first language. The authors highlighted potential of ChatGPT-4 to enhance AWE systems, while stressing the need for careful consideration of fairness and ethical implications. Parker *et al*. (2023) examined the potential of ChatGPT-3 as an AWE tool in nursing education, focusing on providing formative feedback on scholarly writing. The study analyzed 42 graduate nursing students' papers, where ChatGPT-3 evaluated macro-level elements such as organization, development, and thesis clarity. The authors emphasized the importance of using well-constructed prompts to obtain useful feedback from generative AI tools. ChatGPT-3 was found to grade more strictly than a human rater, awarding only one paper a score of 3 (grade A) and giving a score of 2 (grade B) to the rest. It provided detailed feedback with suggestions for improvement in areas like evidence, organization, and mechanics. The authors highlighted the efficiency and individualized feedback that AI tools offer, promoting autonomous learning. The researchers emphasized the need for further research on using ChatGPT effectively in writing instruction and the importance of educating faculty and students on its implementation. The study concluded that ChatGPT holds promise for enhancing writing feedback in nursing education.

Guo and Wang (2024) investigated the potential of ChatGPT (the version was not explicitly stated in the study; however, the date of the data collection corresponds to ChatGPT-3) to assist EFL teachers in providing feedback on learners' argumentative essays. The study involved 50 essays written by Chinese undergraduate students with B2 to C1 English proficiency. Both ChatGPT and five Chinese EFL teachers, varying in experience and technology use, evaluated the essays, focusing on content, organization, and language. ChatGPT provided more feedback than the teachers, offering balanced attention to all aspects, while teachers varied in their focus. ChatGPT's feedback also included summaries that could help learners understand their writing globally, which might aid in revisions. However, ChatGPT occasionally provided off-task feedback, possibly due to its aspect-specific prompting, which was not observed in the teachers' feedback. The teachers generally rated ChatGPT's feedback positively, acknowledging its detailed and structured nature but also noting that its effectiveness depends on learners' ability to understand and apply it. The study's limitations included focusing only on argumentative essays, excluding student perspectives, and not accounting for the teachers' ability to provide more personalized feedback. The study concluded that while ChatGPT has the potential to support EFL writing instruction, its classroom integration requires careful consideration.

Bui and Barrot (2024) explored the use of ChatGPT-3.5 as an AES tool in writing classes. The researchers compared the scores generated by ChatGPT-3.5 with those assigned by a highly experienced human rater, who had 20 years of expertise in essay evaluation. The research analyzed 200 argumentative essays from college students, categorized by proficiency levels from A2 to B2. The results revealed that ChatGPT scores did not strongly correlate with those assigned by human raters, showing weak to moderate correlations and a lack of consistency, as indicated by low intraclass correlation coefficients. The authors suggested that the differences could be due to ChatGPT's scoring algorithm, the data it was trained on, changes in the model, and its inherent randomness. The study also highlighted some limitations, such as relying on a single human rater and not including qualitative feedback in the analysis. The researchers recommended future research that involves multiple raters, considers qualitative feedback, and investigates how student-level factors affect AES performance. Despite these challenges, the paper offers perspectives to ChatGPT's current capabilities as an AES tool and its potential for improvement.

In conclusion, while these studies provided valuable insights into the current capabilities of ChatGPT as an AES tool, they collectively highlighted several limitations and areas for further research. Certain scholars noted that future research should focus on exploring a broader range of essay types (Guo & Wang, 2024) and incorporating multiple human raters for comparison (Bui & Barrot, 2024). In parallel, the current study corresponds to these future research suggestions mentioned, incorporating different essay types and multiple human raters.

2. METHOD

2.1. Research Design

This study adopts a quantitative approach. Quantitative methodology is an approach used in scientific research that focuses on the collection and analysis of numerical data to answer research questions. It involves employing several strategies, techniques, and assumptions to examine various phenomena through the analysis of numerical patterns, enabling researchers to gather and analyze numeric data to conduct statistical analyses ranging from simple to complex, including aggregating data, revealing relationships, and making comparisons across aggregated datasets (Coghlan & Brydon-Miller, 2014). Creswell (2009) proposed that choosing the appropriate research design depends on the nature of the research questions of the study. Because this study seeks to compare the scores generated by two different parties from a statistical point of view, it utilizes a quantitative approach featuring the Wilcoxon signed-rank test and the Spearman correlation coefficient test.

2.2. Materials and Data Collection

A total of 50 EFL essays were collected from 10 learners (6 female and 4 male students) who studied at the B2-level English preparatory class at a state university in Türkiye. The participants and their essays were sampled based on the convenience sampling method. The essays were written by the learners as writing tasks in the context of a writing class without resorting to any type of AI tools. In line with the research aims, various types of essays were included in the sample. The compilation consists of 10 analysis essays, 10 compare and contrast essays, 10 descriptive essays, 10 narrative essays, and 10 opinion essays.

2.2.1. Instrument

The IELTS exam is a significant assessment of English language skills designed to provide proof that test-takers possess the linguistic abilities required by the test user for the specific language context in which they are expected to perform successfully (IELTS, 2019). Advanced to be used in this test to assess writing skills, this research utilized the "IELTS TASK 2 Writing band descriptors (public version)" (IELTS, 2023) as the scoring criteria for the essays. This tool evaluates written texts using four equally important analytical assessment criteria spread across nine performance levels, which include task response, coherence and cohesion, lexical resource, and grammatical range and accuracy. Citing Davies (2008), Pearson (2022) underscored that this feature demonstrates the way that the evaluation of IELTS writing adopts a theoretical approach centered around general proficiency, applicable to various academic domains. Furthermore, Mizumoto and Eguchi (2023) outlined that the rubric enables a comprehensive evaluation covering factors such as the handling of the task, coherence and cohesion, lexical proficiency, and grammatical variety and precision, all assessed on a 10-point scale (band) from 0 to 9. The researchers also noted that this rubric was more favorable over the other renowned rubrics, such as the 5-point scale found in the TOEFL iBT test independent writing rubrics, because the 10-point scale allows for a more detailed evaluation and facilitates a more refined distinction among scores. Taking these points into consideration, this rubric was found to be useful for the purposes of this study and it was chosen as the essay scoring instrument.

2.2.2. Human assessment

A group of three experienced EFL instructors were recruited to assess the collected essays based on the rubric selected. These instructors, experienced in teaching writing, had been working at a university-level preparatory program. One of them had five years of teaching writing experience, one of them had four years, and the other one had three years of experience in EFL contexts. The instructors followed the standardized assessment criteria of “IELTS TASK 2 Writing band descriptors (public version)” to score the essays. Three instructors rated each essay separately for the sake of inter-rater reliability. After the instructors scored the essays, the three scores generated for each essay were examined to determine if there was a score gap more than 20% (1.8 points) of the total score an essay could be assigned. For instance, if an essay was assigned 6, 8, and 9 by the instructors, the instructors were asked to re-evaluate that specific essay to ensure the inter-rater agreement and increase the reliability of the human scores. Out of the 50 essays, only 4 essays required this practice. After this process, the mean scores of the three human scores were calculated for each essay for further analysis.

2.2.3. ChatGPT assessment

A recent version of ChatGPT, ChatGPT-4o mini, was utilized to assess the same set of essays. After several releases of the model such as ChatGPT-3 and ChatGPT-3.5, an enhanced version, ChatGPT-4 is currently in public use, and it is fully available through a subscription service. However, a scaled-down version, ChatGPT-4o mini, is accessible for free to users who sign up for an account. Many people benefit from it because it is freely accessible, which is the reason why this version was selected to be under scrutiny in this study.

On a similar basis with human raters, the essays were scored by the ChatGPT-4o mini system based on the same rubric of “IELTS TASK 2 Writing band descriptors (public version)” for three times with a different prompt for each. Adapting the methodology of Mizumoto and Eguchi (2023, p. 5), the following prompt was typed in ChatGPT’s chat box to generate the first set of scores:

> I would like you to mark a/an [type of the essay] essay written by a B2-level of English as a foreign language learner. The prompt given to the learners for this essay task was [the prompt of the essay task]. The essay should be assigned a rating of 0 to 9, with 9 being the highest and 0 the lowest. You don’t have to explain why you assign that specific score. Just report a score only. The essay is scored based on the following rubric.
>
> [IELTS rubric in a plain text format.]
>
> ESSAY:
>
> [The essay in a plain text format.]

Following the generation of the first set of scores, paraphrased versions of the first prompt were used to re-evaluate the essays. The second prompt was as follows:

> I would like you to grade a/an [type of the essay] which was written by an English as a foreign language learner at B2 level. The instruction given to the learner was [the prompt of the essay task]. This essay should be given a score between 0 to 9 (0 as the lowest and the 9 as the highest). I would like only the score, not the reason why you assigned that score. Rate the essay based on the following rubric.
>
> [IELTS rubric in a plain text format.]
>
> ESSAY:
>
> [The essay in a plain text format.]

Finally, the third set of scores was generated by the subjection of the following prompt:

> I would like you to rate a/an [type of the essay] written by a B2-level English as a foreign language learner. The prompt given was: [the prompt of the essay task]. Assign a score between 0 and 9 (0 being the lowest, 9 being the highest). Only provide the score without an explanation. Use the following rubric to evaluate the essay.
>
> [IELTS rubric in a plain text format.]
>
> ESSAY:
>
> [The essay in a plain text format.]

These three different but very similar prompts were used for each essay to be assigned a score. After three scores were generated for each essay through these prompts, the scores were examined to determine if there was a considerable gap between the three scores (a score difference more than 20% or 1.8 points), a procedure which was also conducted with the human raters to increase the inter-rater agreement and the reliability of the scores. Eventually, no such instances were found. In fact, ChatGPT-4o mini assigned relatively consistent scores across the three assessments. Specifically, it assigned the same score for 31 essays in all three of the assessments. For the other 19 essays, the score differences were minor in that a re-evaluation was not regarded. Finally, the mean scores of the three ChatGPT scores were calculated for comparative analysis.

2.3. Data Analysis

The scores given by human raters and the ChatGPT system were compared to determine the degree of agreement or difference between the two assessment methods. The same set of essays was evaluated by both human raters and ChatGPT, therefore, the data were paired. In paired data, each data point in one group (human scores) corresponds directly to a data point in the other group (ChatGPT scores) because both sets of scores are for the same set of essays. In line with this, the Wilcoxon signed-rank test was operated to determine if there was a statistically significant difference between the two assessment outcomes. Willard (2020) pointed out that the Wilcoxon signed-rank test serves as an appropriate nonparametric alternative to the related samples t-test when parametric assumptions are not met, and it can be applied to matched pairs design. In the case of this study, the Wilcoxon signed-rank test was appropriate because of two main reasons: (1) the data consisted of matched pairs (each essay was assigned a score generated by both human raters and ChatGPT) and (2) the parametric assumptions of normality and variance for a paired samples t-test were not satisfied.

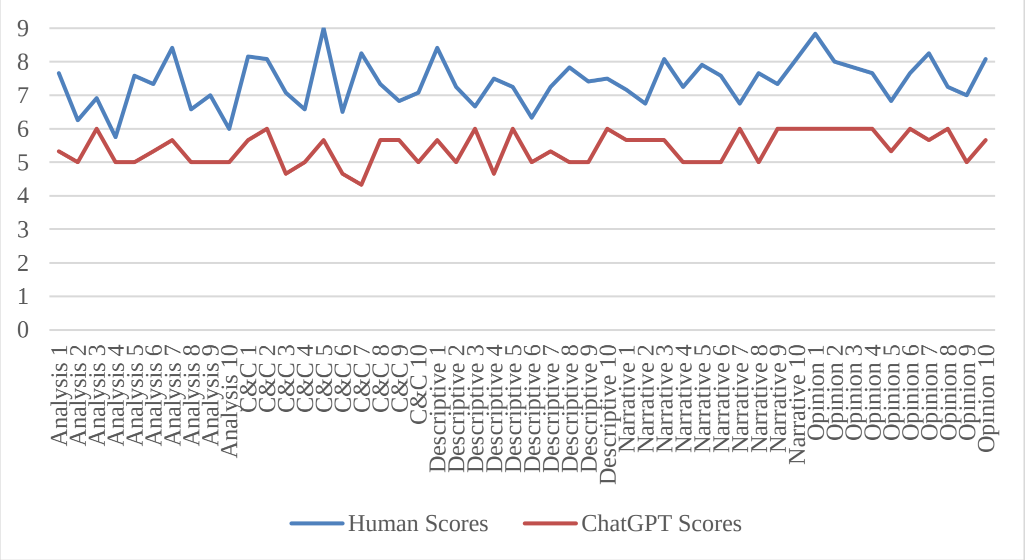
In addition to the Wilcoxon signed-rank test, the Spearman correlation coefficient was computed to reveal the direction of the relationship between the two groups of scores. Correlation tests are operated to reveal the level of the relationship between the data sets (Larson-Hall, 2012). The Spearman correlation coefficient, commonly referred to as Spearman’s rho, serves as the nonparametric alternative to the Pearson correlation coefficient (Willard, 2020). Because the current set of data violated the parametric test assumptions, the

Spearman correlation coefficient test was operated instead of the Pearson correlation test. Finally, the results of these statistical tests were presented and interpreted accordingly.

3. FINDINGS

The first research question aimed to determine if there is a significant difference between the scores generated by human raters and ChatGPT. Results of the Wilcoxon signed-rank test indicated that there is a significant difference between the human scores (Median = 7.3, N = 50) and the ChatGPT scores (Median = 5.5, N = 50), $Z = -6.1504$, $p < 0.001$, with a large effect size ($r = -0.8698$). This set of data indicates that human scores are significantly higher than those assigned by ChatGPT. This finding can be visually represented in Figure 1, where the scores generated by the human raters and ChatGPT for each essay are presented.

Figure 1. *The scores assigned by human raters and ChatGPT for each essay.*



The calculated Z value was -6.1504, and the corresponding p-value was less than 0.001. This result is statistically significant at the $p < 0.05$ level, suggesting that there is a statistically significant difference between the scores assigned by human raters and those assigned by ChatGPT. These findings indicate that the evaluations provided by humans significantly differ from those given by ChatGPT, with humans scoring higher on average. This highlights the potential discrepancy between human and ChatGPT scoring in essay evaluations. The effect size value (r) in the context of the Wilcoxon signed-rank test measures the strength of the relationship between the two samples being compared. In the realm of the current data, the large negative effect size ($r = -0.8698$) suggests that the difference in scores between the human raters and ChatGPT is not only statistically significant but also practically meaningful. This indicates a substantial gap in grading performance between the two scoring parties.

Additionally, the Spearman correlation coefficient test was operated to assess the direction of the relationship between the scores. The results yielded a Spearman correlation coefficient $r_s = 0.30493$, with a p-value of 0.0313. Respectively, the value of $r_s = 0.30493$ indicates a moderate positive correlation between the scores assigned by human raters and ChatGPT in a significant way ($p < 0.05$). This suggests that as the scores from human raters increase, the scores from ChatGPT tend to increase as well. While this is the case in that human and ChatGPT scores tend to rise together, the correlation is not strong enough to imply that they are interchangeable or that they assess the essays in the same manner. The correlation indicates that there is a tendency for higher human scores to align with higher ChatGPT scores, but it also reflects the variability in the scores assigned by both evaluators. Overall, it can be concluded that the findings suggest a significant difference in the evaluation of essays between human raters and ChatGPT.

The second research question aimed to reveal if the genre of the essay plays a significant role in the agreement between the scores produced by human raters and ChatGPT. To investigate whether the genre of the essay influences the agreement and correlation between the scores assigned by human raters and ChatGPT, the same set of tests was conducted for five distinct essay genres each of which contained ten essays: analysis essay, compare and contrast essay (the abbreviation C&C is used in Table 1), descriptive essay, narrative essay, and opinion essay. The results of the Wilcoxon signed-rank tests and the Spearman correlation coefficients for each genre are presented in Table 1.

Table 1. *Comparison of the human and ChatGPT scores based on the genre of the essay.*

	Wilcoxon signed-rank test					Spearman correlation	
	Human *(Median)*	ChatGPT *(Median)*	*Z*-value	*p*-value	Effect size (*r*)	r_s value	*p*-value
Analysis Essay	7	5	-2.7539	0.005	-0.8709	0.56891	0.08611
C&C Essay	7.2	5.3	-2.7539	0.005	-0.8709	0.27777	0.43713
Descriptive Essay	7.3	5.2	-3.0973	0.001	-0.9794	-0.12224	0.73656
Narrative Essay	7.5	5.7	-3.0973	0.001	-0.9794	-0.15504	0.66888
Opinion Essay	7.7	6	-2.7557	0.005	-0.8714	0.32313	0.36244

The Wilcoxon signed-rank test for the analysis essays revealed a significant difference between human and ChatGPT scores, with a *Z*-value of -2.7539 and a *p*-value of 0.005, indicating a statistically significant difference. The median human score (Median = 7) was higher than the median score assigned by ChatGPT (Median = 5), showing that human raters consistently rated the essays higher than ChatGPT. The effect size (r = -0.8709) suggests a strong difference between the two scoring systems. The Spearman correlation coefficient (r_s = 0.56891) indicates a moderate positive relationship between human and ChatGPT scores, but this relationship is not statistically significant (p = 0.08611). Although the scores tend to move in the same direction, this correlation is not strong enough to indicate substantial agreement between human and ChatGPT scores. In other words, while there is some alignment in the scoring, human raters still rated the analysis essays significantly higher than ChatGPT.

For the compare and contrast essays, the Wilcoxon signed-rank test again showed a significant difference between human and ChatGPT scores. The *Z*-value of -2.7539 and a *p*-value of 0.005 indicate a statistically significant disagreement between the two sets of scores. Human scores had a higher median (Median = 7.2) compared to ChatGPT (Median = 5.3), and the large effect size (r = -0.8709) suggests a strong disparity in how human raters and ChatGPT evaluated the essays. Furthermore, the Spearman correlation coefficient (r_s = 0.27777) indicates only a weak positive correlation between human and ChatGPT scores, and this correlation is not statistically significant (p = 0.43713). This result suggests that there is minimal alignment between human and ChatGPT scores for the compare and contrast essays.

Similarly, the test results for the descriptive essays revealed a significant difference between the human and ChatGPT scores, with a *Z*-value of -3.0973 and a *p*-value of 0.001. The median human score was 7.3, while the median ChatGPT score was considerably lower at 5.2. The effect size (r = -0.9794) is very large, indicating a very strong difference between the scores given by human raters and those assigned by ChatGPT. This suggests that ChatGPT systematically scored descriptive essays much lower than human raters did. The Spearman correlation coefficient ($r_s = -0.12224$) shows a very weak negative correlation between human and ChatGPT scores, and this relationship is not statistically significant (p = 0.73656). This result indicates almost no agreement between the scores given by human raters and those assigned by ChatGPT.

For narrative essays, the results again showed a significant difference between human and ChatGPT scores, with a *Z*-value of -3.0973 and a *p*-value of 0.001. The median human score was 7.5, while ChatGPT's median score was 5.7. The effect size (r = -0.9794) indicates a very large difference between the two scoring methods, suggesting that human raters consistently rated narrative essays higher than ChatGPT. The Spearman correlation coefficient (r_s = −0.15504) also indicates a weak negative correlation between human and ChatGPT scores, with no statistically significant relationship (p = 0.66888). This lack of significant correlation suggests that there is no meaningful alignment between human and ChatGPT scores in narrative essays. The results suggest that while human raters scored these essays higher, ChatGPT struggled to evaluate them in a similar way, further highlighting the differences in how the two systems interpret this genre.

The test results for opinion essays also revealed a significant difference between human and ChatGPT scores, with a *Z*-value of -2.7557 and a *p*-value of 0.005. The median human score was 7.7, while the median score assigned by ChatGPT was 6.0, indicating that human raters consistently assigned higher scores. The effect size (r = -0.8714) was large, showing a strong disagreement between human and ChatGPT scores. The Spearman correlation coefficient (r_s = 0.32313) revealed a weak positive correlation between human and ChatGPT scores, but this relationship was not statistically significant (p = 0.36244). This suggests that while there is a slight tendency for human and ChatGPT scores to increase together, the correlation is too weak to assert a meaningful agreement.

Overall, the results indicate that there is a statistically significant difference between human and ChatGPT scores across all essay types, with human raters consistently assigning higher scores than ChatGPT. The large effect sizes observed in all genres suggest that these differences are notable. ChatGPT appears to particularly struggle with more subjective essay types, such as descriptive and narrative essays, where the gap between human and ChatGPT scores is widest. On the other hand, the Spearman correlation results show that there is little to no meaningful correlation between human and ChatGPT scores across most essay types. The moderate positive correlation in the analysis essay suggests some degree of alignment, but the relationship is not statistically significant. For more subjective genres like descriptive and narrative essays, there is no agreement between the two scoring methods, with weak or even negative correlations observed. Eventually, it can be asserted that the genre of the essays does not play a significant role in the agreement between the scores generated by human raters and ChatGPT.

4. DISCUSSION

The findings from this study revealed significant insights into the capabilities and limitations of ChatGPT, or specifically, ChatGPT-4o mini, as an AES tool for assessing essays written by EFL learners. Building on the current findings, while ChatGPT exhibits some degree of alignment with human evaluators, substantial differences in scoring persist, echoing the concerns highlighted in previous research.

The first research question addressed the differences between the scores assigned by ChatGPT and human raters. The results indicated a statistically significant difference, with human raters consistently awarding higher scores. This finding aligns with Huang (2014), who noted that AES systems often fail to capture the nuanced elements of writing that human raters prioritize, such as creativity, depth of analysis, and linguistic fluency. Human raters' higher evaluations may suggest that they appreciate qualitative aspects that may elude AI algorithms, which typically emphasize mechanical correctness over stylistic manners.

The observed difference can also be attributed to the proficiency level of the essays analyzed. Given that this study focused on essays from B2-level learners, it is probable that it involved a level of complexity that AES tools, including ChatGPT, are unable to assess. This is consistent with Wang and Bai (2021), who found that AES systems often underperform with higher-quality writing. These findings suggest a fundamental limitation of ChatGPT: while it may be adept at detecting surface-level errors, it struggles to evaluate the intricate features of more sophisticated writing, which was also evident in earlier AES tools.

Despite the significant differences in the scores, a moderate positive correlation was found between the human and ChatGPT scores. This suggests that while ChatGPT does not replicate human evaluations precisely, it can recognize some general trends in writing quality. This finding resonates with Yancey *et al*. (2023), who reported that ChatGPT-4 could closely match human scoring under certain conditions, indicating that while AI tools may not replace human judgment, they can serve as useful adjuncts in assessing writing quality. However, as this study has shown, the current capabilities of ChatGPT-4o mini are relatively limited.

The second research question examined whether the genre of the essay influenced the agreement between the human and ChatGPT scores. The results indicated consistent differences across all five genres, supporting the assertion that genre does not mitigate the gap between human and ChatGPT evaluations. However, the results showed that the correlation between human and ChatGPT scores varied depending on the essay genre. Notably, this variation in score discrepancies by genre suggests that certain types of essays pose more significant challenges for ChatGPT assessment. The descriptive essays, which garnered the strongest difference, highlight that ChatGPT has difficulty in evaluating writing that relies on sensory details and evocative language. This finding is partly in line with Guo and Wang (2024), who noted that while ChatGPT provided detailed feedback, it occasionally delivered off-topic suggestions due to its specific prompting. Conversely, the strongest correlation was observed in analysis essays, where structured reasoning is more readily assessed by ChatGPT, suggesting that it performs better in genres where explicit criteria are more easily defined.

Overall, the findings from this study underscore that while ChatGPT shows promise as an AES tool, its limitations render it unsuitable as a standalone grading system for EFL essays. The significant differences between human and ChatGPT scores, along with the variable performance across different genres, indicate that ChatGPT is not yet capable of providing evaluations that align closely with human scoring. This can be supported by the conclusions of a very recent study by Bui and Barrot (2024), who highlighted the need for further refinement of AI algorithms and a greater sensitivity to qualitative writing aspects. Currently, while ChatGPT holds the potential to contribute to essay scoring, its limitations necessitate a cautious approach to its implementation in educational contexts. It should primarily be viewed as a supplementary tool rather than a replacement for human raters.

5. CONCLUSION

In this study, it was aimed to explore the feasibility and effectiveness of using ChatGPT, a LLM developed by OpenAI, as an AES tool for evaluating EFL learners' essays. The investigation involved comparing the scores generated by human raters to those assigned by the ChatGPT-4o mini version across various essay types. The findings revealed a significant difference between human and ChatGPT scores, with human raters assigning consistently higher scores. While a small positive correlation was observed, indicating a tendency for scores to increase together, the weak relationship suggested limited agreement between the two assessment methods. Furthermore, it was found that the genre of the essays was not a parameter mitigating between human and ChatGPT scores. While ChatGPT may show promise as an AES tool, its limitations are evident. The observed disparities highlight the complexities of language and subjective interpretation in EFL essays, which pose challenges for current AI models.

5.1. Limitations

The current study was conducted with the account of 50 essays in total, which may be argued to be a small sample. Additionally, the study included a set of essays which were written by

B2-level learners. It is suggested for further research to enrich the sample along with different proficiency levels and to include research questions accordingly.

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**: Sivas Cumhuriyet University, Educational Sciences Ethics Committee, 24.05.2024-431192.

Contribution of Authors

All stages of the study were conducted with equal contribution from both authors.

Orcid

Ahmet Can Uyar https://orcid.org/0000-0003-2438-9877
Dilek Büyükahıska https://orcid.org/0000-0002-4370-7626

REFERENCES

Almusharraf, N., & Alotaibi, H. (2022). An error-analysis study from an EFL writing context: Human and automated essay scoring approaches. *Technology, Knowledge and Learning, 28*, 1015-1031. https://doi.org/10.1007/s10758-022-09592-z

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M.D. Shermis & J.C. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 181-198). Routledge.

Bui, N.M., & Barrot, J.S. (2024). ChatGPT as an automated essay scoring tool in the writing classrooms: how it compares with human scoring. *Education and Information Technologies.* https://doi.org/10.1007/s10639-024-12891-w

Chen, H., & Pan, J. (2022). Computer or human: a comparative study of automated evaluation scoring and instructors' feedback on Chinese college students' English writing. *Asian-Pacific Journal of Second and Foreign Language Education, 7*(34), 1-20. https://doi.org/10.1186/s40862-022-00171-4

Coghlan, D., & Brydon-Miller, M. (2014). *The SAGE encyclopedia of action research*. SAGE.

Creswell, J.W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches.* SAGE Publications.

Davies A. (2008). Assessing academic English language proficiency: 40+ years of U.K. language tests. In Fox J., Wesche M., Bayliss D., Cheng L., Turner C.E., Doe C. (Eds.), *Language testing reconsidered* (pp. 73–86). University of Ottawa Press.

Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies, 29*, 8435–8463. https://doi.org/10.1007/s10639-023-12146-0

Huang, S.J. (2014). Automated versus human scoring: A case study in an EFL context. *Electronic Journal of Foreign Language Teaching, 11*(1), 149-164.

IELTS. (2019). Guide for educational institutions, governments, professional bodies and commercial organisations. Cambridge Assessment English, The British Council, IDP Australia. https://www.ielts.org/-/media/publications/guide-for-institutions/ielts-guide-for-institutions-2015-uk.ashx

IELTS. (2023). IELTS Task 2 Writing band descriptors (Public version). https://takeielts.britishcouncil.org/sites/default/files/ielts_writing_band_descriptors.pdf

Larson-Hall, J. (2012). How to run statistical analyses. In A. Mackey & S.M. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 245-274). Wiley-Blackwell.

Manap, M.R., Ramli, N.F., & Kassim, A.A.M. (2019). Web 2.0 automated essay scoring application and human ESL essay assessment: A comparison study. *European Journal of English Language Teaching, 5*(1), 146-162. https://doi.org/10.5281/zenodo.3461784

Mason, O., & Grove-Stephenson, I. (2002). Automated free text marking with paperless school. In M. Danson (Ed.), *Proceedings of the Sixth International Computer Assisted Assessment Conference* (pp. 216–222). Loughborough: Loughborough University.

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*, 1-13. https://doi.org/10.1016/j.rmal.2023.100050

Page, E. (1966). The imminence of ... grading essays by computer. *Phi Delta Kappan, 47*(5), 238–243.

Parker, J.L., Becker, K., & Carroca, C. (2023). ChatGPT for automated writing evaluation in scholarly writing instruction. *Journal of Nursing Education, 62*(12), 721-727. https://doi.org/10.3928/01484834-20231006-02

Pearson, W.S. (2022). Student Engagement with Teacher Written Feedback on Rehearsal Essays Undertaken in Preparation for IELTS. *Sage Open, 12*(1). https://doi.org/10.1177/21582440221079842

Wang, J., & Bai, L. (2021). Unveiling the scoring validity of two Chinese automated writing evaluation systems: A quantitative study. *International Journal of English Linguistics, 11*(2), 68-84. https://doi.org/10.5539/0jel.v11n2p68

Willard, C.A. (2020). *Statistical methods: An introduction to basic statistical concepts and analysis*. Routledge.

Yancey, K.P., Laflair, G., Verardi, A., & Burstein, J. (2023). Rating short L2 essays on the CEFR scale with GPT-4. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 576-584). Retrieved October 2, 2024, from https://aclanthology.org/2023.bea-1.49

Zribi, R., & Smaoui, C. (2021). Automated versus human essay scoring: A comparative study. *International Journal of Information Technology and Language Studies, 5*(1), 62-71.

International Journal of Assessment Tools in Education
2025, Vol. 12, No. 1, 33–44

https://doi.org/10.21449/ijate.1436616

journal homepage: https://dergipark.org.tr/en/pub/ijate **Research Article**

Validity and reliability of the breast cancer comfort assessment scale in palliative care

Rahime Yöntem Ölmez[1], **İlgün Özen Çınar**[1*]

[1]Pamukkale University, Faculty of Health Science, Department of Public Health Nursing, Denizli, Türkiye

ARTICLE HISTORY

Received: Feb. 15, 2024

Accepted: Oct. 07, 2024

Keywords:
Palliative care,
Breast cancer,
Comfort,
Validity,
Reliability.

Abstract: Breast cancer is a disease that requires palliative care and comfort. The current study aimed to adapt the scale used to assess the comfort level of breast cancer patients receiving palliative care, for the Turkish population, and to contribute to the literature. A total of 340 breast cancer patients who were registered at a university hospital's oncology outpatient clinic, received therapy, and returned for follow-up were included in the study. Data were collected using the Introductory Information Form, Comfort Assessment Breast Cancer Instrument, and General Comfort Scale short form. The International Testing Commission Guide's (2018) suggestions were applied during the scale's modification procedure. The scale's Kaiser Meyer Olkin value was 0.78, and 4454.53 was the Barlett's test result. Fit indices for the confirmatory factor analysis were CFI=0.885, GFI=0.927, and χ^2/df=2.612. The scale's Spearman-Brown correlation value is 0.78, and its Cronbach's alpha coefficient is 0.85. The Comfort Assessment Breast Cancer Instrument's Turkish version provides a reliable and valid tool for assessing the comfort of breast cancer patients. The use of it can help determine the comfort level of breast cancer patients receiving palliative care and inform the development of interventions and care practices throughout each stage of the disease.

1. INTRODUCTION

Cancer is a major global health problem impacting individuals' life quality (Sung *et al*., 2021). Although there have been improvements in diagnosing and treating breast cancer, it remains one of the primary factors contributing to cancer-related fatalities in women in approximately 95% of countries (Hailu *et al*., 2020; WHO 2023a). Breast cancer accounts for 11.7% of total cancer cases and 24.5% of cancers in women (Sung *et al*., 2021). Population growth and aging may cause 3 million new breast cancer cases and 1 million deaths by 2040 (Arnold *et al*., 2022). From 1994 to 2020, breast cancer incidence in Turkey increased 2.5-fold to 23.9% among female cancers (Ferlay *et al*., 2020)

Patients and families with life-threatening diseases benefit from multidisciplinary palliative care. Only 14% of those needing palliative care worldwide access the service (WHO, 2023b). Symptomatic patients with breast cancer need early palliative care (Nuraini *et al*., 2018; Malloy *et al*., 2018). Palliative care is crucial as breast cancer rates rise and life expectancy rises (Zimmermann *et al*., 2014; Ferrell, 2019). It complements therapeutic and lifelong breast cancer

*CONTACT: İlgün ÖZEN ÇINAR ✉ iocinar@pau.edu. Pamukkale University, Faculty of Health Science, Department of Public Health Nursing, Denizli, Türkiye

e-ISSN: 2148-7456

treatment at all ages and stages (WHO, 2023a). Palliative care is improving health indicators, depression, and life expectancy of breast cancer patients (Rugno *et al*., 2014).

Patients with breast cancer and their families lose comfort. Palliative care for all cancers, including breast cancer, includes comfort (Nuraini *et al*., 2018). Comfort is defined by Kolcaba as the absence of discomfort, the resolution of causative conditions, satisfaction, and situations that make life easier and more pleasant. Kolcaba's comfort theory involves determining comfort needs, planning interventions, considering factors, and evaluating (Kolcaba, 1991). Comfort needs are determined holistically and assess the individual's physical, psychospiritual, socio-cultural, and environmental comfort needs (Kolcaba 2003; Kolcaba & Dimarco 2005). An individual's physical comfort is their body perception and affects their disease comfort. Psychospiritual comfort is the combination of spiritual, psychological, and mental health. For instance, surgical intervention causes anxiety and impairs comfort. Environmental comfort is the external factors (noise, heat, etc.) that affect comfort. Socio-cultural comfort is the individual's perception of and relationships with the social and cultural environment. For instance, an individual's traditional approach and social support affect his comfort (Kolcaba, 2003).

Comforted palliative care patients recover faster, rehab better, and handle stress better. Nuraini *et al*. (2018) developed the instrument assessing breast cancer patients' comfort (CABCI) to evaluate their physical, psycho-social, sociocultural, economic, and hospital environment comfort for diagnosis, treatment, and care (Nuraini *et al*., 2018). Previous studies conducted with breast cancer patients in Turkey have frequently used the general comfort scale (Çıtlık *et al*., 2018). There is a need for a specialized tool that holistically assesses the comfort of breast cancer patients. The study aimed to adapt the scale for the Turkish population and contribute to the literature.

2. METHOD

2.1. Study Design and Population

This study was methodological research conducted to validate a Turkish version of the CABCI developed by Nuraini *et al* (2018) to assess the breast cancer patient's comfort. The participants consisted of breast cancer patients who applied to the Oncology and Chemotherapy Clinic of a university hospital for treatment and control purposes. For validity-reliability studies, the sample size should be determined by 5 or 10 times the number of scale items (Grove *et al*., 2013; Erdoğan *et al*., 2017). In this context, the sample of the study consists of 340 breast cancer patients with 10 times the number of scale items. The inclusion criteria: a) Over 18, b) no communication barriers, c) radiotherapy, chemotherapy, or both. Breast cancer patients who met the sampling criteria and volunteered to participate in the study were included in the study.

2.2. Data Collection Methods

Data was collected by researchers via face-to-face survey between September 2019 and March 2020. Data collection time is 10-15 minutes. The introductory Information Form covers age, education, marital status, family structure, employment, residence, income, social security, treatment information, and support. Outpatient clinic records provide diagnosis year, stage, treatment, and hemodynamic status.

Comfort Assessment Breast Cancer Instrument (CABCI); developed by Nuraini *et al*. (2018), aims to assess the breast cancer patient's comfort. The authors' first version has 34 items and five subscales. The sub-dimensions for comfort are physical (1-10), psycho-spiritual (11-22), socio-cultural (23-26), environmental (27-30) and finance (31-34). The scores are based on strongly disagree (1), strongly disagree (2), agree (3), and strongly agree (4), where the highest is 136 and the lowest is 34. Higher scores indicate higher comfort. Cronbach's alpha value is 0.91 (Nuraini *et al*., 2018). In 2019, Nuraini *et al*. (2019) revised the instrument as a single factor and 33 items by combining 5 sub-dimensions. In this study, the first study with permission from the authors was used.

General Comfort Questionnaire- Short form (GCQ-SF); developed by Kolcaba *et al*. (2006), aims to measure the patients' comfort. The instrument has nine items for relief, relaxation, and problem-solving (10 items). The Likert-type scale has 28 items and both positive and negative items (19 items). In the evaluation, negative items are reversed, coded, and summed. To determine the average score, the total score is divided by the number of instrument items. The highest score recorded is 168, while the lowest score recorded is 28. A higher score indicates a higher level of comfort. The scale was adapted to Turkish by Çıtlık *et al*. (2018) and Cronbach's alpha value was 0.82.

2.3. Language Validity of the Scale

The ITC Guidelines for Translating and Adapting Tests (Second Edition) (2018) guided instrument adaptation. It has 18 guidelines in six sections: Pre-condition, Test Development, Confirmation, Administration, Score Scales and Interpretation and Documentation. Each guideline has a description with implementation recommendations (ITC, 2018). The authors received permission from the scale authors in the first section, believing that the scale was necessary for Turkish society and could provide cultural adaptation in assessing the comfort of patients with breast cancer who are in palliative care. Expert translators in the target language and culture were determined (see Table 1). In the second part of the test development, the language adaptation process and examination of the scale's language, forward translation, expert panel utilization, back-translation, and preliminary application of the adapted version, finalization, and documentation recommendations were followed (see Table 2).

Table 1. *Adaptation process of the scale according to the first section of the ITC guideline.*

		ITC guıde 2018	Evidence
First Section Precondition	O1	Obtaining permission from the author to adapt the scale into Turkish.	Scale use permission
	O2	Evaluation of adequacy of scale structure	Researchers
	O3	Choosing the translators selected for the advanced translation of the scale in accordance with the target language and culture	An expert translator and interpreter and an English teacher were determined.

Table 2. *Adaptation process of the scale according to the first section of the ITC guideline.*

		ITC guide 2018	Evidence
Second Section Test Development	T1	Selection of experts with relevant expertise	Creation of the expert panel
	T2	Using appropriate translation design and procedure	Forward translation, expert panel, reverse translation
	T3	Proving that the scale has a similar structure for Turkish society	Expert panel report
	T4	Scale scores, evidence of whether the form of administration was appropriate	Expert panel report
	T5	Pre-application of the adapted test	Pre-application analysis result

Two independent professional native English-speaking translators back-translated the scale. To determine the data collection forms' comprehensibility and applicability, a preliminary application was performed on 20 breast cancer patients. By assessing question comprehensibility, item analysis, and Cronbach's alpha levels, the scale was adapted (Cronbach's alpha: 0.94, spearman-brown correlation coefficient: 0.839, Guttman split-half: 0.829). Forms were not modified because patients understood all expressions and content. Pre-application data were not included. Data analyses were performed in the third section to choose a suitable sample and prove its reliability and validity. The administration section standardized the scale structure and related procedures for the new language and culture. In the last two sections, score scales and interpretation were made, and documentation was created (ITC, 2018; Hernandez *et al*., 2020).

2.4. Content Validity of the Scale

In the ITC (2018) Guidelines, the items' comprehensibility was questioned, and expert opinion was obtained. Content validity was evaluated with the Davis technique. Comparing Turkish and original versions, experts scored each instrument item. The content validity index (CVI) value is expected to be 0.80 and above (Davis, 1992). An expert from the Department of Medical Oncology rejected the original scale's 14th item, "I feel anxious about death," because it mentioned death. With the scale author's permission, this item was changed to "I feel anxious about my future" with expert opinions Expert panel report finalized the scale. In this study, item comprehensibility ranged between 0.88- 1.

2.5. Ethical Considerations

The scale authors permitted for use. The Non-Interventional Clinical Ethics Committee of a university obtained ethical approval (dated 06.08.2019 number 54328). The principles of the Declaration of Helsinki guided the conduct of this study. The data collection institution and study participants gave their consent.

2.6. Statistical Analysis

The validity of the scale was tested using Confirmatory Factor Analysis (CFA). Before starting CFA, whether the data is normally distributed or not determines the estimation method and the type of matrix to be created (Çapık, 2014; Gana & Broc, 2019). Normal distribution was evaluated with skewness and kurtosis coefficients. The Dampened-Weighted Least Squares (DWLS) technique was chosen as it was the preferred technique for estimating Likert-type data in CFA. Analysis was conducted using R-Project (R Core Team, 2020), Lavaan (Rosseel, 2012), and IBM SPSS 26. The margin of error in the study was at 95% confidence level (p<.05).

In validity analysis, the CVI value was calculated for content and scope validity. In construct validity, Barlett's test and Kaiser-Mayer-Olkin (KMO) test assessed sample size and factor analysis suitability. Pearson Product Moment Correlation tested scale construct validity in CFA concurrent validity. In the reliability analysis; item-total score correlation, Cronbach's alpha, spearman-brown coefficient, internal consistency, and two-half reliability were evaluated.

3. RESULTS

The mean age of the patients was 53.08±17.84. Of the patients, 33.2% of them were in the second stage, 55.3% received chemotherapy and 19.4% received radiotherapy (see Table 3).

Table 3. *Descriptive characteristics of breast cancer patients.*

Variables		*n*	%
*Age**	39 and less	92	27.1
	40-64	148	43.5
	65 and over	100	29.4
Educational status	8 years&less	210	61.8
	8 years&over	130	38.2
Marital Status	Single	118	34.7
	Married	222	65.3
Employment Status	Unemployed	206	60.6
	Employed	134	39.4
Getting information about treatment	Yes	244	71.8
	No	96	28.2

Type of treatment	Chemotherapy	188	55.3
	Radiotherapy	66	19.4
	Chemotherapy and radiotherapy	86	25.3
Stage of cancer	Stage I	101	29.7
	Stage II	113	33.2
	Stage III	85	25.0
	Stage IV	41	12.1

*The average age:53.08 ± 17.84

3.1. Validity Findings of CABCI

Construct validity was assessed after language and content validity. The scale's KMO was 0.78 and Bartlett's test of Sphericity was 4454.53 (p<0.001). Since the data were Likert-type, DWLS was preferred for CFA estimation. The CFA statistics revealed that all sub-items of CABCI were statistically significant (p<0.05) (see Table 4).

Table 4. *CFA statistics of the scale.*

Category	Items	Beta	SE	z value	*p*
Physical	S1	1			
	S2	0.79	0.051	15.54	<0.001
	S3	0.66	0.050	13.50	<0.001
	S4	0.68	0.046	14.92	<0.001
	S5	1.12	0.065	17.49	<0.001
	S6	0.80	0.055	14.76	<0.001
	S7	0.69	0.050	13.89	<0.001
	S8	0.46	0.047	9.82	<0.001
	S9	0.46	0.036	12.98	<0.001
	S10	0.26	0.031	8.55	<0.001
Psycho-spiritual	S11	1			
	S12	0.90	0.051	17.96	<0.001
	S13	0.84	0.047	17.91	<0.001
	S14	0.68	0.043	16.15	<0.001
	S15	0.74	0.046	16.28	<0.001
	S16	0.25	0.032	7.89	<0.001
	S17	0.65	0.043	15.03	<0.001
	S18	0.45	0.040	11.31	<0.001
	S19	0.45	0.030	14.92	<0.001
	S20	0.91	0.050	18.07	<0.001
	S21	0.72	0.047	15.48	<0.001
	S22	0.67	0.039	17.56	<0.001
Socia-cultural	S23	1			
	S24	0.71	0.075	9.51	<0.001
	S25	0.54	0.063	8.67	<0.001
	S26	0.44	0.052	8.50	<0.001
Finance	S27	1			
	S28	1.01	0.067	15.00	<0.001
	S29	1.04	0.070	15.00	<0.001
	S30	0.52	0.047	11.21	<0.001
Environmental	S31	1			
	S32	0.75	0.090	8.37	<0.001
	S33	0.52	0.065	8.15	<0.001
	S34	0.23	0.042	5.53	<0.001

SE: Standart Error

The CFA graphical structure showed all items had standardized loadings above 0.20 (see Figure 1). The goodness of fit index values was $\chi^2/df = 2.612$, GFI = 0.927, AGFI = 0.916, CFI = 0.885, TLI = 0.876, RMSEA = 0.069 and SRMR = 0.083 (see Table 5).

Figure 1. *CFA graphical structure.*

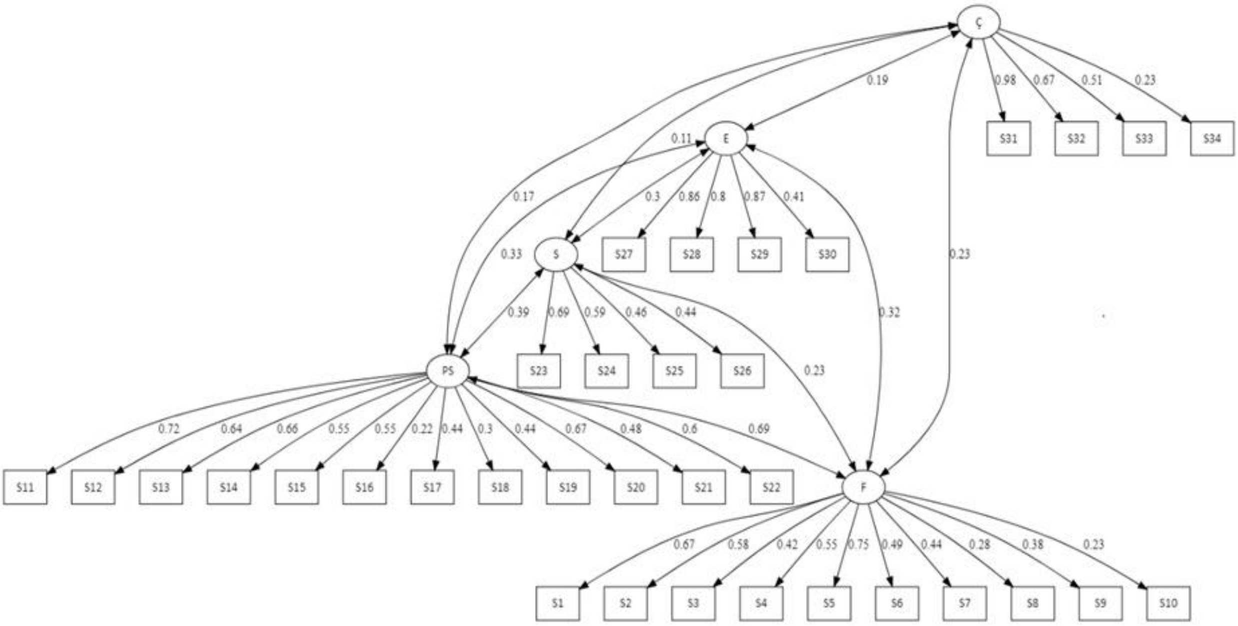


Table 5. *Fit index of CFA findings of the scale.*

Goodness-of-fit indices	
χ^{2*}	1350.516
χ^2/df^{**}	2.612
RMSEA	0.069
TLI	0.876
SRMR	0.083
CFI	0.885
AGFI	0.916
GFI	0.927

RMSEA, root mean square error of approximation; TLI, Tucker-Lewis index; SRMR, standardized root mean square residual; CFI, comparative fit index; AGFI, Adjusted goodness of fit index GFI goodness of fit index, *df* (degree of freedom)=517, $^*p$<.001, $^{**}p$<.05

3.2. Reliability Findings of CABCI

Table 6 shows the mean scale score and sub-scores. The scale's total score and sub-dimensions' skewness and kurtosis values were normal.

Table 6. *Descriptive statistics and normality tests of total scores of the scale and its sub-dimensions.*

Category	$\bar{X}$±SD	Min-Max	Skewness	Kurtuosis
Physical	22.10±5.61	10.000-35.000	0.115	-0.661
Psycho-spiritual	26.04±6.95	12.000-46.000	0.351	0.024
Socia-cultural	7.95±2.65	4.000-16.000	0.332	-0.554
Finance	9.20±3.60	4.000-16.000	0.227	-1.095
Environmental	10.10±2.79	4.000-16.000	-0.051	-0.887
CABCI	75.40±14.16	36.00-111.00	0.236	-0.304

$\bar{X}$±SD: Mean± Standard Deviation, Min-Max: Minimum-Maximum

The scale's items were examined, and Cronbach's alpha coefficient was determined for internal consistency and homogeneity reliability. The scale's item means, and standard deviation were 1.571±0.782 and 3.083±1.135. The item means showed no zero-standard deviation items. Removing items from subscales did not significantly increase the reliability coefficient. All subscale item corrected correlation values were positive. The Cronbach’s alpha coefficients for the subscales were 0.76, 0.82, 0.64, 0.81, and 0.71, respectively (see Table 7).

Table 7. *Reliability analysis results of the scale.*

Category	Items	$\bar{X}$	SD	AC	AID	Alpha
Physical	S1	2.356	1.013	0.442	0.742	0.76
	S2	2.179	0.944	0.501	0.734	
	S3	2.300	1.075	0.415	0.746	
	S4	1.718	0.853	0.470	0.739	
	S5	2.171	1.022	0.578	0.722	
	S6	2.509	1.122	0.447	0.741	
	S7	2.685	1.072	0.488	0.735	
	S8	2.529	1.117	0.373	0.753	
	S9	2.024	0.834	0.370	0.751	
	S10	1.638	0.821	0.186	0.771	
Psycho-spiritual	S11	2.138	1.045	0.601	0.795	0.82
	S12	2.418	1.076	0.581	0.796	
	S13	2.168	0.968	0.642	0.792	
	S14	1.941	0.945	0.598	0.796	
	S15	1.800	1.034	0.562	0.798	
	S16	1.547	0.873	0.184	0.827	
	S17	2.844	1.103	0.417	0.812	
	S18	3.083	1.135	0.299	0.823	
	S19	1.935	0.773	0.438	0.810	
	S20	2.018	1.019	0.538	0.801	
	S21	2.300	1.144	0.398	0.814	
	S22	1.865	0.851	0.452	0.808	
Socia-cultural	S23	2.509	1.138	0.398	0.610	0.64
	S24	1.891	0.942	0.575	0.467	
	S25	1.979	0.913	0.479	0.540	
	S26	1.571	0.782	0.278	0.663	
Finance	S27	2.100	1.068	0.737	0.712	0.81
	S28	2.129	1.160	0.755	0.697	
	S29	2.097	1.099	0.706	0.724	
	S30	2.876	1.183	0.361	0.885	
Environmental	S31	2.335	0.937	0.605	0.582	0.71
	S32	2.818	1.032	0.603	0.577	
	S33	2.447	0.947	0.460	0.670	
	S34	2.500	0.901	0.336	0.737	

$\bar{X}$: Mean, SD: Standard Deviation, AC: Adjusted Correlation, AID: Alpha when ıtem is deleted (Hotelling’s T-Squared 223.2 *p*=0.000)

Regarding internal consistency, CABCI's total mean score was 75.409±14.167, the Spearman-Brown correlation coefficient was 0.78, and Cronbach's alpha coefficient was 0.85, (Table 8).

Table 8. *Internal consistency values of scales (n=340).*

Mean±SS	Cronbach’s Alpha	Spearman-Brown Correlation Coefficient	Guttman Split-Half
75.409±14.167*	0.85	0.78	0.78

*Hotelling’s T-Squared *F*=43.41, *p*<0.001

4. DISCUSSION and CONCLUSION

The final version was created after the ITC Guide (2018) language validity was performed. In instrument adaptation studies, language validity should be supported by content validity (ITC, 2018). 10 academics with diverse expertise provided expert opinions for the study. Expert consensus and scale content validity are indicated by a CVI index above 0.80. Pre-application analysis values are excellent or acceptable, indicating item validity and reliability. If the results are unsatisfactory, adapt by improving the problematic items (Hernandez *et al*., 2020). No issues were found in patient's perception and response to the CABCI during language validity testing. The pre-application analysis' excellent item correlation coefficient and Cronbach's alpha values guided the scale's adaptation study applicability.

In construct validity, the KMO test was conducted to assess the entire model and its variables' adequacy for sampling adequacy and suitability for analysis before CFA. The 0.90-1.00 KMO value is evaluated as excellent, 0.50-0.59 poor, 0.60-0.69 fair, 0.70-0.79 good, 0.80-0.89 very good (Sarmento & Costa 2017; Nia *et al*., 2023). This value was determined at a good level in our study. Barlett's test determined whether the data was normal and whether the correlation matrix was a unit matrix (Caycho -Radriguez *et al*., 2021). Our study's KMO (0.78) and Barlett's value are significant, and the sample size is good for factor analysis.

Construct validity determines how well an instrument measures the concept or event and how well its items relate to each other. Factor analyses evaluate construct validity, and the measurement tool should have high construct validity (Gana & Broc, 2019). Instead of EFA, a factor analysis method, CFA, the most common model verification method, should be used in instrument adaptation (Erdoğan *et al*., 2017; Seçer, 2018). So, CFA was performed in the instrument adaptation process. The results of the fit indexes of the CABCI are well-compatible (CFI = 0.885, GFI = 0.927, AGFI = 0.916, SRMR = 0.083, TLI = 0.846, χ^2/df = 2.42, RMSEA = 0.069). In the first instrument development study, Nuriani *et al*. (2018) did not specify fit index values, but CFA was performed, and instrument validity was confirmed. Some of the fit indexes in the construct validation of the scale in 2019 are given (Nuriani *et al*., 2019). Based on the statistically significant χ^2 value, the fit between the model and the data is not perfect. However, χ^2 is not a reliable and robust model fit indicator. This value is also sensitive to the sample size. It is therefore recommended to look at other fit indices. Examination of these indices (e.g., CFI, RMSEA, SRMR) shows that the model fits the data well (Gana & Broc, 2019). The χ^2/df value, called the initial fit index, shows the difference between the observed and expected covariance matrices (Gunzler & Morris, 2016). Higher values indicate that the model does not fit the data, while lower values indicate a better fit (Costa & Sarmento, 2019). A value of three or less, which is also expressed as a poor fit index, is an indicator of excellent fit (Çokluk *et al*., 2014; Seçer, 2018). Our study's CABCI value (2.42) was within the excellent fit, but Nuriani *et al*. (2019) found a high χ^2/df value in their instrument construct validity (χ^2=283.65, df=10). The theoretical model's adequacy is shown by strict fit indexes. For optimal fit, a few parameters should be estimated. The most recommended index in this category is the RMSEA with a 90% confidence interval (Gana & Broc, 2019). RMSEA tries to correct the chi-square value's tendency to reject instruments with large samples. RMSEA is very good if it is equal to or below 0.05, good between 0.05 and 0.08, moderate between 0.08 and 0.10, and unacceptable if above 0.10 (Costa & Sarmento, 2019). The RMSEA value of the scale (0.069) shows a good fit. Nuriani *et al*. (2019) reported a good fit with RMSEA=0.000. One of the absolute fit indexes, Root Mean Square Residual (RMR) or Standardized RMR measures observed and predicted correlation errors. RMR and SRMR decrease as model element deviations decrease. The SRMR value should be between 0.00 and 1.00. When this value is close to 0.00, the fit is better (Gana & Broc 2019; Costa & Sarmento, 2019). In our study, the CABCI's SRMR value is a good fit. Other absolute fit indexes are the Goodness of Fit Index (GFI) and Adjusted Goodness of Fit Index (AGFI). These index degrees of freedom increase with sample size (Costa & Sarmento, 2019, Gunzler & Morris, 2016). These values of 0.90 and above indicate a perfect fit (Gana & Broc, 2019). In our study, these values were found to be perfectly compatible. Incremental fit indexes (TLI, CFI) analyze model fit by examining the comparing data to the proposed model while assessing the chi-square sample size, and these values between 0 and 1 show excellent fit (Gana & Broc, 2019). According to Costa and Sarmento (2019), CFI and TLI values are very good if they are equal to or above 0.95, good between 0.9-0.95, moderate between 0.8 - 0.9, and poor below 0.8. Brown (2015) states that these indexes being equal to or above 0.80 indicate an acceptable fit. In our study, CFI and TLI were considered moderate fit indices. In Nuriani *et al*.'s study (2019), the CFI value was found to be 1.000 and it was stated to have a good fit index value. Despite a statistically significant χ^2 value, the values of the other fit indices indicate that the model is compatible with the data.

Factor analysis calculates factor loadings by grouping variables that measure the same dimension and calculating their correlation using sample group responses. Factor loading coefficients explain item-factor relationships (Harrington, 2009; Gana & Broc, 2019). The CFA result's graphical structure shows that four scale items (items 8, 10, 16, and 34) have factor loadings above 0.20 and others above 0.30. The factor loading value should be above 0.30 (Çokluk *et al*., 2014; Seçer, 2018), but it can also be above 0.20 (Grove *et al*., 2013), and another suggestion is that more samples may reduce factor loadings (Gana & Broc, 2019). The Turkish version's factor structure of the CABCI matches the structure in the original instrument. In the CFA statistics, all CABCI sub-items were significant.

Concurrent validity compares a Turkish-adapted instrument to a validated and reliable scale (Erdoğan *et al*., 2017). GCQ-SF concurrent validity showed a positive and moderately significant relationship in our study. When the patients' comfort is high in GCQ-SF, an increase is seen in CABCI measurement. This shows the validity of the CABCI scale when applied together with the previously validated scale. This shows the validity of the CABCI scale when applied together with the previously validated scale

4.2. Discussion of the Reliability Findings of the Scale

When the sample size is 300 or more, absolute skewness and kurtosis values are taken into account in evaluating the normality of the data. For a normal distribution, absolute skewness ≤2 and absolute kurtosis ≤4 are reference values (Kim, 2013). In our study, the data showed a normal distribution. It is important to specify that the distribution of the normal constitutes a convenient model serving a technical benchmark (Gana & Broc, 2019). Reliability is a crucial feature of any scale (Streiner *et al*., 2015), and is typically determined by Cronbach's alpha, which measures the internal consistency of instrument items. A value between 0.00 and 0.40 indicates low reliability, 0.40 to 0.59 suggests moderate reliability, 0.60 to 0.79 reflects good reliability, and 0.80 to 1.00 signifies high reliability (Grove *et al*., 2013). In this study, the CABCI subdimensions' Cronbach's alpha coefficients ranged from 0.64 to 0.82, and the total alpha value was 0.85, indicating high reliability. Nuriani *et al*. (2018) also found Cronbach's alpha to be highly reliable ($\alpha = 0.91$), with item mean and standard deviation distributions between 1.57 ± 0.78 and 3.08 ± 1.13.

Item-total correlation is commonly used to test the homogeneity of a scale with several items. Any item with a low correlation value measures a different characteristic than other instrument items. Literature suggests that item-total correlation values above 0.20 are considered acceptable. The item-total score correlation coefficient starts at 0.20, and item scores between 0.30-0.40 are good and above 0.40 are very good (Streiner *et al*., 2015). Items with a correlation coefficient below 0.20 should be removed from the scale, but only if their removal improves or does not affect the overall Cronbach's alpha (Grove *et al*., 2013). In our study, all items had good item-total correlation coefficients. The mean CABCI score indicated moderate comfort in breast cancer patients, with moderate scores across all subdimensions, highlighting the need to address patients' comfort in all areas. The applied test was divided into two equal halves to estimate split-half reliability, with the Spearman-Brown coefficient used to assess the correlation between participants' scores on each half (Erdoğan *et al*., 2017). The Spearman-Brown correlation coefficient for the CABCI was 0.78, meeting the recommended reliability threshold of 0.75 or higher (Grove & Cipher, 2019). This suggests that the scale has high internal consistency and stability.

In conclusion, the CABCI is a valid and reliable tool for assessing the comfort of breast cancer patients receiving palliative care within the Turkish context (Appendix A1). Given the critical role of palliative care in breast cancer, this scale can be used clinically to assess patient comfort at any stage of the disease. It evaluates economic, socio-cultural, physical, psycho-spiritual, and environmental dimensions of comfort, supporting holistic care. Nursing interventions to improve breast cancer patients' palliative care comfort should use the scale. This scale will contribute to the individual, family, and society by using it in application areas and future research. In future studies, it is recommended to repeat the scale in patients at different stages. The scale was developed and customized for breast cancer patients. In our study, we validated the scale specifically for breast cancer patients, a group disproportionately affected by the disease both globally and in our country. While general comfort scales have been used for other cancers and chronic diseases, future research could explore disease-specific comfort scales for other chronic conditions.

Acknowledgments

The authors thank all breast cancer patients and healthcare professionals for their support. The study was presented orally at the 1st International Congress of Palliative Care in Nursing, 06-08 November 2022, Burdur-Türkiye.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Pamukkale University, Non-Interventional Clinical Research Ethics Committee, 60116787-020/54328.

Contribution of Authors

Rahime Yöntem Ölmez: Conception, Design, Investigation, Literature review, Methodology, Data collection, Data interpretation, and Writing-original draft. **İlgün Özen Çınar**: Conception, Design, Supervision, Methodology, Formal Analysis, Finding, Critical review, and Writing-original draft.

Orcid

Rahime Yöntem Ölmez https://orcid.org/0000-0003-0727-3151
İlgün Özen Çınar https://orcid.org/0000-0001-5774-5108

REFERENCES

Arnold, M., Morgan, E., Rumgay, H., Mafra, A., Singh, D., Laversanne, M., … Soerjomataram, I. (2022). Current and future burden of breast cancer: Global statistics for 2020 and 2040. *Breast*, *66*, 15-23. https://doi.org/10.1016/j.breast.2022.08.010

Brown, T.A. (2015). *Confirmatory factor analysis for applied research*. Guilford Press.

Çapık, C. (2014). Geçerlik ve Güvenirlik Çalışmalarında Doğrulayıcı Faktör Analizinin Kullanımı [Use of Confirmatory Factor Analysis in Validity and Reliability Studies]. *Journal of Anatolia Nursing and Health Sciences*, *17*(3),196-205.

Caycho-Rodríguez, T., Rojas-Jara, C., Ventura-León J, Noe-Grijalva M, Cabrera-Orosco I., & Reyes-Bossio M. (2021). Single item to assess for worry for cancer: Initial evidence of validity and reliability. *Enfermería Clínica*, (Engl Ed.). *31*, 203-210. https://doi.org/10.1016/j.enfcle.2019.11.002

Çıtlık S.C., Çevik, S., & Özden, G. (2018). Genel konfor ölçeği kısa formu türkçe geçerliliği ve güvenilirliği [Validity and reliability study of the Turkish version of the short general comfort questionnaire]. *Diyabet, Obezite ve Hipertansiyonda Hemşirelik Forumu Dergisi 10*(2), 16-22. http://www.tdhd.org/pdf/Diyabet_Hems_Forumu_2018_2.pdf

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (Eds). (2014). *Sosyal bilimler için çok değişkenli istatistik SPSS ve LISREL uygulamaları, yapısal eşitlik modeli* [*Multivariate statistics for the social sciences, SPSS and LISREL applications, structural equation modeling*]. Pegem Yayıncılık.

Davis, L.L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research, 5*(4), 194–197. https://doi.org/10.1016/S0897-1897(05)80008-4

Erdoğan, S., Nahcivan, N., & Esin, M. (Eds) (2017). Hemşirelikte araştırma, süreç uygulama ve kritik, veri toplama yöntem ve araçları & Veri toplama araçlarının güvenirlik ve geçerliliği [Research in nursing, process, practice and critique, data collection methods and tools & Reliability and validity of data collection tools]. Nobel Tıp Kitapevleri

Ferlay, J., Ervik, M., Lam, F. Colombet, M., Mery, L., Piñeros, M., … Bray, F. (2020). *Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer*. Available from (accessed 14 March 2023): https://gco.iarc.fr/today

Ferrell, B. (2019). National consensus project clinical practice guidelines for quality palliative care: ımplications for oncology nursing asia-pacific. *Journal of Oncology Nursing, 6*(2), 151-153. https://doi.org/10.4103/apjon.apjon_75_18

Gana, K., & Broc, G. (2019). *Structual Equation Modeling with lavaan*. Great Britain and United States by ISTE and John Wiley&Sons, Inc ISBN 978-1-78630-369-1.

Grove, S.K., Burns, N., & Gray, J. (2013). *The practice of nursing research: appraisal, synthesis, and generation of evidence*, (7th ed.), St. Louis, Missouri: Elsevier Saunders 72.

Grove, S.K., & Cipher, D.J. (2019*). Statistics for nursing research e-book: a workbook for evidence-based practice*. Elsevier Health Science.

Gunzler, D.D., & Morris, N.A. (2015). Tutorial on structural equation modeling for analysis of overlapping symptoms in co-occurring conditions using Mplus. *Statistics in Medicine, 34*(24), 3246–3280. https://doi.org/10.1002/sim.6541

Hailu, H.E., Mondul, A.M., Rozek, A.S., & Geleta, T. (2020). Descriptive epidemiology of breast and gynecological cancers among patients attending Saint Paul's Hospital Millennium Medical College. *Plos One 15*(3), e0230625. https://doi.org/10.1371/journal.pone.0230625

Harrington, D. (2009). *Confimatory Factor Analysis*. Oxfort Universty Press 21-35.

Hernández, A., Hidalgo, M.D., Hambleton, R.K., & Gómez-Benito, J. (2020). International Test Commission guidelines for test adaptation: A criterion checklist. *Psicothema, 32*(3), 390–398. https://doi.org/10.7334/psicothema2019.306

International Test Commission. (2018). ITC guidelines for translating and adapting tests (Second edition). *International Journal of Testing, 18*(2), 101-134. https://doi.org/10.1080/15305058.2017.1398166

Kim, H.Y. (2013). Statistical notes for clinical researchers: Assessing normal distribution using skewness and kurtosis. *Restorative Dentistry and Endodontics, 13*(38), 52-54. https://doi.org/10.5395/rde.2013.38.1.5

Kolcaba, K., Tilton, C., & Drouin, C. (2006). Comfort theory: A unifying framework to enhance the practice environment. *Journal of Nursing Administration, 36*(11), 538–544. https://doi.org/10.1097/00005110-200611000-00010

Kolcaba, K. (1991). Taxonomik structure for the concept comfort. *Journal of Nursing Scholarship, 23*(4), 237-240. https://doi.org/10.1111/j.1547-5069.1991.tb00678.x

Kolcaba, K. (2003). *Comfort theory and practice: A vision for holistic health care and research*. Springer Publishing Company. Canada.

Malloy, P., Takenouchi, S., Kim, H.S., Lu, Y., & Ferrell, B. (2018). Providing Palliative Care Education: Showcasing Efforts of Asian Nurses. *Asia Pacific Journal of Oncology Nursing*, *5*(1), 15-20. https://doi.org/10.4103/apjon.apjon_55_17

Nia, H.S., Somasundram, S., Fomani, F.K., Kaveh, O., & Hosseini, L. (2023). Validity and Reliability of Persian Version of the 12-Item Expectations Regarding Aging Survey. The *International Journal of Aging and Human Development*, *96*(2), 248-262. https://doi.org/10.1177/00914150221084650.

Nuraini, T., Antirijono, Gayatri, D., Irawaty, D., Umar, J., & Gayatri, D. (2019). Construct and criterion validity of The Comfort Assessment Breast Cancer Instrument. *Enfermería Clínica*, *29*(2), 826-830. https://doi.org/10.1016/j.enfcli.2019.04.124

Nuraini, T., Gayatri, D., & Irawaty, D. (2018). Validity And Reliability of the Comfort Assessent Breast Cancer Instrument in Breast Cancer Palliative Care. *Enfermería Clínica*, *28*(1), 162-166. https://doi.org/10.1016/S1130-8621(18)30059-7

R Core Team. (2020). *A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rosseel, Y. (2012). Lavaan: An R ackage for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Rugno, F.C., Paiva, B.S., & Paiva, C.E. (2014). Early integration of palliative care facilitates the discontinuation of anticancer treatment in women with advanced breast or gynecologic Cancers. *Gynecologic Oncology*, *135*(2), 249-254. https://doi.org/10.1016/j.ygyno.2014.08.030.

Sarmento, R., & Costa, V. (2017). Comparative approaches to using R and Python for statistical data analysis. (p:124-156) IGI Global. https://doi.org/10.4018/978-1-68318-016-6

Seçer, İ. (2018). *Psikolojik test geliştirme ve uyarlama süreci* [*Psychological test development and adaptation process*]. Anı Yayıncılık.

Streiner, D., Norman, G., & Cairney, J. (2015). *Health measurement scales*: A practical guide to their development and use 5th ed. Oxford, UK: Oxford University Press, (p;234-248). https://doi.org/10.1093/med/9780199685219.001.0001

Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomatatam, I., Cemal, A., & Bray F. (2020). Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *A Cancer Journal for Clinicians*, *71*(3), 209-249. https://doi.org/10.3322/caac.21660

World Health Organization (WHO). (2023a). *Global Breast Cancer Initiative Implementation Framework (GBCI) 2023*: Assessing, strengthening and scaling-up of services for the early detection and management of breast cancer. Geneva: World Health Organization; Licence: CC BY-NC-SA 3.0 IGO

World Health Organization (WHO). (2023b). *Palliative-care*. Available from (accessed; 15.06.2023): https://www.who.int/health-topics/palliative-care

Zimmermann, C., Swami, N., Krzyzanowska, M., Hannon, B., Leighl, N., Oza A., … Lo, C. (2014). Early palliative care for patients with advanced cancer: a cluster-randomised controlled trial. *Lancet*, *383*(9930), 1721–1730. https://doi.org/10.1016/S0140-6736(13)62416-2

APPENDIX

A1. Turkish Form of Comfort Assessment Breast Cancer Instrument

Meme Kanseri Konfor Değerlendirme Ölçeği

Maddeler	Kesinlikle Katılıyorum	Katılıyorum	Katılmıyorum	Kesinlikle Katılmıyorum
1. Güçsüz hissediyorum				
2. Mide bulantısı hissediyorum				
3. Sağlık durumum nedeniyle ailemin ihtiyaçlarını karşılamakta zorlanıyorum (yemek yapmak, çocuklara bakmak gibi)				
4. Tedavinin yan etkileri beni rahatsız etti				
5. Kendimi hasta hissediyorum				
6. İştahım yok				
7. Sık sık başım dönüyor				
8. Cildimin ve ağzımın çok kuru olduğunu hissediyorum				
9. Yatak istirahati için çaba gösteriyorum				
10. Hemen yoruluyorum				
11. Mutsuz hissediyorum				
12. Hastalığımla mücadele etme konusunda ümitsizim				
13. Kendimi huzursuz hissediyorum.				
14. Geleceğim konusunda endişeliyim				
15. Durumum kötüleşir diye korkuyorum				
16. Ailemdeki bireylerinde aynı hastalığa yakalanmasından endişe duyuyorum				
17. Kızgın hissediyorum				
18. Yalnız hissediyorum				
19. Kendimi iyi hissetmediğim bazı değişiklikler yaşıyorum				
20. Tedaviden korkuyorum				
21. Tedaviyi sürdürmekten sıkıldım				
22. Kendimi daha hassas hissediyorum.				
23. Kendimi diğer insanlara bağımlı hissediyorum				
24. Hastalığım başka insanların hayatını etkilediği için üzülüyorum				
25. Başkalarına yük olmaktan korktuğum için hastalığımı konuşmak istemiyorum				
26. Ailemi korkutuyorum				
27. Tedavinin maliyeti beni endişelendiriyor				
28. Hastaneye ulaşım maliyeti konusunda endişeliyim				
29. Tedavim boyunca oluşan maliyet konusunda endişeliyim				
30. Hastalık gelirimi kaybetmeme neden oluyor				
31. Hastane ortamından rahatsız oluyorum				
32. Hastane ortamında kalmaya katlanamıyorum				
33. Hastane ortamının kokusundan hoşlanmıyorum				
34. Hastane ortamında rahat hissedebiliyorum				

International Journal of Assessment Tools in Education
2025, Vol. 12, No. 1, 45–61
https://doi.org/10.21449/ijate.1472939
journal homepage: https://dergipark.org.tr/en/pub/ijate
Research Article

How valid and reliable are teachers' assessments of gifted students?

Sümeyye Arkan[1*], Sema Tan[2]

[1]Zonguldak Bülent Ecevit University, Eregli Faculty of Education, Department of Special Education, Türkiye
[2]Sinop University, Faculty of Education, Department of Special Education, Türkiye

ARTICLE HISTORY
Received: Apr. 24, 2024
Accepted: Nov. 19, 2024

Keywords:
Questionnaire,
Scale,
Systematic literature review,
Teacher evaluation.

Abstract: Teachers' perceptions, attitudes, and opinions about students, curricula, or evaluation methods contribute to the development of students' talents. Thus, researchers often collect data from teachers to identify gifted students, determine educational practices to meet the students' needs and assess gifted education programs. Researchers often develop measurement tools or utilize existing ones to collect valid and reliable data from teachers. This systematic literature review screened online databases to investigate measurement tools for teachers developed from 2017 to 2024. We combined the keywords "scale", "instrument", "questionnaire", "inventory", "gifted," and "teacher" to screen Web of Science (WoS) and Scopus databases. We categorized the measurement tools based on their intended use and analyzed seventeen instruments across themes including identification/nomination, attitude-behavior-perception, and knowledge and opinion. Nearly half of these studies employed exploratory or confirmatory factor analysis for construct validity, although some relied on the more superficial face validity. Overall, the studies demonstrated high reliability and validity, but simple analyses should be repeated to further enhance the robustness of measurement instruments.

1. INTRODUCTION

Assessment is considered one of the basic building blocks used in special education to collect information from students (Lockwood *et al.*, 2021). Therefore, researchers develop and use measurement tools to identify students' educational needs and psychomotor characteristics and assess and explore their many characteristics (Maison *et al.*, 2020). In order to achieve this purpose, researchers develop scales, inventories, and questionnaires according to the field of study and research topic. Developing a measurement tool to measure a particular construct correctly takes a long time. Therefore, they sometimes use existing valid and reliable measurement tools (Güngör, 2016). According to Karakoç and Dönmez (2014), researchers interested in obtaining a valid and reliable measurement tool should study and interpret an existing or developed scale according to many criteria and standards. Furthermore, the American Psychological Association (APA, 2014) has published standards for scales developed

*CONTACT: Sümeyye ARKAN ✉ sumeyyearkan97@gmail.com Zonguldak Bülent Ecevit University, Eregli Faculty of Education, Department of Special Education, Türkiye

e-ISSN: 2148-7456

in education and psychology. In addition, the APA 7 publication guidelines include content on reporting validity and reliability for qualitative and quantitative research (APA, 2020).

Researchers have developed and continue to refine intelligence tests, creativity scales, teacher evaluation instruments, and psychometric assessments to gauge the achievements of gifted students (Acar *et al*., 2016; Kaufman *et al*., 2011; Peters & Gentry, 2010; Renzulli *et al*., 2021; Sak *et al*., 2016). Research pertaining to the development of measurement tools in gifted education is predominantly categorized under the subfield of identification, given that student identification remains one of the most extensively studied areas in this domain (Dai *et al*., 2011). In order to gather information about students, researchers employ various measurement tools, including self-report instruments (Şencan, 2003) and criterion-referenced assessments (Renzulli, 2011). Moreover, researchers frequently engage teachers in the nomination process, as teachers offer vital, albeit potentially biased, insights into student performance and the effectiveness of teaching and evaluation processes (Siegle *et al*., 2011). Consequently, researchers focusing on teaching (Nel *et al*., 2011; Österling & Christiansen, 2022) and gifted education (Bildiren & Kargın, 2019; Idsøe *et al*., 2022; McCoach & Siegle, 2007; Park *et al*., 2016) routinely consider and assess teachers' opinions, attitudes, and competencies.

Numerous assessment tools have been developed for teachers including specialized instruments for teachers working with gifted students. A few researchers have systematically examined the assessment instruments developed for gifted students. Jarosewich *et al*. (2002) examined three assessment scales: Gifted and Talented Evaluation Scales (Gilliam *et al*., 1996); Gifted Evaluation Scale (McCarney & Anderson, 1989); and Scales for Rating the Behavioral Characteristics of Superior Students (Renzulli *et al*., 1976; Renzulli *et al*., 1997). They examined them in detail regarding subscales, age range, duration, and validity and reliability analysis. They found that within the nomination scales, students could be screened based on federal definition which includes and relates to giftedness, leadership, artistic talent, or creativity. In addition, the internal consistency and test-retest reliability of these scales were generally adequate, but the inter-rater reliability of scales is not adequately reported. Also, the researchers concluded that validity of scales (content and construct) was limited. Cao *et al*. (2017) conducted a literature review on assessing gifted students between 2005 and 2016. They categorized the types of assessments used in the research published between these years. They concluded that there had been advances in assessment over the years, and several assessment tools have been developed. Farah and Chandler (2018) examined eight measurement tools used for observation. They conducted a detailed review of the instrument's purpose, validity and reliability analysis, and development process. They underlined the need of a new instrument for observation. Pfeiffer and Jarosewich (2007) looked at giftedness multidimensionally and analyzed a teacher rating scale already developed for identification (The Gifted Rating Scales-School; GRS-S). They concluded that it was a valid screening scale, and that this analysis could provide additional support for the test manual. These instruments, in conjunction with other measurement tools such as tests and surveys, provide a framework for the collection of quantitative data in the field of educational research.

Researchers employ various measurement tools to gather quantitative data, which can be categorized into tests, surveys, and scales within the framework of measurement tools (Terzi, 2020). Surveys serve as effective research methods for comparing participants' knowledge, attitudes, beliefs, and behaviors (Woodcock, 2011), while scales are commonly utilized to measure abstract concepts like attitudes. Likert-type scales are generally developed to explore latent variables such as attitudes, fears, and perceptions (Terzi, 2020). Although surveys and scales are often used interchangeably, surveys offer the advantage of studying interrelationships among multiple topics. Many surveys integrate one or more scales as separate sections, which are then analyzed together or separately.

Tests, on the other hand, are typically designed to assess knowledge or skill (Trochim *et al*., 2016), emphasizing the importance of field-specific evaluation of these instruments. While researchers have systematically reviewed measurement instruments used in gifted identification and classroom observation (Cao *et al*., 2017; Jarosewich *et al*., 2002), there remains a gap in the literature regarding systematic reviews of tools developed to examine and assess teachers' views, attitudes, or competencies.

Researchers have employed various methods to explore a range of measurement tools and select the most appropriate ones for data collection. One such method is the systematic review, defined as a scientific process guided by precise and rigorous guidelines to ensure comprehensiveness, impartiality, accountability, and transparency in both methodology and execution (Dixon-Woods, 2016). Rammsted and Matthias (2019) argue that systematic literature reviews and meta-analyses should evaluate quality indicators, such as objectivity, reliability estimates, construct validity, factorial validity, and predictive validity of measurement instruments.

One advantage of systematic reviews is their ability to identify the strengths and weaknesses within the literature on a particular topic (Cook & West, 2012). While previous systematic reviews have examined measurement instruments for gifted identification and classroom observation (Cao *et al*., 2017; Jarosewich *et al*., 2002), a gap remains in the literature regarding reviews of tools designed to assess teachers' views, attitudes, or competencies.

To address this, we conducted a systematic review of teacher-focused instruments for assessing gifted students. The rationale for including publications from 2017 to 2024 is that Cao *et al*. (2017) conducted an analysis of publications up to 2016. Our goal was to document the validity and reliability of teachers' assessments when evaluating gifted students and to provide a roadmap for researchers interested in evaluating teachers' opinions, attitudes, or competencies. By examining the measures identified in this review, researchers can adapt the tools to suit their needs and gain insights into the subject areas most commonly involving teachers.

In this context, the following research questions guided our systematic literature review:

Research Question 1: What measurement tools, such as scales, instruments, questionnaires, and inventories, were developed between 2017 and 2024 for assessing gifted students, specifically designed for use or engagement by teachers?

Research Question 2: What validity and reliability criteria do researchers report when they develop a new measurement tool intended for use or engagement by teachers in assessing gifted students?

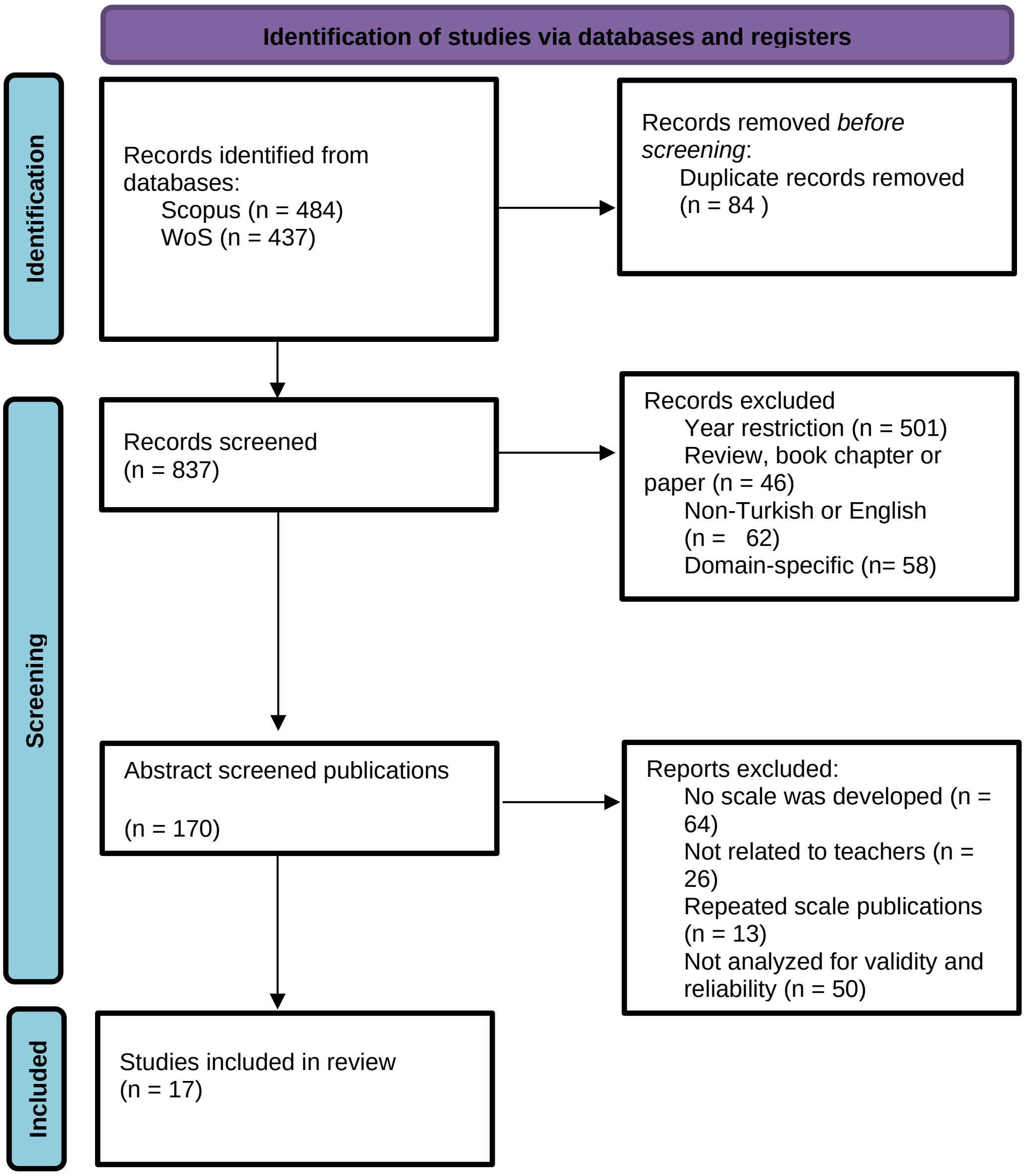
2. METHOD

We conducted a systematic literature review to examine the measurement tools developed for teachers in the gifted literature. The systematic literature review was based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA, 2021). We conducted a literature search in Web of Science and Scopus, databases between November-December 2022, January 2023 and March 2024 based on the keywords "scale," "instrument," "questionnaire," "inventory," "gifted," "teacher". The literature review yielded 921 publications.

We set inclusion and exclusion criteria to examine the publications in detail. The inclusion criteria were as follows: a) having been published between 2017 and 2024, b) written in Turkish or English, c) full text available, d) published in a peer-reviewed journal, d) a measurement tool developed for teachers, and e) validity and reliability research. The exclusion criteria were as follows: a) publications published before 2017, b) not in the field of giftedness, c) systematic review, meta-analysis, book chapter, and paper. Figure 1 shows the PRISMA diagram for the screening according to the criteria.

Furthermore, this systematic literature review did not include intelligence tests and nomination scales for three reasons. First, researchers have conducted test reviews. Second, the instructions and contents of intelligence tests are usually published in book form. Third, several other researchers have previously conducted systematic literature reviews to evaluate nomination scales (Jarosewich *et al*., 2002; Pfeiffer & Jarosewich, 2007). Therefore, we only included nomination instruments developed for teachers (Alnaim, 2023; Bildiren & Kargın, 2019; Idsøe *et al*., 2022).

Figure 1. *PRISMA diagram.*


Identification of studies via databases and registers
Identification
Records identified from databases:
Scopus (n = 484)
WoS (n = 437)
Records removed *before screening*:
Duplicate records removed (n = 84)
Screening
Records screened (n = 837)
Records excluded
Year restriction (n = 501)
Review, book chapter or paper (n = 46)
Non-Turkish or English (n = 62)
Domain-specific (n= 58)
Abstract screened publications (n = 170)
Reports excluded:
No scale was developed (n = 64)
Not related to teachers (n = 26)
Repeated scale publications (n = 13)
Not analyzed for validity and reliability (n = 50)
Included
Studies included in review (n = 17)

2.1. Data Collection

Seventeen publications were recruited according to the inclusion and exclusion criteria. In the coding procedure, we used inclusion and exclusion criteria to select a publication that is appropriate for our research aim and questions. We examined and included the publications according to whether they were published in gifted education literature, whether the measurement tool was developed, and reported validity and reliability criteria. We identified 921 publications in our initial database search. First, duplications were removed, reducing the number of publications to 837. Second, the remaining publications were restricted to Turkish and English languages, and keywords were checked in the titles and abstracts. The publications that were unsuitable for the review were removed, reducing the number of publications to 170. Third, the abstracts and method sections of the remaining publications were examined in detail. The publications that did not employ a measurement tool, did not have a full text, and did not have a measurement tool for teachers were removed. We checked whether the publications that developed measurement tools conducted validity and reliability analyses according to the criteria in Table 1. Self-report instruments are generally reflective scales. In this context, studies reporting at least one of the validity and reliability analysis criteria needed for reflective scales were included in the sample.

Table 1. *Validity and reliability criteria (Şencan, 2003).*

Validity	Reliability
Face	Split-half
Content	Item total score correlation
Nomological network	Cronbach Alpha coefficient
Concurrence	Parallel form
Predictive validity	Test-retest
Factor analysis within the framework of construct validity	Exploratory common factor analysis
Merger-separation	
Multiple feature multiple methods	

2.2. Data Analysis

We analyzed the publications descriptively with the aim of providing readers with a comprehensive source of information on the measurement tools, including their strengths and limitations, to help them make informed decisions when selecting a tool that is appropriate for their specific needs and context. For this reason, after determining what the measurement tools we examined were used for, we analyzed these measurement tools thematically according to their intended use. Therefore, the sample consisted of 17 publications (see Table 2). Although there are many types of measurement tools, we only included 17 publications because one of our objectives was to reveal the validity and reliability of the measurement tools. This is because researchers do not conduct validity and reliability analyses for inventories, questionnaires, and instruments. In the findings section, we reported the measurement tools, their purpose, sample, and validity and reliability analyses in more detail.

Table 2. *Reviewed publications.*

N	Publication	Measurement Tool	Classification	Purpose of Use	Sample	Validity	Reliability
1	Alnaim (2023)	Special Questionnaire	Survey	Identification/Nomination	108 teachers of gifted students	Face validity was reported.	Internal reliability (Cronbach's alpha)
2	Cheung *et al.* (2022)	Teacher Behavior Scale (TBS) Teacher Attitude Scale (TAS) Teacher Knowledge Scale (TKS)	Scale	Behavior Attitude Knowledge	2031 teachers (not specified)	EFA/CFA and factor loadings were reported for the developed scales. Same datasets were used for factor analysis.	KR-20 was reported.
3	Szymanski *et al.* (2022)	Determining Attitudes Toward Ability (DATA)	Scale	Attitude	350 teachers (not specified)	Construct validity was reported	Internal reliability (Cronbach's alpha)
4	Goddard & Evans (2018)	Teacher Attitudes	Survey	Attitude	50 elementary school teachers	Face validity was reported.	Internal reliability (Cronbach's alpha)
5	Idsoe *et al.* (2022)	Teacher Nomination Scale Parent Nomination Scale	Scale	Identification/Nomination	Parents and teachers of 243 students	PCA, CFA and concurrence validity were reported. Different datasets were used for factor analyses.	Inter-item correlation was calculated.
6	Al-Mamari *et al.* (2020)	Self-Awareness Scale (SAS)	Scale	Belief	60 teachers of students with LD	Face validity was reported.	Internal reliability (Cronbach's alpha)
7	Kandemir *et al.* (2019)	Creative Teaching in Mathematics Class scale	Scale	Behavior	423 math teachers	EFA/CFA, convergent and discriminant validity were reported. Different datasets were used for factor analyses.	Internal reliability (Cronbach's alpha)
8	Bildiren & Kırgın (2019)	Nomination Form	Survey	Identification/Nomination	Pre-school teachers	Face validity and factor loadings were reported.	KR-20 was reported.
9	Alshammari & Rababah (2019)	Scale for Teachers to Identify Gifted Students with Learning Disabilities in the Primary Stage	Scale	Identification/Nomination	Developed for elementary school teachers	Content, concurrence, factor, construct, and discriminatory validity were reported. Same datasets were used for factor analyses.	Test-retest Cronbach Alfa
10	Jarrah & Almarashdi (2019)	Teachers' perceptions toward their competency to teach gifted and talented students	Scale	Perception	66 math teachers	Face and content validity were reported.	Internal reliability (Cronbach's alpha)
11	Dağlıoğlu *et al.* (2019)	Classroom Practices in Inclusive Preschool Education	Scale	Belief	156 pre-school teachers	EFA and CFA were reported. Same datasets were used for factor analyses.	Internal reliability (Cronbach's alpha)

		Environment with Talented and Gifted Children Scale					
12	Gonzalez & Jung (2021)	Survey	Survey	Attitude	252 elementary school teachers	Construct validity and factor loadings were reported.	Internal reliability (Cronbach's alpha)
13	Westphal *et al.* (2017)	Perceived Knowledge About Grade Skipping		Belief			Internal reliability (Cronbach's alpha)
		Acceptance of Grade Skipping	Survey	Attitude	316 teachers (not specified)	Content and factor validity were reported.	Internal reliability (Cronbach's alpha)
		Beliefs About Students Development After Skipping a Grade		Belief			
14	Dersch *et al.* (2022)	The Math-Gender Misconception Questionnaire	Survey	Knowledge	303 teachers (different spezialization)	Construct validity and factor structure were reported.	McDonald's omega was reported.
15	Aljughaiman *et al.* (2017)	The Profile of Gifted Students	Instrument	Identification/Nomination	195 gifted student teachers and elementary school teachers	Content validity was reported.	Test-retest was applied.
16	Weyns *et al.* (2021)	Likability Emotional Demand Questionnaire	Scale Questionnaire	Belief Belief	522 teachers in training	Item loadings and PCA were reported. .	Internal reliability (Cronbach's alpha)
17	Wadaani (2023)	Math Teachers' Attitudes Toward Nurturing Creativity	Scale	Attitude	93 math teachers	Content validity was reported.	Internal reliability (Cronbach's alpha)

Note. EFA=Expolary Factor Analyses, CFA=Confirmatory Factor Analyses, PCA=Principal Component Analyses, KR-20=Kuder-Richardson 20.

3. RESULTS

The publications were analyzed according to the intended use of the measurement tools. Table 3 presents the themes developed from these analyses. Some researchers created more than one measurement tool. In total, eighteen tools were grouped under six distinct themes.

Table 3. *Themes created according to purposes of using measurement tools.*

Theme	Measurement Tools
Identification/Nomination	Special Questionnaire (Alnaim, 2022)
	Teacher Nomination Scale (Idsoe *et al*., 2022)
	Nomination Form (Bildiren & Kargın, 2019)
	Scale for Teachers to Identify Gifted Students with Learning Disabilities in the Primary Stage (Alshammari & Rababah, 2019)
	The Profile of Gifted Students (Aljughaiman *et al*., 2017)
Behavior	Teacher Behavior Scale (TBS) (Cheung *et al*., 2022)
	Creative Teaching in Mathematics Class (Kandemir *et al*., 2019)
Attitude	Acceptance of Grade Skipping (Westphal *et al*., 2017)
	Teacher Attitude Scale (TAS) (Cheung *et al*., 2022)
	Determining Attitudes Toward Ability (DATA) (Szymanski *et al*., 2022)
	Teacher Attitudes (Goddard & Evans, 2018)
	Survey (Gonzalez & Jung, 2021)
	Math Teachers' Attitudes Toward Nurturing Creativity (Wadaani, 2023)
Perception	Teachers' perceptions toward their competency to teach gifted and talented students (Jarrah & Almarashdi, 2019)
	Perceived Knowledge About Grade Skipping (Westphal *et al*., 2017)
Knowledge	Teacher Knowledge Scale (TKS) (Cheung *et al*., 2022)
	The Math-Gender Misconception Questionnaire (Dersch *et al*., 2022)
Belief	Beliefs About Students Development After Skipping a Grade (Westphal *et al*., 2017)
	Classroom Practices in Inclusive Preschool Education Environment with Talented and Gifted Children Scale (Dağlıoğlu *et al*., 2019)
	Likability and Emotional Demand Questionnaire (Weyns *et al*., 2021)
	Self-Awareness Scale (SAS) (Al-Mamari *et al*., 2020)

The studies reviewed span several countries, with a notable frequency in research from Saudi Arabia (Aljughaiman *et al*., 2017; Alnaim, 2022; Alshammari & Rababh, 2022; Jarrah, 2022), followed by Turkiye (Bildiren & Kargın, 2022; Dağlıoğlu, 2022; Kandemir *et al*., 2019), and Germany (Dersch, 2022; Westphal *et al*., 2022; Weyns *et al*., 2022). Other countries represented include China (Cheung *et al*., 2022), the USA (Szymanski, 2022; Wadaani, 2022), Australia (Goddard & Evans, 2022), Norway (Idsøe *et al*., 2022), Oman (Al Mamari, 2022), and Mexico (González Jung, 2022). Teacher specializations include mathematics teachers (Jarrah, 2022; Kandemir *et al*., 2019; Wadaani, 2022), primary school teachers (Goddard & Evans, 2022; González Jung, 2022), and preschool teachers (Bildiren & Kargın, 2022; Dağlıoğlu, 2022). There is also research on teachers of gifted students and students with learning disabilities (Alnaim, 2022; Al Mamari, 2022; Aljughaiman *et al*., 2017). Several studies did not specify the type of teachers involved (Alshammari & Rababh, 2022; Idsøe *et al*., 2022; Szymanski, 2022; Westphal *et al*., 2022).

3.1. Identification/Nomination

Assessment is critical to meeting the educational needs of gifted students. Researchers often focus on this topic and use different assessment tools to evaluate gifted students. This theme documented and analyzed assessment tools developed for the purpose of identification/nomination, intended for use by teachers to help identify students for further evaluation.

Researchers developed the Special Questionnaire (Alnaim, 2022) and the Scale for Teachers to Identify Gifted Students with Learning Disabilities (Alshammari & Rababah, 2019) for teachers to use to nominate students. The items in the Scale for Teachers to Identify Gifted Students with Learning Disabilities were based on the Al-Hajri (2015) scale, which was developed to determine giftedness/learning disability and the characteristics of gifted students with learning disabilities in the literature (Alshammari & Rababah, 2019). The Special Questionnaire (Alnaim, 2022) items, on the other hand, were created based on qualitative data collected through interviews with teachers about the challenges faced by gifted people with ADHD and the relevant educational literature. Alnaim (2022) also established content validity and calculated Cronbach's alpha coefficient for reliability (.761-.926) for the Special Questionnaire. Bildiren and Kargın (2019) developed and used the Nomination Form to enable teachers to guide students in a program. In the process of developing the form, a comprehensive review of the pertinent national and international literature was undertaken by the researchers to inform the selection of the items. Following this, the form was subjected to a rigorous assessment by a panel of experts to gauge its content validity and ensure its adequacy for the intended purpose. The last form consists of 14 items and two subscales. The researchers reported factor loadings and assessed internal consistency for reliability (KR-20=.92).

Idsoe *et al*. (2022) aimed to nominate students for a project. To do this, they developed and analyzed the Teacher Nomination Scale and the Parent Nomination Scale. The instrument has seven items that are rated on a four-point Likert-type scale. Firstly, researchers reviewed the existing scales in the literature and after that, they examined the characteristics identified by professionals in the field to decide on the scale items. They modified these scales according to the local screening instruments for parents and teachers because the Norwegian Early Childhood Education and Care (ECEC) system does not include cognitive tasks that could reveal high intellectual abilities among these children. The items on this scale correlate more than those developed for parents. They included these correlations under the heading of concurrent validity. They used confirmatory factor analysis to explain the items' mean and standard deviation scores. The scale developed for teachers is more consistent for screening purposes.

Aljughaiman *et al*. (2017) developed gifted student profiles and then presented them to teachers. They ask teachers to nominate the eight profiles which constitute of giftedness behavior. Their aim was revealing which student was suitable for the identification. For the content validity they presented the cases to the seven professors from giftedness and creativity domain. For the reliability of the cases, they used test-retest reliability coefficient (.81). Teachers'nominations of students were biased towards students who achieved high grades, while students who achieved low grades were disregarded.

Based on scales and questionnaires in the identification/nomination theme, researchers have developed valid and reliable tools for teachers. Since these tools have demonstrated both reliability and validity, it can be concluded that they are practical and suitable for use in the identification/nomination process.

3.2. Attitude-Behavior-Perception

The scales developed by the studies included in our sample are Likert-type scales. Cheung *et al*. (2022) developed the Teacher Attitude Scale (TAS) to obtain teachers' views on gifted students. The scale consists of 12 items and three subscales: teacher support, attitude toward gifted education, and support for gifted education. All but one of the items were normally distributed. Therefore, the researchers conducted an exploratory factor analysis on 17 items. They removed five items from the scale because they had low factor loadings.

Cheung *et al*. (2022) developed the Teacher Behavior Scale (TBS) to assess teachers' instructional practices in three dimensions: nurturing gifted students, differentiated instruction, and learning support for undiagnosed students. This scale also has 12 items loaded on three dimensions. The researchers conducted exploratory and confirmatory factor analyses. They reported factor loadings for the items in the subscales. For reliability, they calculated pretest and posttest Cronbach's alpha values for each subscale (.84-.71, .75, .85, .78-.79).

Jarrah and Almarashdi (2019) developed a measurement tool for perception measurement due to reviewing the literature on giftedness. They conducted pilot studies and used the scales to measure teachers' teaching-related competencies. In their survey, the researchers used 19 statements. For content validity, they sent the scale to six faculty members specializing in gifted education and math education. Additionally, the scale was reviewed by eight specialists, including mathematics teachers and supervisors, for feedback and comments. The scale measured teachers' perceptions using two subscales: (1) the Competency to Teach Gifted and Talented Students scale (nine items) and (2) the Teaching Gifted and Talented Students scale (ten items). They reported Cronbach's alpha coefficient (.93), which indicated that the scales were highly reliable.

Westphal *et al*. (2017) used and developed several scales for grade-skipping among gifted students. They developed four scales. The items in the scales were drawn from the authors' experiences in teacher training for gifted education, as well as from the relevant research literature. Subsequently, researchers evaluated the items for content validity, specifically focusing on their clarity, comprehensibility, and whether they accurately reflected the intended construct. They presented them online to teachers to collect data. They used the Perceived Knowledge About Grade Skipping scale to assess teachers' perceptions of students' grade skipping. They reported internal consistency for reliability (Cronbach's alpha .86). For validity, they conducted an exploratory factor analysis on four items and removed one item from the scale. After examining teachers' attitudes toward gifted education, they developed the Acceptance of Grade Skipping scale because they needed another scale for the study. The four-item response scale measures teachers' attitudes toward grade-skipping for gifted students. For validity, they conducted an exploratory factor analysis. For reliability, they calculated internal consistency (Cronbach's alpha .89).

Szymanski *et al*. (2022) developed the Determining Attitudes Toward Ability (DATA) scale to measure teachers' attitudes toward various issues related to gifted education because no questionnaire provided a wide range of information about gifted students. The scale measures attitudes toward grade skipping, acceleration, diagnosis, and curriculum. The scales developed and used after the scale developed by Gagné and Nadeau (1991) were examined, and the items were decided accordingly. The DATA scale consists of 92 items rated on a four-point Likert scale. For content validity, the scale was reviewed by four domain experts and then teachers. The researchers conducted a pilot study for the DATA scale and administered it to 124 participants. They removed 18 items. The final version of the scale consists of five subscales and 74 items. The final version included both exploratory and confirmatory factor analyses. However, the researchers did not recommend the scale for use due to the low sample size and some low factor loadings.

Goddards and Evans (2018) developed the Teacher Attitudes questionnaire to examine pre-service teachers' attitudes toward inclusion. The questionnaire has two parts. The first part consists of questions about demographic characteristics. The second part consists of three sub-dimensions to determine pre-service teachers' attitudes. According to the pilot study results, the questionnaire's final version consists of 40 items rated on a five-point Likert-type scale. They reported face validity for validity and internal consistency for reliability (Cronbach's alpha .761).

Gonzalez and Jung (2021) detected that we needed a questionnaire to determine teachers' attitudes toward acceleration. Therefore, they developed an 80-item questionnaire to assess attitudes toward acceleration and its predictors. They reported factor analyses for validity and reliability. They calculated Cronbach's alpha values for the subscales; support for acceleration (.64), communication with gifted students (.75), support from school administrators (.73), socio-emotional impact (.73), perception of elitism (.59), and self-perception of gifted students (.82)

Kandemir *et al*. (2019) argued that creativity is content-based, and measuring the behaviors that promote teachers' discipline-specific creativity is important. Therefore, they developed a scale with six subscales. They developed the scale for mathematics and aimed to assess teachers' behaviors. The final version of the scale consists of 31 items. The scale has high factor loadings, which indicates validity. They calculated Cronbach's alpha values of the subscales for reliability. Cronbach’s alpha coefficients demonstrated the scale’s reliability, with values of .91 for Teaching Style, .88 for Confidence, .91 for Classroom Climate, .74 for Overcoming Barriers, .75 for Asking Questions, and .89 for Innovative Teaching Practices.

Weyns *et al*. (2021) used two additional scales, in addition to previously developed and implemented questionnaires. One of these scales assessed likability, using a self-constructed questionnaire consisting of three items: 'I like him/her', 'I would like to spend time with him/her', and 'I would like to teach him/her'. Principal component analysis was used, and the results showed that all items had loadings above 0.40. The Likability scale's reliability was reported as 0.74 using Cronbach's alpha. Another questionnaire, the Emotional Demand, also reported a reliability of 0.75 using Cronbach's alpha. This questionnaire aimed to measure how engaged the student was and what their feelings were towards the student.

Wadaani (2023)’s questionnaire comprises sections for collecting data on preservice education and professional development independent variables, evaluating teachers' attitudes towards creativity and mathematics gifted education, and assessing the availability of support features for enhancing creativity and developing mathematical giftedness. Participants rated their level of agreement with statements using a 5-point Likert scale. The questionnaire’s validity was ensured by connecting items to relevant literature and utilizing existing validated instruments. Refinement of the instrument was achieved through feedback from teachers and experts, as well as focus group discussions. The instrument's reliability was assessed using Cronbach's Alpha coefficient, resulting in a high value of 0.88 for the overall scale, indicating its reliability. Item-total statistics showed that no item significantly affected the reliability.

The measurement tools categorized under this theme were developed to address the need for new tools in the assessment of teachers' attitudes, behaviors, and perceptions towards educational programs for gifted students. The primary purpose of these tools was to evaluate the effectiveness of these programs. Factor analyses and assessments of internal consistency reliability were conducted to ensure the reliability of the items in these measurement tools.

3.3. Knowledge

In the study conducted by Cheung *et al*. (2022), multiple measurement tools were developed, and the same samples were utilized in these tools. The Teacher Knowledge Scale (TKS) is another measurement tool that was developed. Teacher Knowledge Scale (TKS) based on myths about gifted students. Teachers evaluate myths as true-false-don't know. The scale consists of 10 items. The researchers reported content validity. For reliability, they calculated the pretest and posttest KR-20 internal consistency (KR-20=0.44, 0.52).

Dersch *et al*. (2022) developed the Math-Gender Misconception Questionnaire to examine whether three potential misconceptions about giftedness are related to theoretically relevant constructs. The questionnaire consists of 30 items rated on a five-point Likert-type scale. Fifteen items address misconceptions related to mathematics-gender, while the remaining items address misconceptions related to research hypotheses. The researchers reported McDonald's Omega for reliability. The empathizing-systemizing (ω = .88) and compensating for girls (ω = .76) subscales have good reliability, while the noncompensating for girls (ω = .72) subscale has acceptable reliability. They conducted a factor analysis for validity and found that all three factors were consistent.

The study of myths is a common focus in the field of giftedness, with researchers attempting to assess the level of knowledge of teachers and individuals in this area (Kaya *et al*., 2015; O'Connor, 2012; Sak, 2011). To assess teachers' comprehension of myths, Dersch *et al*. (2022) developed measurement tools that concentrate on the connection between mathematics and gender. Similarly, Cheung *et al*. (2022) created measurement tools to assess general myths in the field.

3.4. Opinion

Westphal *et al*. (2017) created another scale to assess teachers' opinions about the potential impact of grade skipping on students' development. As stated above, the scale's items were developed from the authors' experiences in teacher training for gifted education and from the existing research literature. Researchers then evaluated the items for content validity. The scale consists of 17 items rated on a four-point Likert-type scale. Westphal *et al*. (2017) performed an exploratory factor analysis to evaluate the scale's validity. For reliability, they calculated Cronbach's alpha coefficients, which were .71 for opinions and .86 for academic development.

Dağlıoğlu *et al*. (2019) developed a scale to explore the teaching approaches applied by teachers in inclusive preschool classrooms with typically developing and gifted children together. The instrument consists of 22 items rated on a five-point Likert-type scale. The confirmatory factor analysis showed that the scale agreed with the model. One of the project's aims was to identify the primary educational and instructional elements preschool teachers use in inclusive education settings. Based on this purpose, the researchers constructed items on the educational and instructional elements that preschool teachers use in inclusive education settings. For validity, they conducted an exploratory factor analysis and confirmatory factor analysis and reported factor loadings. They calculated Cronbach's alpha for reliability. The whole scale has a Cronbach's alpha of .88, while the first, second, and third subscales have Cronbach's alpha values of .76, .83, and .80, respectively.

Al-Mamari *et al*. (2020) developed the Self-Awareness Scale to assess teachers' awareness of gifted students with learning difficulties. They concluded that the scale was suitable enough to assess teachers' awareness. They reported face validity for validity and calculated Cronbach's alpha for reliability. The "knowledge awareness," "skill awareness," and "individual awareness" have Cronbach's alpha values of .94, .96, and .95, respectively.

This theme explored three measurement tools designed to evaluate and assess teachers' opinions. In the educational literature, teachers are frequently consulted for their opinions on various topics within the field. Therefore, it is crucial to be familiar with existing measurement tools to facilitate their reuse.

4. DISCUSSION and CONCLUSION

This systematic literature review examined measurement tools designed for teachers to evaluate gifted students. The sample included 17 publications featuring 13 scales, 7 questionnaires, and one other instrument. We analyzed the validity and reliability of these tools across six thematic areas. The results show that researchers have generally developed measurement tools for teachers in the theme of "attitude." Researchers have developed up-to-date, valid, and reliable scales to replace the scale previously developed by Gagne and Nadeau (1991), which provided a wide range of assessment opportunities for teachers and parents in the field of giftedness

(Jarrah & Almarashdi, 2019; Szymanski *et al*., 2022). This scale allowed for the assessment of attitudes in gifted children across various dimensions, including needs and support, resistance to objections, social value, rejection, ability grouping, and school. Researchers have also developed measurement tools related to "identification-nomination." Researchers have developed only two measurement tools under the themes of "perception" and "behavior."

The majority of scales in our sample assess educational adaptations (*n* = 8) (Cheung *et al*., 2022; Idsøe *et al*., 2021; Westphal *et al*., 2017). Given that the literature primarily focuses on identification in studies of giftedness (Dai *et al*., 2011), one might expect researchers to concentrate on identification when developing measurement tools for teachers. However, our findings revealed that even though several measurement tools related to identification-nomination were designed for the teachers, the researchers also mainly focused on other topics such as attitudes as well.

Farah *et al*. (2018) conducted a similar systematic literature review and focused on publications that did not conduct validity and reliability analysis. The studies we reviewed, including Cheung *et al*. (2022), Alshammari and Rababah (2019), and Dağlıoğlu *et al*. (2019), used the same dataset for exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). Furthermore, several studies have emphasized the importance of using different datasets for exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) (Hurley *et al*., 1997: Knekta *et al*., 2019). Only the studies of Idsoe *et al*. (2022) and Kandemir *et al*. (2019) in our sample conducted factor analyses using different datasets. Therefore, it is safe to say that out of the studies employing factor analysis as a method, Idsøe *et al*. (2022) and Kandemir *et al*. (2019) followed a more methodologically sound approach than the others. Currently, the most widely used index for assessing scale reliability is Cronbach's coefficient alpha (Raykov& Marcoulides, 2019). Our results showed that most publications have calculated Cronbach's alpha to report internal reliability. According to Nunnally (1978), a scale and its subscales with a Cronbach's alpha coefficient of .70 and above are reliable. In this context, most publications in our sample have reported Cronbach's alpha values above .70. Studies that did not report Cronbach's alpha were evaluated for reliability using KR-20, McDonald's omega, and test-retest methods. Almost half of the publications have conducted exploratory and confirmatory factor analyses for construct validity (*n* =12) or they only checked for face validity (*n* =6). However, face validity is the most superficial level of validity. Şencan (2003) suggests that researchers report construct validity for more robust validity analyses. In general, the researchers have reported high validity and reliability. However, researchers should repeat simple analyses to increase the validity and reliability of their instruments.

Among the studies we reviewed for this study, we found that measurement tools were generally developed for attitude (*n* =6) and identification/nomination (*n* =6) purposes. Identification represents a particularly prominent topic within the field of giftedness literature (Dai *et al*., 2011), reflecting a clear research focus on this area and the consequent development of measurement tools for educators. Furthermore, the evaluation of an individual's beliefs, attitudes and perceptions regarding various aspects of education, including courses, enrichment activities and differentiation activities, represents another key area of interest within gifted education (Akgül, 2021; Kim, 2016; Laine *et al*., 2019). A bibliometric analysis of these tools could help clarify their overall distribution more effectively. This would enable a more detailed examination of the current measurement tools developed for teachers of gifted students, using an alternative method. The measurement tools within these themes can also be applied in other studies to assess teachers' attitudes, behaviors, and perceptions. Further comments on the validity and reliability of these tools can be made in the future.

As suggested by Rammsted and Matthias (2019), researchers should conduct meta-analyses to quantitatively analyze the validity and reliability of measurement tools for teachers in the field of giftedness. In addition, researchers should conduct both systematic literature reviews and meta-analyses for the validity and reliability analysis of measurement tools developed for gifted students. In this way, they can evaluate the objectivity of scales and the criteria for measuring instruments. An in-service training can be designed to help teachers to choose appropriate assessment tools for solid evaluation. Teachers should be trained in research methods to help them design appropriate interventions and develop or select measurement tools tailored to their specific needs, rather than relying solely on pre-developed tools.

This study has several limitations. First, only the Scopus and Web of Science databases were accessed, which may restrict the range of relevant studies. Additionally, publications after 2017 are limited in these databases, and as 2024 is not yet complete, the results may vary due to future additions.

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.

Contribution of Authors

Sümeyye Arkan: Literature Review, Conceptualization, Visualization, Formal Analysis, and Writing-original draft. **Sema Tan**: Methodology, Supervision and Critical Review.

Orcid

Sümeyye Arkan https://orcid.org/0000-0001-7788-5917
Sema Tan https://orcid.org/0000-0002-9816-8930

REFERENCES

Acar, S., Sen, S., & Cayirdag, N. (2016). Consistency of the performance and nonperformance methods in gifted identification: A multilevel meta-analytic review. *Gifted Child Quarterly*, *60*(2), 81–101. https://doi.org/10.1177/0016986216634438

Akgül, G. (2021). Teachers' metaphors and views about gifted students and their education. *Gifted Education International*, *37*(3), 273-289. https://doi.org/10.1177/0261429421988927

Aljughaiman, A.M., & Ayoub, A.E.A. (2017). Giftedness in Arabic environments: Concepts, implicit theories, and the contributed factors in the enrichment programs. *Cogent Education*, *4*(1), 1364900. https://doi.org/10.1080/2331186X.2017.1364900

Al-Mamari, S.S., Al-Zoubi, S.M., Bakkar, B.S., & Al-Mamari, K.H. (2020). Effects of a training module on omani teachers' awareness of gifted students with learning disabilities. *Journal of Education and E-Learning Research*, *7*(3), 300-305. https://doi.org/10.20448/journal.509.2020.73.300.305

Alnaim, F.A. (2023). The services provided to students with attention deficit hyperactivity disorder in primary schools from the special education teachers' perspectives. *International Journal of Learning, Teaching and Educational Research*, *12*(4), 20-42. https://doi.org/10.36941/jesr-2022-0107

Alshammaria, M.M., & Rababahb, A.A. (2019). Developing a scale for teachers to identify gifted students with learning disabilities in the primary stage in the eastern province of Saudi Arabia. *International Journal of Innovation, Creativity and Change*, *8*(4), 128-154.

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). https://doi.org/10.1037/0000165-000

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bildiren, A., & Kargın, T. (2019). Proje temelli yaklaşıma dayalı erken müdahale programının üstün yetenekli çocukların problem çözme becerisine etkisi [The effect of early intervention programme based on project-based approach on gifted children's problem-solving skills]. *Education and Science, 44*(198), 343-360. http://doi.org/10.15390/EB.2019.7360

Cao, T.H., Jung, J.Y., & Lee, J. (2017). Assessment in gifted education: A review of the literature from 2005 to 2016. *Journal of Advanced Academics, 28*(3), 163-203.

Cheung, A.C.K., Shek, D.T.L., Hui, A.N.N., Leung, K.H., & Cheung, R.S.H. (2022).

Professional development for teachers of gifted education in Hong Kong: Instrument validation and training effectiveness. *International Journal of Environmental Research and Public Health, 19(15), 9433*. https://doi.org/10.3390/ijerph19159433

Cook, D.A. and West, C.P. (2012). Conducting systematic reviews in medical education: a stepwise approach. *Medical Education, 46*(10), 943-952. https://doi.org/10.1111/j.1365-2923.2012.04328.x

Dağlıoğlu, H.E, Ömeroğlu, E., Turupcu Doğan, A., Şahin, M.G., Sarıcı Bulut, S., Sabancı, O., ... Karataş, S. (2019). The reliability and validity study of 'classroom practices in inclusive preschool education environment with talented and gifted children scale. *Pegem Journal of Education and Instruction, 9*(2), 413-434. http://doi.org/10.14527/pegegog.2019.013

Dai, D.Y., Swanson, J.A., & Cheng, H. (2011). State of research on giftedness and gifted education: A survey of empirical studies published during 1998—2010 (April). *Gifted Child Quarterly, 55*(2), 126–138. https://doi.org/10.1177/0016986210397831

Dersch, A.S., Heyder, A. & Eitel, A. (2022). Exploring the nature of teachers' math-gender stereotypes: The math-gender misconception questionnaire. *Frontier in Psychology, 13*, 820254. https://doi.org/10.3389/fpsyg.2022.820254

Dixon-Woods, M. (2016). *Systematic reviews and qualitative studies.* D, Silverman (Ed.), Qualitative Research (4th ed.). Sage.

Farah, Y.N., & Chandler, K.L. (2018). Structured observation instruments assessing instructional practices with gifted and talented students: A review of the literature. *Gifted Child Quarterly, 62*(3), 276–288. https://doi.org/10.1177/0016986218758439

Gagné, F., & Nadeau, L. (1991). *Opinions about the gifted and their education.* Unpublished instrument.

Goddard, C., & Evans, D. (2018). Primary pre-service teachers' attitudes towards inclusion across the training years. *Australian Journal of Teacher Education, 43*(6), 122-142. https://doi.org/10.14221/ajte.2018v43n6.8

Güngör, D. (2016). Psikolojide ölçme araçlarının geliştirilmesi ve uyarlanması kılavuzu [Guide to the development and adaptation of measurement instruments in psychology]. *Turkish Psychology Writings, 19*(38), 104-112.

Idsøe, E., Campbell, J., Idsøe, I., & Størksen, I. (2021). Development and psychometric properties of nomination scales for high academic potential in early childhood education and care. *European Early Childhood Education Research Journal*, 1-14. https://doi.org/10.1080/1350293X.2021.2007969

Jarrah, A.M., & Almarashdi, H.S. (2019). Mathematics teachers' perceptions of teaching gifted and talented learners in general education classrooms in the UAE. *Journal for the Education of Gifted Young Scientists, 7*(4), 835-847. http://doi.org/10.17478/jegys.628395

Jarosewich, T., Pfeiffer, S.I., & Morris, J. (2002). Identifying gifted students using teacher rating scales: A review of existing instruments. *Journal of Psychoeducational Assessment, 20*(4), 322–336. https://doi.org/10.1177/073428290202000401

Kandemir, M.A., Tezci, E., Shelley, M., & Demirli, C. (2019). Measurement of creative teaching in mathematics class. *Creativity Research Journal, 31*(3), 272-283. https://doi.org/10.1080/10400419.2019.1641677

Karakoç, F.Y.& Dönmez, L. (2014). Ölçek geliştirme çalışmalarında temel ilkeler [Basic principles of scale development]. *World of Medical Education, 13*(40), 39-49.

Kaufman, J.C., Plucker, J.A., & Russell, C.M. (2012). Identifying and assessing creativity as a component of giftedness. *Journal of Psychoeducational Assessment*, *30*(1), 60–73. https://doi.org/10.1177/0734282911428196

Kim, M. (2016). A meta-analysis of the effects of enrichment programs on gifted students. *Gifted Child Quarterly*, *60*(2), 102-116. https://doi.org/10.1177/0016986216630607

Knekta, E., Runyon, C., & Eddy, S. (2019). One size doesn't fit all: using factor analysis to gather validity evidence when using surveys in your research. *CBE life sciences education*, *18*(1). https://doi.org/10.1187/cbe.18-04-0064

Laine, S., Hotulainen, R., & Tirri, K. (2019). Finnish elementary school teachers' attitudes toward gifted education. *Roeper Review*, *41*(2), 76-87. https://doi.org/10.1080/02783193.2019.1592794

Lockwood, A.B., Farmer, R.L., Bohan, K.J., Winans, S., & Sealander, K. (2021). Academic achievement test use and assessment practices: A national survey of special education administrators. *Journal of Psychoeducational Assessment*, *39*(4), 436-451. https://doi.org/10.1177/0734282920984290

Maison, D., Astalini, D., Agus, K., & Sumaryanti, R.P. (2020). Supporting assessment in education: E-assessment interest in physics. *Universal Journal of Educational Research*, *8*(1), 89- 97. https://doi.org/10.13189/ujer.2020.080110

McCoach, D.B., & Siegle, D. (2007). What predicts teachers' attitudes toward the gifted? *Gifted Child Quarterly*, *51*(3), 246-254. https://doi.org/10.1177/0016986207302719

Nel, N., Muller, H., Hugo, A., Helldin, R., Bãckmann, Õ., Dwyer, H., & Skarlind, A. (2011). A comparative perspective on teacher attitude-constructs that impact on inclusive education in South Africa and Sweden. *South African Journal of Education*, *31*(1), 74-90.

Nunnally, J.C. (1978). *Psychometric theory. 2nd Edition*, McGraw-Hill.

Österling, L., & Christiansen, I. (2022). Whom do they become? A systematic review of research on the impact of practicum on student teachers' affect, beliefs, and identities. *International Electronic Journal of Mathematics Education*, *17*(4), em0710. https://doi.org/10.29333/iejme/12380

Palacios Gonzalez, P., & Jung, J.Y. (2021). The predictors of attitudes toward acceleration as an educational intervention: Primary school teachers in Mexico. *High Ability Studies*, *32*(1), 27–49. https://doi.org/10.1080/13598139.2019.1692649

Pajares, M.F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research*, *62*(3), 307-332.

Park, S., Callahan, C.M., & Ryoo, J.H. (2016). Assessing gifted students' beliefs about intelligence with a psychometrically defensible scale. *Journal for the Education of the Gifted*, *39*(4), 288–314. https://doi.org/10.1177/0162353216671835

Peters, S.J., & Gentry, M. (2010). Multigroup construct validity evidence of the HOPE scale: Instrumentation to identify low-income elementary students for gifted programs. *Gifted Child Quarterly*, *54*(4), 298–313. https://doi.org/10.1177/0016986210378332

Peters, S.J., & Gentry, M. (2013). Additional validity evidence and across-group equivalency of the HOPE teacher rating scale. *Gifted Child Quarterly*, *57*(2), 85-100. https://doi.org/10.1177/0016986212469253

Pfeiffer, S.I., & Jarosewich, T. (2007). The gifted rating scales-school form: An analysis of the standardization sample based on age, gender, race, and diagnostic efficiency. *Gifted Child Quarterly*, *51*(1), 39–50. https://doi.org/10.1177/0016986206296658

Rammstedt, B., & Bluemke, M. (2019). Measurement instruments for the social sciences. *Measurement Instruments for the Social Sciences*, *1*(1), 1-3. https://doi.org/10.1186/s42409-018-0003-3

Raykov, T., & Marcoulides, G.A. (2019). Thanks coefficient alpha, we still need you!, *Educational and Psychological Measurement*, *79*(1), 200-210. https://doi.org/10.1177/0013164417725127

Renzulli, J. (2010). *Scales for Rating the Behavioral Characteristics of Superior Students: Technical and Administration Manual (3rd ed.).* Routledge. https://doi.org/10.4324/9781003237808

Sak, U., Bal Sezerel, B., Ayas, B., Tokmak, F., Özdemir, N., Demirel Gürbüz, Ş., & Öpengin, E. (2016). *Anadolu Sak Zekâ Ölçeği (ASİS) uygulayıcı kitabı.* Anadolu Üniversitesi ÜYEP Merkezi.

Siegle, D., Moore, M., Mann, R.L., & Wilson, H.E. (2010). Factors that influence in-service and preservice teachers' nominations of students for gifted and talented programs. *Journal for the Education of the Gifted, 33*(3), 337-360. https://doi.org/10.1177/016235321003300303

Staff, A.I., Oosterlaan, J., van der Oord, S., Hoekstra, P.J., Vertessen, K., de Vries, R., van den Hoofdakker, B.J., & Luman, M. (2021). The validity of teacher rating scales for the assessment of ADHD symptoms in the classroom: A systematic review and meta-analysis. *Journal of Attention Disorders, 25*(11), 1578-1593. https://doi.org/10.1177/1087054720916839

Şencan, H. (2005). *Güvenilirlik ve geçerlilik* [*Reliability and validity*]. Seçkin Publishing.

Terzi, R. (2020). Nicel veri toplama teknikleri [Quantitative data collection techniques]. In S. Şen, & İ. Yıldırım (eds.), *Eğitimde araştırma yöntemleri* (2.ed, pp. 357-382), Nobel Publishing.

Uzunboylu, H., Akçamete, G., Sarp, N., & Demirok, M. (2022). Primary school teachers' opinions about gifted education programmes in distance education. *Sustainability, 14*, 17031. https://doi.org/10.3390/su142417031

Wadaani, M. (2023). The influence of preservice education and professional development in mathematics Teachers' attitudes toward nurturing creativity and supporting the gifted. *Journal of Creativity, 33*(1), 100043. https://doi.org/10.1016/j.yjoc.2023.100043

Westphal, A., Vock, M., & Stubbe, T. (2017). Grade skipping from the perspective of teachers in Germany: The links between teachers' decisions, acceptance, and perceived knowledge. *Gifted Child Quarterly, 61*(1), 73-86. https://doi.org/10.1177/0016986216670727

Weyns, T., Preckel, F., & Verschueren, K. (2021). Teachers-in-training perceptions of gifted children's characteristics and teacher-child interactions: An experimental study. *Teaching and Teacher Education, 97*, 103215. https://doi.org/10.1016/j.tate.2020.103215

Woodcock, S. (2011). A cross-sectional study of pre-service teacher efficacy throughout the training years. *Australian Journal of Teacher Education, 36*(10), 23-34. https://doi.org/10.14221/ajte.2011v36n10.1

Yetim-Karaca, S., & Türk, T. (2020). Ortaokul matematik dersi öğretim programının üstün yetenekli öğrencilerin eğitimi açısından öğretmen görüşlerine göre değerlendirilmesi [Evaluation of secondary school mathematics curriculum in terms of education of gifted students according to teachers' opinions]. *Turkish Journal of Computer and Mathematics Education, 11*(1), 241-279. https://doi.org/10.16949/turkbilmat.526817

International Journal of Assessment Tools in Education

2025, Vol. 12, No. 1, 62–77

https://doi.org/10.21449/ijate.1496193

journal homepage: https://dergipark.org.tr/en/pub/ijate **_Research Article_**

ChatGPT-3.5 as an automatic scoring system and feedback provider in IELTS exams

Xinming Chen[1*], Ziqian Zhou[2], Malila Prado[3]

[1]University College London, England
[2]The University of Hong Kong, Hong Kong (SAR) China
[3]BNU-HKBU United International College, China

ARTICLE HISTORY

Received: June 6, 2024

Accepted: Nov. 17, 2024

Keywords:
ChatGPT-3.5,
Automatic Essay Scoring,
AI Proofreader,
IELTS,
L2 learning.

Abstract: This study explores the efficacy of ChatGPT-3.5, an AI chatbot, used as an Automatic Essay Scoring (AES) system and feedback provider for IELTS essay preparation. It investigates the alignment between scores given by ChatGPT-3.5 and those assigned by official IELTS examiners to establish its reliability as an AES. It also identifies the strategies employed by ChatGPT-3.5 in revising essays based on the four IELTS rubrics: task achievement, coherence and cohesion, lexical resources, and grammatical range and accuracy. Based on pre-rated essays from an official IELTS preparatory book as a control measure to ensure objectivity, the findings indicate a discrepancy, with ChatGPT-3.5 typically assigning lower scores compared to official raters. However, ChatGPT-3.5 shows a robust capability to revise essays across all four descriptors. In addition, the effectiveness of ChatGPT-3.5 as a feedback provider may be attributed to the essay type and its widely accepted rubrics. Our study contributes to the understanding of the application of AI tools in second language writing and suggests that future studies should focus on evaluating the capacity and effectiveness of such tools in pedagogical applications.

1. INTRODUCTION

Even though technology is becoming increasingly present in education, it does not appear to have dramatically changed the way we teach. Though technology is at every teacher's disposal, old pedagogical concepts appear to meet the needs of most teachers (Chiu *et al.*, 2023). Regarding English as a Second Language (ESL) writing, there is resistance among teachers against employing technology such as Grammarly (Huang *et al.*, 2020), machine translation (Lee, 2023), or even using digitally available multilingual resources (Prado & Huggins, 2023). Chiu *et al.* (2023) report that "some teachers described the technologies as difficult to control, lacked an understanding of how the technologies operated, and were concerned about ethical issues, such as bias and breaches of privacy." This probably explains why the response to the launch of ChatGPT (Open AI, 2022) at the end of 2022 was not widely embraced in the education realm, particularly in higher education.

*CONTACT: Xinming CHEN ✉ q030025010@mail.uic.edu.cn University College London, Faculty of Education and Society, Department of Psychology and Human Development, England

e-ISSN: 2148-7456

As suggested by Bai *et al*. (2022), Artificial Intelligence (AI) applications in Education (AIEd) are a trending research topic. ChatGPT, an artificial intelligence chatbot, uses natural language processing to create humanlike conversations based on large amounts of digital content (Boa Sorte *et al.*, 2021; Pavlik, 2023; Fryer *et al.*, 2020). It can compose texts in a variety of written genres, including articles, social media posts, essays, and emails, all generated in a conversation-like style (Boa Sorte *et al.*, 2021). However, the introduction of ChatGPT in academia has sparked debates regarding authorship and concerns over plagiarism (Dergaa *et al.*, 2023) and raised the concern that teachers might be substituted (Warschauer *et al.*, 2023).

Yet ChatGPT is having a significant impact on language education research, particularly in second language (L2) writing (Artiles Rodríguez *et al.*, 2021; Barrot, 2023; Baskara, 2023; Dergaa *et al.*, 2023; Han *et al.*, 2023; Warschauer *et al.*, 2023). Four major advantages of ChatGPT as a writing assistant tool have been considered: i) providing instant and realistic interactions with learners; ii) designing personalized learning materials based on different proficiency levels; iii) stimulating learners' interests; and iv) providing timely and adaptive feedback and assessments (Barrot, 2023; Fryer *et al*., 2020; Huang *et al*., 2022; Kuhail *et al.*, 2023). While ChatGPT has been shown to be a productive tool for students whose English is not their first language (L1), a few scholars have argued against it because it will either cut down on the practice of good writing demands or hinder creative or critical thinking skills (Liang *et al.*, 2023).

The workload of writing classes for teachers consists of a large amount of assessment, including review, feedback, and grading. In large classes, the task becomes impractical. A solution to this problem may be the use of AI technology such as ChatGPT (Kohnke *et al.*, 2023), which enables the provision of autonomous feedback to students (Artiles Rodríguez *et al.*, 2021; Ranalli, 2018). However, reducing the teacher's workload through automated marking or teaching students to grade themselves poses several challenges, including issues of reliability, consistency, and quality. While educational and linguistic software packages are available for automated assessment and grading, such as Pigaiwang and Coh-Metrix (Zhou & Prado, 2024), the functionality of chatbots allows for easier consultation between the student and the tool and, as such, more effective use of these tools, thus aiding in the management of assessments. In response, we suggest that using chatbots can significantly simplify the task of grading, thereby lessening teachers' workload.

This study explores the use of ChatGPT-3.5 as automated feedback on writing system (Cotos, 2023) and a proofreader for assessing and revising students' essays. In pursuit of objectivity and reliability in our analysis, this study makes use of essays sourced from an official preparatory book for the International English Language Testing System (IELTS), one of the world's most widely accepted English proficiency exams. These essays, previously assessed and selected by IELTS examiners for publication, served as a benchmark for evaluating ChatGPT-3.5's scoring reliability. The choice to use pre-rated essays aims to mitigate the potential subjectivity associated with individual rater judgments. By relying on essays with established scores, we created a more controlled environment to investigate the consistency and reliability of ChatGPT-3.5 as a scoring mechanism as against the standardized criteria set by IELTS, whose descriptors (task achievement, coherence and cohesion, lexical resources, and grammatical range and accuracy) are already embedded in ChatGPT. This methodological approach ensures that the evaluation of ChatGPT-3.5's effectiveness as an Automated Essay Scoring (AES) system is grounded in comparison with authoritative, pre-validated assessments, thus providing a foundation for our analysis. The study manually and qualitatively classifies the strategies used by ChatGPT-3.5 in revising examinees' essays in terms of the four descriptors in the IELTS rubrics, identifying the strengths and weaknesses of ChatGPT-3.5 for revising the essays against different descriptors. To this end, the study investigated the following research questions:

- To what extent do scores on essays differ (or are consistent) between ChatGPT-3.5 and official raters?
- What strategies are used by ChatGPT-3.5 to revise students' IELTS essays?

The results of this study will serve to advance educators' awareness of the pros and cons of ChatGPT as an AES and proofreader. Furthermore, the study will provide directions for future research in the application of ChatGPT to L2 writing. The findings will also shed light on the pedagogical implications of the use of AI tools in future education.

2. LITERATURE REVIEW

2.1. Automatic Scoring and Evaluation of Writing

Traditional classroom-based teaching of writing lacks individual attention to students' learning, resulting in a lack of autonomy and self-initiative, with students passively waiting for teachers to assign essays to be later graded (Yang & Dai, 2015). Automated Essay Scoring (AES) refers to the use of specialized computer programs to evaluate and score the characteristics of compositions based on validity, impartiality, and reliability (Shermis & Burstein, 2003). The development of such systems is the embodiment of the development of machine-assisted language testing (Yang & Dai, 2015), which, technically, is usually based on mathematical formulas and equations for linguistic decodings of the textual features (Zhou & Prado, 2024). In the 1960s, the development of Page Essay Grade (PEG), a program that used multiple regression analysis of measurable text features to build a scoring model based on a corpus of essays previously graded by hand, marked the beginning of AES (He, 2016; Mizumoto & Eguchi, 2023). A large number of AES programs, such as Criterion, My Access, Writing Roadmap, and Pigaiwang, followed suit. These programs were equipped with a number of functions, including a scoring engine, an editing tool that offered grammar and spelling feedback, and a dictionary (He, 2016). As proposed by Bai *et al*. (2022), AES systems are able to lower teachers' workload, especially in situations where learning needs are highly specific.

He (2016) classified the research in AES systems into three types: i) validity of the software; ii) learning outcomes and improvements to learners' writing skills; and iii) use of writing software tools in classroom settings. One of the most recent research projects, carried out by Mizumoto and Eguchi (2023), was representative of the first type. They collected 12,100 Test of English as a Foreign Language (TOEFL) essays and compared the scores given by ChatGPT-3.0 with the benchmark levels, aiming to explore the reliability and accuracy of using ChatGPT-3.0 as an AES along with the linguistic features that influenced the system itself. Their results showed that ChatGPT had a certain level of accuracy and reliability. Moreover, Mizumoto and Eguchi considered several linguistic features at the level of lexis, phraseology, syntax, and cohesion based on previous research that investigated linguistic correlates of human rating scores. They found that the more linguistic features of a text were taken into consideration while evaluating, the more accurate this was reflected in the scoring.

Studies of the second type, namely research in students' learning outcomes, are exemplified by the longitudinal research carried out by Huynh-Cam *et al.* (2023) on students' writing quality. These researchers collected the English writing scores of 82 university students before and after the intervention of an AES tool named Marking Mate in a course of English as a Foreign Language (EFL) writing. A self-report survey was also conducted to explore the attitude of students toward studying with this AES tool. The study found a rise in writing scores using the AES tool as well as favorable opinions from students toward the usefulness of the tool. As regards the third type of AES research, namely its implementation in the classroom, Li (2021) investigated how teachers perceived ESL writing classes supported by Criterion, an automated writing evaluation system. The research found that different teachers tended to take different approaches to implementing the same evaluation tool in classrooms, which in turn reflected observable differences in writing quality. This advocates for the value and significance of teacher agency and cognition in the AES-assisted English teaching classroom.

Having derived from AES, Automated Writing Evaluation (AWE) tools "support the process of writing by providing formative feedback that is typically displayed on an engaging graphic interface" (Cotos, 2023, pp. 347–348). Such tools, considered formative while AES tools are summative (Cotos, 2023), go several steps further in that they employ AI to generate feedback on lexical, semantic, syntactic, and discourse elements on students' writing. AWE tools allow students to draft a text as many times as they wish and be agentive in their selection of feedback, which can vary from global writing skills to language mechanics (Stevenson & Phakiti, 2014).

However, the capabilities offered by AWE tools may not be easily accessed by students. In his L2 writing qualitative study of three students engaging with AWE feedback on their own writing, Zhang (2020) observed that even with a machine designed for the task of analyzing both micro- and macro-level issues, students had their attention drawn almost exclusively to micro-level changes such as spelling and grammar mistakes. In contrast, macro-level changes such as redundancy were attended to only once in Zhang's study, which may reflect a mutual correspondence with higher proficiency levels. Thus, according to Zhang, there is a need for a radical change in how we view L2 revision, which should diverge from an error-reduction activity in favor of more global development.

As regards the field of AIEd, Chiu *et al.* (2023) list several critical areas, among which is the implementation of AI technologies for automating student assessment and predicting their performance. According to their study, priority should be given to the development of a new pedagogical framework centered on AI learning and teaching, particularly in supporting teachers' assessment by "providing automatic marking and predicting students' performance" (p. 9) along with the application of personalized learning. Conditional on this objective is the importance of teachers themselves possessing sufficient knowledge of AI tools and their pedagogical applications. To this end, the authors suggest that future studies should concentrate on the evaluation of the capacity and effectiveness of AI tools applicable to pedagogy.

2.2. Chatbots to Support Writing Feedback and Improvement

Bašić *et al.* (2023) tested ChatGPT-3 as essay-writing assistance for students. The authors compared 18 second-year masters students' essay writing performance with or without employing ChatGPT-3 as a writing assistant tool. Results showed no evidence that using ChatGPT-3 improved the quality of students' essays. This result was consistent with the findings of Fyfe (2022), which tested students' use of GPT-2 and found that students regarded writing independently as easier than writing with GPT-2 as they would be distracted by the texts generated by GPT-2 for the writing task. The study concluded that the use of ChatGPT as an assistance tool could not reduce students' writing time. However, it is worth mentioning that in the study conducted by Bašić *et al*. (2023), the essays were written in Croatian rather than in English. Given that ChatGPT was predominantly fed with English content and thus may have generated higher-quality information in English for students who used it as an essay-writing assistant tool, the results may have been different if English essays had been used instead.

However, some studies support the view that ChatGPT may be beneficial to L2 writing. Han *et al*. (2023) investigated the integration of ChatGPT into L2 writing courses by creating a learning platform called RECIPE (Revising an Essay with ChatGPT) on an Interactive Platform with 213 EFL undergraduate and graduate learners. ChatGPT played the role of a personalized English writing teacher and instructed the students step by step on revising their writing. The results showed that this kind of learning could improve students' writing ability as the steps reminded students of the lecture content and helped them receive a more class-relevant response from ChatGPT. At the end of the course, students reflected that they had a positive experience working with ChatGPT.

Although the effectiveness of ChatGPT-2.0 or 3.0 in grading students' essays and being an assistant to students has been investigated, the quality and nature of improvements to reviewed texts remain to be explored. It is important to examine the characteristics of the suggestions made by chatbots, such as ChatGPT, along with their reliability.

3. METHOD

ChatGPT-3.5, currently a free version, was employed to verify how consistent its suggestions are and to review the feedback it provides. To ensure data consistency, this study made use of one of the most widely used large-scale ESL tests with a writing test component, namely IELTS, the International English Language Testing System, a highly popular exam worldwide as well as in China. The writing section of IELTS contains two types of assignments. The first is a short essay that usually requires candidates to write about 150 words to describe data from a chart or table, and the second is an argumentative essay of about 250 words (for a critical review, see Uysal, 2010).

Bai *et al*. (2022) reviewed 13 studies of the assessing power and accuracy of AES tools in 2021 and found that different studies used different measures. They concluded that the simplest measures consist of focusing on the correlation between human and machine scoring (Pearson correlation coefficient R) and exact accuracy (i.e., the percentage of cases when both human and machine agree on the exact score). Following the same prompt, our study used a quantitative method that references the correlation between human IELTS examiners' grading and ChatGPT-3.5 scores to investigate any differences through experimental comparisons with Pearson's R. Furthermore, a qualitative method was also used focusing on the observation of the strategies used by ChatGPT-3.5 in revising the essays.

3.1. Resources

A total of 23 essays officially scored between band 5.5 and 6.5 were taken from Cambridge IELTS volumes 1 to 17 (see Table 1). The Cambridge IELTS consists of a selection of official examination papers from the University of Cambridge ESOL Examinations with the purpose of preparing candidates for the tests.

Table 1. *Selected essays from Cambridge IELTS Volumes 1-17.*

Publisher	Number	Volume	Year of First Publication	Test No.	Word Count	Score
Cambridge University Press & Cambridge English Language Assessment	1	3	2002	4	317	6
	2	3	2002	Training B	260	6
	3	4	2005	Training A	334	6
	4	5	2006	3	369	6
	5	6	2007	Training A	285	6
	6	8	2011	2	250	5.5
	7	8	2011	4	378	6.5
	8	9	2013	Training A	302	6
	9	10	2015	4	224	5.5
	10	11	2016	1	264	5.5
	11	11	2016	4	276	5.5
	12	12	2017	5	269	6
	13	13	2018	1	313	6.5
	14	13	2018	3	282	6
	15	13	2018	4	276	6
	16	14	2019	3	240	5.5
	17	15	2020	2	350	6
	18	15	2020	4	269	6.5
	19	16	2021	1	284	6
	20	17	2022	1	243	6.5
	21	17	2022	2	280	6.5
	22	17	2022	3	280	6.5
	23	17	2022	4	254	6

The texts were written by candidates and assessed by official IELTS examiners based on four descriptors: task achievement, coherence and cohesion, lexical resources, and grammatical range and accuracy. They are employed as examples or samples to be used by future candidates. These essays correspond to IELTS Writing Task 2, which aims to assess students' ability to provide solutions to problems, clearly presenting and justifying their opinions and supporting them with explicit, logical, and related evidence. Based on IELTS Test Demographic Data (Test Statistics, 2022),[†] which states that the largest proportion (62%) of IELTS scores received by candidates seeking a higher education course was between band 5.5 and 6.5, we selected scores ranging from bands 5.5 to 6.5.

3.2. IELTS Descriptors

As mentioned above**,** the IELTS writing exam consists of four descriptors: task achievement, coherence and cohesiveness, lexical resources, and grammatical range and accuracy. Grammatical range and accuracy are first and foremost a descriptor that emphasizes the accuracy and range of the grammar in the essay. For instance, candidates are expected to use complex structures, appropriate tenses, comparatives, conditionals, and modal verbs in their writing. Second, the lexical resources descriptor highlights the range and accuracy of vocabulary, including synonyms, collocations, and parts of speech. Coherence and cohesiveness, the third descriptor, refers to the flow of texts and how the paragraphs are structured. Finally, task achievement is concerned with how fully the exam question has been answered.

3.3. Instruments

To collect sufficient and useful data to answer the research questions, ChatGPT-3.5 and R were used as the instruments in this study. ChatGPT-3.5 was used to score the 23 essays and revise them to band 7. The suggestions generated by the chatbot were individually compared, and submitted to R for the descriptive data calculation. R is a computational language and a data processing, calculation, and mapping software system that is increasingly being used in research in many disciplines (Crawley, 2012). A further explanation of its use will be included in the next subsections.

3.4. Procedure

The research procedure was divided into two parts. The first part aimed to answer the first research question. After we collected a total of 23 sample essays with scores ranging from bands 5.5 to 6.5, we inserted them into ChatGPT for scoring.

The following steps were replicated with each of the 23 sample essays. First, we gave the chatbot a single prompt, consisting of the request, "Please give a score to this essay in terms of the four descriptors of IELTS writing rubrics", followed by each of the IELTS writing prompts and writing samples. The input is brief as we aimed to imitate how students or teachers, as real-life users, would make use of ChatGPT. For each essay, we input five times, and since, in some cases, the output results of the grade of the same essay were different, the average score of the grades provided in the five rounds was adopted as the grade for later data analysis. We then copied the average band score of each essay given by ChatGPT-3.5, and altogether, there were 23 scores given by ChatGPT-3.5. A t-test between the 23 official scores and the 23 ChatGPT-provided scores was performed through the R language software to ascertain whether there was a significant difference between the gradings of the two groups, namely the samples rated in the resource book and ChatGPT-3.5. In addition, we repeated these steps by inputting "Please give a score to this essay in terms of the four descriptors of IELTS writing rubrics" and the essay again, but this time, we did not provide GPT with the IELTS writing prompt, or the

[†] Text Statistics (2022): https://ielts.org/researchers/our-research/test-statistics#Demographic

required essay topic from the question. A paired t-test was performed again with this group of data and human ratings. This process helped us find whether GPT read and considered the required writing topic for grading.

The second part of the study addressed the second research question by analyzing the revision strategies adopted by ChatGPT. All the selected essays were inserted into ChatGPT-3.5 along with the new prompt "Please revise this IELTS essay to make it achieve a band score of 7 referring to the IELTS writing rubric." Subsequently, we selected 10 of the 23 revised essays through a systematic sampling method by publication year (Table 2), analyzed the revisions suggested by ChatGPT-3.5, coded and classified each revision in terms of the four descriptors from IELTS benchmark (task achievement, coherence and cohesion, lexical resources, and grammatical range and accuracy), with more detailed sub-categories under each descriptor. The analysis and classification achieved by the coding method were implemented through Microsoft Word, particularly the Highlight and Comment functions, to facilitate our collaborative analysis. We first conducted text analyses and coding independently for the ten essays, then discussed until we reached a baseline of 80% intercoder reliability, given that an 80% intercoder reliability is advocated as reliable by scholars such as Miles and Huberman (1994). A more detailed classification of the revisions was then made under each descriptor. Finally, we calculated the strategies most frequently used by ChatGPT-3.5 for further explanation.

As an additional step, despite sampling ten essays for further text analysis, we input all 23 essays into ChatGPT-3.5 for proofreading and revision, after which we input the revised essays again into ChatGPT-3.5 on a separate new page, asking it to assess and grade the revised essays. This helped us explore if the proofreading of ChatGPT-3.5 was effective from the view of ChatGPT-3.5 itself, as we would verify if there was a difference in grades between the original essays and the revised essays.

Table 2. *Selected essays for data analysis.*

Publisher	Number	Series	Year of First Publication	Test Number	Word Count	Score
Cambridge University Press & Cambridge English Language Assessment	1	3	2002	4	317	6
	2	4	2005	Training A	334	6
	3	6	2007	Training A	285	6
	4	8	2011	4	378	6.5
	5	10	2015	4	224	5.5
	6	11	2016	4	276	5.5
	7	15	2020	2	350	6
	8	16	2021	1	284	6
	9	17	2022	2	280	6.5
	10	17	2022	4	254	6

3.5 Data Analysis

We now outline the statistical methods used to analyze the data collected from the 23 IELTS essays, focusing on comparing the scores provided by ChatGPT-3.5 and official IELTS raters, as well as analyzing the revisions made by ChatGPT-3.5 in response to the essays.

The primary method for analyzing the scores given by ChatGPT-3.5 and official IELTS raters was the paired samples t-test, which was used to compare the scores of each essay between the two groups (ChatGPT-3.5 vs. IELTS official raters). The t-test helped us assess whether the differences between the two sets of scores were statistically significant. A paired t-test provides us with the gap between grades of every essay from the two groups rather than an overall distribution of scores of the two groups. This ensures that we focus on each essay in terms of the difference between the two raters, and the t-tests work as an investigator of the scoring gaps of all 23 essays rated by the two raters.

4. RESULTS

4.1. ChatGPT as an Automatic Essay Scoring (AES) System

The numerical data on the grading of the 23 essays are displayed in Table 3, which also shows the mean scores given by ChatGPT-3.5 (with the input of the required topic) and the Cambridge official examiners. Additionally, the table displays the *p*-value of students' t-tests comparing the scores given by ChatGPT-3.5 and those on the official resource book.

Table 3. *Mean scores and t-test (1).*

ChatGPT (input with topic)	Examiners	*p*-value
5.65	6	0.038

The *t*-test checked the degree of difference in the scores given by ChatGPT-3.5 (with input IELTS instructions) and those given by the official IELTS examiners. Results revealed a significant difference between the scores given by the two approaches: ChatGPT-3.5 (with instructions) (*M*=5.65, *SD*=0.93) and IELTS examiners (*M*=6.00, *SD*=0.34), *t*=-1.8606, *p*=.03843.

As mentioned earlier, to check whether ChatGPT-3.5 considered the instructions provided, a new round of testing was performed by inputting *without instructions* for each essay. The results of a *t*-test comparing the scores of ChatGPT-3.5 and those of IELTS examiners are shown in Table 4.

Table 4. *Mean scores and t-test (1).*

ChatGPT (input without instructions)	Examiners	*p*-value
5.75	6	0.077

The results also reveal a difference between the scores given by the two approaches with a 90% confidence interval. However, the difference between the scores provided by ChatGPT-3.5 (*M*=5.75, *SD*=0.86) and those of the examiners (*M*=6.00, *SD*=0.34) was smaller (*t*=-1.4735, *p*=.07772) compared with the difference shown in Table 4.

To test the difference in scores given by ChatGPT-3.5 with and without inputting instructions, a third t-test was performed by R. The two groups of grades are 1) GPT's grading with our input of the instructions from the writing question and 2) GPT's grading without our input of the writing instruction but only the request of grading and the sample essay.

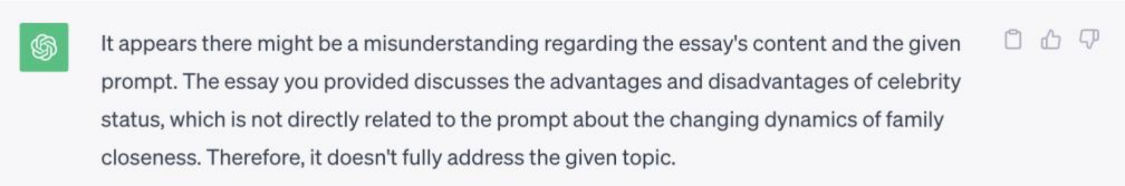
Table 5. *t-test by R.*

Group	Mean Score (*M*)	Standard Deviation (*SD*)	*t*-value	Degree of freedom	*p*-value
ChatGPT-3.5 (with instructions)	5.65	0.93	-1.0058	21	0.326
ChatGPT-3.5 (without instructions)	5.75	0.86			

The results show no significant difference between the scores given by the two approaches. To be specific, there was no evidence of a significant difference between the scores provided by GPT-3.5 *without* instructions (*M*=5.75, *SD*=0.86) and *with* instructions (*M*=5.65, *SD*=0.93) over short-term learning outcomes (*t*=-1.0058, *p*=.326). This indicates that whether inputting the required writing topic or not, GPT will grade the essay similarly, with almost the same scores.

During the interaction with ChatGPT-3.5, two responses were noted. First, even though there was no significant difference between providing and not providing instructions for the essays, this does not imply that ChatGPT-3.5 disregards the instructions. When there was a mismatch between instructions and essay, i.e., when an essay with an instruction differed from a different writing task, ChatGPT-3.5 responded with the identification of the mismatch between instruction and essay, as shown in Figure 1.

Figure 1. *Response to mismatch.*



It appears there might be a misunderstanding regarding the essay's content and the given prompt. The essay you provided discusses the advantages and disadvantages of celebrity status, which is not directly related to the prompt about the changing dynamics of family closeness. Therefore, it doesn't fully address the given topic.

4.2. ChatGPT as Proofreader

During the interactions with ChatGPT-3.5 with requests to revise the IELTS essays, we found that ChatGPT-3.5 tended to re-write the essays rather than simply correcting the problematic areas. That is, ChatGPT changed the structure of sentences, the structure of paragraphs, and even the content of the essays.

Among all the modifications performed by ChatGPT-3.5 in the 10 selected essays, Lexical Resources was the most often revised descriptor (see Table 6).

Table 6. *Modifications to lexical resources.*

Strategy	Occurrences
change a word	179
add an adjective	8
add a phrase	7
correct spelling	3
add a clause	3
total	200

Among the recorded modifications of lexical resources, the most used strategy by ChatGPT-3.5 in revising lexical resources was to "change a word", which was found 179 times in the revisions to the ten essays. Based on further analysis of these modifications, we found that the tool usually uses synonyms to replace original words. In most cases, the revised words were more infrequent or complex, as in changing the expression "some people dead" to "fatalities." However, there were also occasions where the revision did not appear to significantly enhance the difficulty level of the words, as in changing "in my opinion" to "in my view." Examples of word changes are shown in Table 7.

Table 7. *Word modifications by ChatGPT-3.5.*

Original	Revised
in our rather futuristic society	in today's rapidly evolving society
getting more interested	developing a keen interest
in my opinion	in my view
some people dead	fatalities
hometowns	homes and neighborhoods
help	assistance
have been drawn to the attention	has garnered the attention
thus	consequently

The second most revised descriptor was Cohesion and Coherence, with 50 occurrences identified in the revised 10 essays, as displayed in Table 8.

Table 8. *Modifications to Cohesion and Coherence.*

Strategy	Occurrences
add a topic sentence	11
add a connective	11
change a connective	9
restructure	8
add a conclusion	6
clarify reference	3
subject unification	1
add a recap	1
total	50

As shown in Table 8, the most used strategy for revising Cohesion and Coherence in the essays was "add a topic sentence" and "add a connective," with both appearing 11 times in the revisions of the 10 sample essays. "Add a topic sentence" refers to the original essay lacking an overall statement of key ideas at the beginning (or elsewhere) in a paragraph, in response to which ChatGPT-3.5 generated a topic sentence to make up for this deficiency. Examples of topic sentences added by ChatGPT-3.5 are displayed in Table 9.

Table 9. *Examples of topic sentences added by ChatGPT-3.5.*

1. "Raising a child is a profound responsibility that demands love, care, and readiness."
2. "Today, the scenario has undergone a profound transformation."
3. "This essay delves into the reasons behind this growing interest and explores various means by which individuals can research the history of their dwellings."

Regarding the Add Connective strategy, which comes under the Cohesion and Coherence descriptor, Table 10 shows the specific connective words that were added to the 10 selected essays.

Table 10. *Record of added connectives.*

Connective	Occurrences
furthermore	3
however	2
not only; but also	1
additionally	1
moreover	1
in turn	1
secondly	1
conversely	1
total	11

The descriptors of "Task Response" and "Grammar" recorded the same amounts of revisions, with 37 occurrences in total. Three strategies were identified by ChatGPT-3.5 under Task Response, namely "add details", "clarification", and "rationalization," as shown in Table 11. "Add details" refers to ChatGPT-3.5 adding new content to enrich the original text, and the added content is primarily not involved in the original essays. "Clarification" refers to revisions made by ChatGPT to present the original content more clearly. The difference between "Clarification" and "Add details" is that "Clarification" does not add new ideas but only chooses a clearer way to express the author's original idea. In the analysis of the 10 sample essays, "Add details" was found 19 times and "Clarification" 15 times.

The third strategy under "Task Response" is "Rationalization," which refers to providing a rationale for the writer's idea. In some cases, the writer uses strong but unsupported arguments that express ideas powerfully, as in "something must happen" or "it is never possible." In such cases, ChatGPT decreased the (unsupported) strength of the argument, thus enhancing the rationality of the idea, a strategy found on 3 occasions in the 10 essays.

Table 11. *Modifications under Task Response.*

Strategy	Occurrences
add detail	19
clarification	15
rationalization	3
total	37

Table 12 shows the strategies adopted by ChatGPT-3.5 to revise essays in terms of the Grammar descriptor.

Table 12. *Modifications to Grammar.*

Strategy	Occurrences
complication	14
change voice	10
change subject	3
word re-order	7
change sentence structure	4
total	38

The most used strategy was defined as "complication," which refers to grammar being made more complex. To distinguish "complication" from the other strategies under this descriptor, the criterion we chose was the enhancement of grammatical complexity. For example, in one essay, the original sentence "… my view is elaborated further" was revised to "I will elaborate on …." In this case, we classified the revision as "change voice" rather than "complication" since the level of grammatical complexity was not enhanced. An example of "complication" was found in another sentence from a sample essay, in which the original opening was "In this essay, I will try to discuss…" and the revised text was "…, which I will discuss in this essay." Here, the original simple sentence was combined with the previous sentence by transforming it into an attributive clause, which can be considered a step further in grammatical complexity.

Table 13 displays the distribution of the 14 occurrences of complications involving four types of revisions.

Table 13. *Complication.*

Strategy	Occurrences	Year of First Publication	
		Original	Revised
Change independent sentence to attributive clause	5	… and their levels of health and fitness are decreasing.	…, accompanied by a decline in overall health and fitness levels.
Change independent sentence to adverbial clause	5	…, as you do not have to go to a pharmacy …	…, sparing individuals the financial burden …
Change attributive phrase to parentheses	2	The smartphone connected with the internet opens up …	Smartphones, when connected to the internet, open up …
Change independent sentence to parentheses	2	Usually we have to pay around $30 for admissions.	The cost of entry, often around $30, can …

Table 14 displays the scores given by ChatGPT-3.5 to both original and revised essays, revealing a sharp difference between the two groups of scores.

Table 14. *Scores for revised essays given by ChatGPT-3.5.*

Descriptor	Original Essay	Revised Essay
Task Response	5.9	7.7
Coherence & Cohesion	5.7	7.8
Lexical Resources	5.5	7.8
Grammar	5.6	7.9
Overall Band	5.6	7.8

As Table 14 shows, although the grades given by ChatGPT-3.5 differ from the official scores, thus addressing our first research question, based on the scores given to the revised essays, it can be concluded that ChatGPT-3.5 was effective as a proofreader, at least to some extent. However, since the scores for the revised essays were given by ChatGPT-3.5 itself, the next step in the research should be to invite real IELTS examiners to evaluate the revised essays and compare their scores with the original essays.

An interesting phenomenon is that although the instruction to ChatGPT was to "revise the essay to a band 7 score," the tool generally revised all the essays to an average score of 7.8, which did not meet our requirement but exceeded the expected score.

5. DISCUSSION

The research found that an AES system such as ChatGPT-3.5 cannot be regarded as an ideal grader of IELTS exams since scores were generally lower than those given by official raters, with a significant gap in the grading outcomes. Thus, the inaccuracies in ChatGPT-3.5's grading outcomes might, at least for now, mitigate the concern the over total replacement of human raters or teachers (Warschauer *et al*., 2023). Moreover, the findings illustrate the difference in the scores generated by ChatGPT depends on whether or not an instruction was issued along with the essay inputs. The results imply that ChatGPT can read and consider instructions while assessing the essays. However, providing instructions does not make the scoring output more accurate but rather more different from the official scores. This indicates a limitation of ChatGPT-3.5 to take the writing instruction from the IELTS question we provided into appropriate consideration since our provision of this information did not help ChatGPT-3.5 grade more accurately. Moreover, the data showed no significant difference between having instructions input or not. Thus, ChatGPT can only be considered an inconsistent assessor, which makes it unsuited to what Yang and Dai (2015) call machine-assisted language testing. However, since the gap in average scores between ChatGPT and official scores was less than 0.5, ChatGPT can still be used as a supplementary tool in self-study, as in Huynh-Cam *et al*. (2024) and Mizumoto and Eguchi (2023), or a machine-assisted human rating.

As a proofreader, ChatGPT-3.5 showed comprehensive abilities in revising all the descriptors of the IELTS benchmark, as suggested in Stevenson and Phakiti (2014) about AEW tools. This finding is based on a qualitative perspective, with the researchers doing text analysis and manually coding the revisions. However, a much higher average score was given by ChatGPT itself after revising all the sample essays, a positive outcome that is in sharp contrast with the results from Bašić *et al*. (2023), who found GPT-3.0 to be ineffective in assisting students' essay writing. Three possible reasons for this finding can be suggested. The first may be the difference between GPT-3.0, the version used by Bašić *et al*. (2023), and ChatGPT-3.5, which was employed in this research. Second, even though GPT can revise essays, it may not be readily adopted by students, as He's (2016) study. Third, the essays in the study by Bašić *et al*. (2023) were not official exams and thus, unlike IELTS, had no acknowledged rubrics. Thus, the effectiveness of GPT-3.5 as a reliable proofreader can be attributed to the type of essays under consideration as well as its use of popular rubrics such as that used by IELTS and similar exams. This finding aligns with ethical concerns raised by Chiu et al. (2023) regarding textual appropriation and plagiarism in academic writing. ChatGPT's improved performance with well-established, often-studied exams such as IELTS, which focus more on rhetorical strategies than the content itself, highlights potential risks as familiarity with these exams could make it easier for students to rely solely on AI to produce more accurate responses without truly engaging with the content or developing their writing skills.

Regarding ChatGPT-3.5's ability to revise English essays, there was a sharp difference with previous studies that denied the effectiveness of ChatGPT 2 or 3 (Bašić *et al*., 2023; Fyfe, 2022). This suggests two main reasons for the differences between the studies. One of the potential causes may be the gap between theoretical and practical research. Our study explored the effectiveness of ChatGPT from qualitative aspects through our interactions with the tool itself (see Fyfe, 2022; Kuhail *et al*., 2023; Pavlik, 2023) along with our analysis of the output. However, previous studies were mostly of a practical or empirical type, utilizing the tool with students and analyzing their performance (Huynh-Cam *et al*., 2024; Li, 2021; Mizumoto & Eguchi, 2023; Zhang, 2020). This methodological difference could thus be the cause of the inconsistency noted above. Another aspect, as noted above, could be the difference in the version of ChatGPT used, as previous studies investigated earlier versions. Thus we strongly recommend that future research adopt ChatGPT-3.5 (even ChatGPT-4 for the latest technology) in teachers' and students' practices.

The fact that we have experience of ChatGPT places us on an unusual path. For example, we were able to observe how global writing skills and language mechanics (Stevenson & Phakiti, 2014) and common L2 writing mistakes (Liang *et al*., 2023) could both be tackled by ChatGPT. For example, when ChatGPT pointed out issues regarding strong assumptions, we could identify how the way we express ideas might be misinterpreted, including ideas we often do not see as problematic but as enriching our texts. Moreover, we were able to verify what strategies students might have come across when choosing the suggestions given by ChatGPT (Barrot, 2023; Cotos, 2023; Huynh-Cam *et al*., 2024; Stevenson & Phakiti, 2014). Such strategies might inform pedagogical practices that aim to promote students' autonomy (Artiles Rodríguez *et al*., 2021; Barrot, 2023; Baskara, 2023; Chiu *et al*., 2023; Fyfe, 2022; Warschauer *et al*., 2023). They may also be useful in reducing teachers' essay correcting workload (Bai *et al*., 2022; Han *et al*., 2023; Li, 2021; Ranalli, 2018; Yang & Dai, 2015), particularly in the earlier phases of writing (such as drafting).

With regards to the limitations of this study, the coding of the proofreading, though monitored by a teacher, was conducted by two human researchers. Even though this has shown to provide high intercoder reliability, there may be some disputable points regarding categorizing the strategies used in revisions. Second, the sample involved only 23 essays, which may compromise the findings of our quantitative research. Furthermore, as we point out earlier in this paper, there should be another round of human raters, preferably IELTS raters, to assess the output of ChatGPT.

6. CONCLUSION

This study investigated two functions of ChatGPT-3.5 in addressing L2 writing. As a scoring system, ChatGPT-3.5 demonstrates the ability to provide referable scores but lacks the consistency needed to replace human raters entirely. Given the statistically significant gap between AI-generated scores and official rater scores, we should highlight the need for the cautious application of AI in grading high-stakes assessments. In contrast, as a proofreading tool, ChatGPT-3.5 shows significant potential, offering valuable revisions that help students improve lexical resources, cohesion, and overall writing quality. These findings suggest that while ChatGPT-3.5 may not yet be a solution for automated grading, it can effectively support teachers and students in the writing process, particularly in the formative stages. Our research provides a reference to teachers and learners on how reliable and useful ChatGPT is, which in their future teaching and learning will act as a parameter for deciding whether to trust it or not or at least the extent of one's responsibility while using the tool. Future research should explore the integration of advanced AI versions of the tool in practical classroom applications in order to refine their reliability and maximize their pedagogical benefits. By addressing the limitations identified in this study, including the need for larger sample sizes and additional human rater evaluations, researchers can attempt to elucidate the role of AI tools in fostering autonomous learning environments.

Acknowledgments

We would like to express our gratitude to Yangfan Xu, who assisted us with the statistical analysis. This work is supported in part by the UIC Research Grant with No. of UICR0700033–22 at BNU–HKBU United International College, Zhuhai, China.

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.

Contribution of Authors

Xinming Chen: Study conception and design, data collection, analysis and interpretation of results, and manuscript preparation. **Ziqian Zhou:** Study conception and design, data collection, analysis and interpretation of results, and manuscript preparation. **Malila Prado:** Study conception and design, supervision of data collection, analysis and interpretation of results, and manuscript preparation.

Orcid

Xinming Chen https://orcid.org/0009-0008-6860-8633
Ziqian Zhou https://orcid.org/0009-0004-6937-1667
Malila Prado https://orcid.org/0000-0001-6281-6759

REFERENCES

Artiles Rodríguez, J., Guerra Santana, M., Aguiar Perera, V., & Rodríguez Pulido, J. (2021). Agente conversacional virtual: La inteligencia artificial para el aprendizaje autónomo. *Pixel-Bit, Revista de Medios y Educación*, *62*, 107–144. https://doi.org/10.12795/pixelbit.86171

Bai, J.Y.H., Zawacki-Richter, O., Bozkurt, A., Lee, K., Fanguy, M., Cefa Sari, B., & Marin, V.I. (2022, September). Automated Essay Scoring (AES) Systems: Opportunities and challenges for open and distance education. Tenth Pan-Commonwealth Forum on Open Learning. https://doi.org/10.56059/pcf10.8339

Barrot, J.S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, *57*, 100745. https://doi.org/10.1016/j.asw.2023.100745

Bašić, Z., Banovac, A., Kruzic, I., & Jerkovic, I. (2023). *Better by you, better than me: ChatGPT3 as writing assistance in student essays*. https://doi.org/10.48550/ARXIV.2302.04536

Baskara, R. (2023). Integrating ChatGPT into EFL writing instruction: Benefits and challenges. *International Journal of Education and Learning*, *5*(1), 44-55. https://doi.org/10.31763/ijele.v5i1.858

Boa Sorte, P., Farias, M.A. de F., Santos, A. E., Santos, J. do C.A., & Dias, J.S. dos S.R. (2021). Artificial intelligence in academic writing: What is the CPT-3 algorithm? *Revista EntreLinguas*, *7*, e021035.

Chiu, T.K.F., Xia, Q., Zhou, X., Chai, C.S., & Cheng, M. (2023). Systematic literature review of opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *4*, 100118. https://doi.org/10.1016/j.caeai.2022.100118

Cotos, E. (2023). Automated feedback on writing. In O. Kruse, C. Rapp, C.M. Anson, K. Benetos, E. Cotos, A. Devitt, & A. Shibani (Eds.), *Digital writing technologies in higher education* (pp. 347–364). Springer International.

Crawley, M.J. (2012). *The R book*. John Wiley & Sons.

Dergaa, I., Chamari, K., Zmijewski, P., & Ben Saad, H. (2023). From human writing to artificial intelligence generated text: Examining the prospects and potential threats of ChatGPT in academic writing. *Biology of Sport*, *40*(2), 615-622. https://doi.org/10.5114/biolsport.2023.125623

Fryer, L.K., Coniam, D., Carpenter, R., & Lăpușneanu, D. (2020). Bots for language learning now: Current and future directions. *Language Learning & Technology*, *24*(2), 8–22. http://hdl.handle.net/10125/44719

Fyfe, P. (2022). How to cheat on your final paper: Assigning AI for student writing. *AI & Society*, *38*, 1395–1405. https://doi.org/10.1007/s00146-022-01397-z

Han, J., Yoo, H., Kim, Y., Myung, J., Kim, M., Lim, H., Kim, J., Lee, T. Y., Hong, H., Ahn, S.-Y., & Oh, A. (2023). RECIPE: How to integrate ChatGPT into EFL writing education. *Proceedings of the Tenth ACM Conference on Learning @ Scale*, 416–420. https://doi.org/10.1145/3573051.3596200

He, H. (2016). A survey of EFL college learners' perceptions of an on-line writing program. *International Journal of Emerging Technologies in Learning (Online)*, *11*(4), 11-15. https://doi.org/10.3991/ijet.v11i04.5459

Huang, H.-W., Li, Z., & Taylor, L. (2020). The effectiveness of using Grammarly to improve students' writing skills. *Proceedings of the 5th International Conference on Distance Education and Learning*, 122–127. https://doi.org/10.1145/3402569.3402594

Huang, W., Hew, K.F., & Fryer, L.K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, *38*(1), 237-257. https://doi.org/10.1111/jcal.12610

Huynh-Cam, T.-T., Agrawal, S., Bui, T.-T., Nalluri, V., & Chen, L.-S. (2023). Enhancing the English writing skills of in-service students using Marking Mate automated feedback. *Asia Pacific Education Review*, *25*(2), 459–474. https://doi.org/10.1007/s12564-023-09904-7

Kohnke, L., Moorhouse, B.L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal*, *54*(2), 537–550. https://doi.org/10.1177/00336882231162868

Kuhail, M.A., Alturki, N., Alramlawi, S., & Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, *28*(1), 973–1018. https://doi.org/10.1007/s10639-022-11177-3

Lee, S.M. (2023). The effectiveness of machine translation in foreign language education: A systematic review and meta-analysis. *Computer Assisted Language Learning*, *36*(1-2), 103–125. https://doi.org/10.1080/09588221.2021.1901745

Li, Z. (2021). Teachers in automated writing evaluation (AWE) system-supported ESL writing classes: Perception, implementation, and influence. *System*, *99*, 102505. https://doi.org/10.1016/j.system.2021.102505

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, *4*(7), 100779. https://doi.org/10.1016/j.patter.2023.100779

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

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

Open AI. (2022). *Introducing ChatGPT*. https://openai.com/blog/chatgpt

Pavlik, J.V. (2023). Collaborating with ChatGPT: Considering the implications of generative artificial intelligence for journalism and media education. *Journalism & Mass Communication Educator*, *78*(1), 84–93. https://doi.org/10.1177/10776958221149577

Prado, M.C.A., & Huggins, T.J. (2023). Technological approaches to student participation while studying the history of psychology in an EMI institution. In J. Corbett, E.M.Y. Yan, J. Yeoh, & J. Lee (Eds.), *Multilingual Education Yearbook 2023* (pp. 49–69). Springer International. https://doi.org/10.1007/978-3-031-32811-4_4

Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it? *Computer Assisted Language Learning*, *31*(7), 653-674. https://doi.org/10.1080/09588221.2018.1428994

Shermis, M.D., & Burstein, J.C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Routledge.

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing*, *19*, 51–65. https://doi.org/10.1016/j.asw.2013.11.007

Test Statistics. (2022). IELTS. https://ielts.org/researchers/our-research/test-statistics#Demographic

Uysal, H.H. (2010). A critical review of the IELTS writing test. *ELT Journal*, *64*(3), 314–320. https://doi.org/10.1093/elt/ccp026

Warschauer, M., Tseng, W., Yim, S., Webster, T., Jacob, S., Du, Q., & Tate, T. (2023). The affordances and contradictions of AI-generated text for writers of English as a second or foreign language. *Journal of Second Language Writing*, *62*, 101071. https://doi.org/10.1016/j.jslw.2023.101071

Yang, X., & Dai, Y. (2015). An empirical study of college English autonomous writing teaching mode based on www.pigai.org. *Technology Enhanced Foreign Language Education*, *162*(02), 17-23. (Translated from Chinese) https://doi.org/10.3969/j.issn.1001-5795.2015.02.003

Zhang, Z.V. (2020). Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student perceptions and revisions. *Assessing Writing*, *43*, 100439. https://doi.org/10.1016/j.asw.2019.100439

Zhou, Z., & Prado, M. (2024). A corpus-based comparative study of readability of passages in compulsory Chinese English textbooks and exams for middle school students. *Proceedings of the 13th Int. Conf. on Educational and Information Technology*. pp. 279-83. http://doi.org/10.1109/ICEIT61397.2024.10540975

International Journal of Assessment Tools in Education
2025, Vol. 12, No. 1, 78–92

https://doi.org/10.21449/ijate.1371162

journal homepage: https://dergipark.org.tr/en/pub/ijate **_Research Article_**

Examining the cut-off score of the English B1 progression exam according to different standard setting methods

Rümeysa Kaya[1*], Bayram Çetin[2]

[1]Gaziantep University, Gaziantep, Türkiye
[2]Gaziantep University, Faculty of Education, Department of Educational Sciences, Gaziantep, Türkiye

ARTICLE HISTORY

Received: Oct. 5, 2023

Accepted: Apr. 16, 2024

Keywords:
Angoff,
Angoff Y/N,
Nedelsky,
Ebel.

Abstract: In this study, the cut-off scores obtained from the Angoff, Angoff Y/N, Nedelsky and Ebel standard methods were compared with the 50 T score and the current cut-off score in various aspects. Data were collected from 448 students who took Module B1+ English Exit Exam IV and 14 experts. It was seen that while the Nedelsky method gave the lowest cut-off score, Angoff Y/N method gave the highest cut-off score. The z test was used to determine the difference between the percentages of students who were considered successful according to the methods, and all *z* values were found to be significant. The classification of students according to their achievement status was examined with the Cohen's Kappa test. Spearman Brown Rank Differences Correlation coefficient was calculated to examine the relationship between the MPSs of the experts according to the methods, and the highest correlation was found between the Angoff-Ebel methods. Wilcoxon test was used to examine the significance of the difference between the MPS of the methods. Because of the test, the difference between Angoff-Nedelsky, Angoff-Ebel, Angoff Y/N-Nedelsky and Nedelsky-Ebel methods was found to be significant. Among the expert decisions, it was seen that there was a moderate level of agreement in the Angoff, and a high level of agreement in the Ebel and Nedelsky methods. A significant difference was found between the current cut-off score, the 50 T score, and the percentages of students considered successful according to the methods.

1. INTRODUCTION

Measurement tools are used when determining the impact of educational activities on individuals. The measurement tool can be written or oral. Evaluation is made when the measurement result obtained from the measurement tool is compared with a criterion, and a decision is made about the individual's success. Having common goals and criteria in the assessment - evaluation process will ensure standardization in education. This standardization will develop a common language even at the international level. For example, for the English language level, an individual at the B1 level is expected to be able to talk about experiences in daily life, daily events, and topics of interest.

*CONTACT: Rümeysa KAYA ✉ rumeysakayarusen@hotmail.com Gaziantep University, Gaziantep, Türkiye

e-ISSN: 2148-7456

The cut-off score is used to determine the level of language skills an individual possesses according to his/her performance. Before determining the cut-off score, it would be more appropriate to determine and define the performance levels. The cut-off score and performance levels do not have to be determined by the same experts.

The steps and methods used in the cut-off point determination require a certain process called the standard-setting process. There are many methods that can be used in the standard-setting process. The method of application may differ in terms of analysis and interpretation of the obtained data. Jeager (1989) divided these methods into two groups: test-centered methods and student-centered methods. In test-centered methods, experts form the minimum passing score based on their judgments about the test items, while in student-centered methods, they create a cut-off point based on the knowledge and skills of the individuals who answered the test. The test-centered methods that are commonly used are Angoff, Angoff Y/N, Nedelsky, Ebel, and Marking methods, while the student-centered methods mostly utilized are the Boundary Group method and Opposite Groups methods. One of the advantages of these methods is that the cut-off score from the test-centered methods can be obtained without applying the test to the students and that the experts are not affected by the characteristics of the student groups while determining the cut-off score. The test-centered methods used in this study are briefly mentioned below.

1.1. Angoff Method

In this method, developed by William H. Angoff in 1971, experts are asked to predict how many of the 100 students on the pass-fail limit will be able to answer the item correctly for each item in the test. The minimum passing score of that expert is obtained by adding the probability values given by the expert for the items, dividing by the number of items in the test, and multiplying the result with the evaluation score of the test (the highest score that can be obtained from the test). The mean score of the test is obtained by taking the average of the MPS (minimum passing score) found in this way.

1.2. Angoff Y/N Method

In this method developed by Impara and Plake in 1997, experts are asked to give one point for each item in the test if they think an individual on the pass-fail limit will answer that item correctly, and zero points if they think they will answer incorrectly. After adding the points given by the expert for the items and dividing by the number of items in the test, the expert's MPS is obtained by multiplying the result with the evaluation score of the test. The cut-off score of the test is found by taking the mean of the MPS.

1.3. Nedelsky Method

In this method developed by Leo Nedelsky in 1954, experts are asked to estimate the number of options that a pass-fail student can eliminate when reaching the correct answer for each item in the test. The probability of answering the item correctly is found with the formula '1/number of remaining options'. This method can only be applied in tests containing multiple-choice items. The expert's MPS is by adding these probability values calculated based on expert judgments, dividing by the number of items in the test, and multiplying the result by the evaluation score of the test. The cut-off score of the test is obtained by averaging the MPSs.

1.4. Ebel Method

In this method, developed by Ebel in 1972, experts are asked to evaluate each item in the test in two stages. In the first stage, the experts examine the items in two dimensions, namely convenience and difficulty, and place them in a 3x4 table. There are four subgroups in the dimension of relevance: necessary, important, acceptable, and debatable. In the difficulty dimension, there are three subgroups as easy, medium, and difficult. In the second stage, they predict how many of the 100 students on the pass-fail limit will be able to answer the items in each cell correctly. A score is obtained for the cell by multiplying the number of items in the cell with the percentage determined for that cell. The result obtained by adding the cell scores and dividing by the number of items in the test is multiplied by the evaluation score of the test, and the expert's MPS is found. The cut-off score of the test is obtained by averaging the MPSs.

The standard-setting method to be used should be understandable by experts, and the results should be interpretable. Working with a large group of experts will provide a more accurate cut-off score. The expert group should be informed about the method of application, the purpose, and the characteristics of the test.

Studies comparing different standard-setting methods are avaliable in the literature (Berk, 1986; Boduroğlu, 2017; Buckendahl *et al*., 2002; Livingston & Zieky, 1983; Norcini *et al*., 1987; Ömür & Selvi, 2010). In this study, it was aimed to examine how the cut-off points changed according to the four test-centered standard-setting methods, how the obtained cut-off scores affected the percentage of students who were considered successful, how the decisions of the experts about the items changed according to the methods, and the consistency between the expert decisions. In addition, the cut-off score obtained from the standard-setting methods and the 50 T score as a norm-based assessment method, were compared in various aspects. In this study, answers were sought for the following problem statements:

1. What are the cut-off scores for Module B1+ Exit Exam IV using Angoff, Angoff Y/N, Nedelsky, and Ebel standard-setting methods?
2. Is there a significant difference between the percentages of successful students according to the cut-off points obtained from the standard-setting methods used?
3. Is there a consistency between the standard-setting methods used to classify students as successful or unsuccessful according to the methods?
4. Is there a consistency between the standard-setting methods used regarding minimum passing scores among experts?
5. What are the relationships between the actual difficulty values of the items, the estimated item response probabilities given by the experts using the Angoff method, and the estimated item response probabilities given by the experts using the Ebel method?
6. What is the level of agreement between the experts' decisions on the items according to the standard-setting methods used?
7. Do the percentages of students who score above the current cut-off score of Module B1+ Exit Exam IV and the cut-off scores obtained by the standard-setting methods used in the research differ?
8. What is the cut-off score obtained according to the 50 T score, the number of students accepted as successful according to this score, and the percentage of students, and is there a significant difference between the 50 T score and the percentage of students who are considered successful according to the cut-off scores obtained from the standard-setting methods used in this study?
9. Is there harmony in classifying students as successful or unsuccessful according to the standard-setting methods used in this study with a T score of 50?

2. METHOD

This study aimed to obtain cut-off points from different standard-setting methods and examine the obtained cut-off scores in different centers. In this context, it is a descriptive and relational study. Excel, JASP 0.16.1.0, and SPSS Statistics v23 x64 programs were used during the tests and analyses. The significance value was accepted as .05 in all analyses in the study.

2.1. Study Group

In this study, data were collected from two different groups. The 1st group consisted of 448 students who answered the Module B1+ Exit Exam IV. The second group was 14 lecturers working at the School of Foreign Languages and filling out the standard-setting methods forms. While determining the number of experts, previous studies on this subject were taken into account (Hurtz & Hertz, 1999).

2.2. Data Collection Tools

This study used Module B1+Exit Exam IV, which was held at the end of the 2021-2022 academic year of the School of Foreign Languages of a state university, was used. There are 62 items in the exam, which consists of four sub-sections: Listening, use of English, vocabulary and reading. Student scores were calculated in accordance with the exam guidelines. As a result of the analyses made on these scores, it was seen that the difficulty and distinctiveness of the test were moderate (KR20=0.69, test difficulty ($\bar{P}$)=0.51). Student responses showed a normal distribution (kurtosis=0.02, skewness=0.20).

While obtaining data from the experts, expert evaluation forms were given to the experts along with the exam questions. Experts filled out the forms following the instructions. In this study, pass-fail students were identified as individuals with B1-level characteristics made by the Common European Framework of Reference for Languages.

B1 Level (Intermediate-Independent User):

- He/she can convey the events and experiences he/she has lived; can talk about their dreams, hopes, and wishes, and briefly explain their views and plans with their reasons.
- Can handle most situations encountered when traveling, where the language is spoken. Can understand the main lines of written expressions based on familiar topics in daily life.
- Can express himself/herself in line with his/her interests or on the subjects he/she knows through simple texts with links between ideas.

2.3. Analysis of Data

For the first sub-problem of the study, expert evaluation forms prepared in accordance with the application of the methods used in the study were given to the experts. While 14 expert forms were used for Angoff, Angoff Y/N, and Nedelsky methods, the forms belonging to 4 experts were deemed invalid in the Ebel method and 10 expert forms were used.

In the solution of the second sub-problem of the study, the student scores were classified as successful or unsuccessful according to the cut-off points obtained from the methods. The number and percentage of successful students were determined and the significance of the difference between these percentages was examined with the *z*-test. The *z*-test is used to check the significance of the difference between two dependent percentages in sample numbers larger than 30.

Cohen's Kappa test was used to determine the compatibility between the classification of students' achievement status according to the methods in the solution of the third sub-problem of the study. In order to make the scores suitable for the test, the cut-off score of the method and above were converted to 1 and other scores to 0. The fit rating scale suggested by Landis and Koch (1977) was used to interpret the results. This scale is as follows:

0.00 - 0.20 = slight
0.21 - 0.40 = fair
0.41 - 0.60 = moderate
0.61 - 0.80 = substantial
0.81 - 1.00 = almost perfect

In the solution of the 4th sub-problem of the study, the relationship between the expert MPS was examined by calculating the Spearman-Brown Rank Differences Correlation Coefficient. The Spearman-Brown Rank Correlation Coefficient is a statistical method used to examine the relationship between variables when the data is less than 30. The following rating scale was used to interpret this correlation coefficient (İlhan, 2022).

$r < 0.20$ = no relationship
$0.20 < r < 0.39$ = weak relationship
$0.40 < r < 0.59$ = moderate correlation
$0.60 < r < 0.79$ = high level of association
$0.80 < r < 1.00$ = very high correlation

In the continuation of the solution, the Friedman chi-square test was performed to examine the significance of the difference between the mean of the MPSs obtained from the methods. Friedman chi-square test is a non-parametric test used to check whether the mean scores of two or more groups differ significantly Wilcoxon Signed Ranks test was used to see the difference between the mean of MPS and which methods were significant.

In the solution of the fifth sub-problem of the research, the average of the percentage estimates of the experts for answering the items based on the Angoff and Ebel method (considering the percentages obtained in the Ebel method on an item basis). With these averages, descriptive statistics based on students' exam results were found. Pearson Product Moments Correlation Coefficient was calculated since the data showed normal distribution.

In the solution of the sixth sub-problem of the study, the expert evaluation forms were transferred to Excel according to the methods filled by the experts. Kendall's W fit coefficient was calculated by considering the agreement between the expert decisions, Kendall's W fit coefficient in Angoff method, Cochran Q test in Angoff Y/N method, Intraclass Correlation Coefficient for Nedelsky method and the percentage values given by the experts about cells in Ebel method on an item basis. Kendall's W concordance coefficient is used when the number of raters is more than two and a single cohesion coefficient is desired to be obtained from the data. The scale used in the interpretation of this coefficient is given below (Rovai *et al*., 2014):

0.00 – 0.20 = very weak effect
0.21 – 0.40 = weak effect
0.41 – 0.60 = medium effect
0.61 – 0.80 = strong effect
0.81 – 1.00 = very strong effect

Since the Cochran Q test examines the agreement between expert evaluations in two categories, such as 1-0 or positive-negative, this test was preferred in the Angoff Y/N method.

For the solution of the seventh sub-problem of the study, the passing grade of the B1 level of the School of Foreign Languages, where the study was carried out, was 60, and it was assumed in this study that the passing grade was created only according to Module Exit Exam IV. The number and percentages of students who got the current cut-off score and above of the methods and the exam were found. Then, the significance of the difference between these percentages was examined with the formula of the *z*-test.

In the solution of the eighth sub-problem of the study, the scores obtained by the students from the exam were converted into T scores. In this study, 50 T score was determined as a criterion as a norm-based assessment. The number and percentage of students considered successful according to the 50 T score were found. The significance of the difference between the percentages of students who were considered successful according to the methods and those who were considered successful according to the 50 T score was examined by performing the *z*-test.

In the solution of the ninth sub-problem of the study, the scores of the students who were considered successful according to the 50 T score and the cut-off point of the methods were converted to 1 and the other scores to 0. Then, Cohen's Kappa test was performed on these data.

3. RESULTS

In the solution of the first sub-problem of the study, MPSs of the methods were calculated based on the standard-setting methods forms filled by the experts. Since four expert forms were deemed invalid in the Ebel method, the MPS of four experts could not be calculated for this method. In Table 1, the MPSs of the experts according to the methods are given:

Table 1. *MPS of experts by methods*.

Experts	Minimum Passing Score (MGP) for Angoff Method	Minimum Passing Score (MGP) for Angoff Y/N Method	Minimum Passing Score (MGP) for Nedelsky Method	Minimum Passing Score (MGP) for Ebel Method
Expert 1	73.71	72.58	64.06	73.15
Expert 2	94.48	77.42	33.00	83.63
Expert 3	49.76	45.16	40.94	48.65
Expert 4	56.05	64.52	51.18	52.10
Expert 5	63.23	82.26	65.11	-
Expert 6	72.10	62.90	53.23	70.56
Expert 7	72.02	64.52	64.19	70.48
Expert 8	67.82	46.77	39.19	52.58
Expert 9	58.06	72.58	39.02	41.53
Expert 10	58.39	62.90	34.66	50.48
Expert 11	57.34	74.19	39.29	41.53
Expert 12	39.81	58.06	42.03	-
Expert 13	56.69	56.45	37.66	-
Expert 14	53.95	61.29	57.65	-

As can be seen in Table 1, since the MPPs of the Angoff method contain extreme values, the cut-off scores of the methods were obtained by taking the mean of the corrected (pruned) mean in this method and the MPS of the other methods, since the MPS of the other methods did not contain extreme values. The cut-off points calculated according to the MPSs obtained from the experts are given in Table 2.

Table 2. *Cut-off scores of Angoff, Angoff Y/N, Nedelsky, and Ebel methods.*

Methods	Angoff	Angoff Y/N	Nedelsky	Ebel
Cut-off Score by Method	61.59	64.40	47.23	58.47

When Table 2 is examined, the highest cut-off score in this study was obtained from the Angoff Y/N (64.40) method, while the lowest cut-off score was obtained with the Nedelsky method (47.23). It was observed that there was a difference of 14.36 points between the highest cut-off score and the lowest cut-off score. This may be due to the way the methods are applied. It is possible that the Nedelsky method, which involves focusing on all options together with the item root, may have been overlooked in this instance. This may have resulted in the clues provided by the correct option being misinterpreted, leading experts to consider the items in question to be more challenging than they actually were. In the Angoff Y/N method, on the other hand, it may be due to the decrease in the judgment options related to the items by evaluating the items according to only two value judgments (1-0). The cutoff scores of the Angoff and Ebel methods are close to each other because both methods contain an estimate of the percentage of students at the minimum proficiency level. The fact that the lowest cut-off score belongs to the Nedelsky method also coincides with the results of the studies conducted by Tanrıverdi (2006), Taşdemir (2009), and Yıldırım Kan (2019).

For the second sub-problem of the study, the cut-off points obtained from the methods and the number and percentages of students who scored above were calculated. Then, a z-test was performed to test the significance of the difference between these percentages. Table 3 gives the percentage of students who are considered successful according to the methods and the results of the z-test.

Table 3. *The number of students deemed successful according to the methods, their percentage, and z-test results.*

Methods	*N*	%	*z*
Angoff	79	17.63	5.1*
Angoff Y/N	53	11.83	
Angoff	79	17.63	13.68*
Nedelsky	26	59.38	
Angoff	79	17.63	4.58*
Ebel	100	22.32	
Angoff Y/N	53	11.83	14.60*
Nedesky	266	59.38	
Angoff Y/N	53	11.83	6.86*
Ebel	100	22.32	
Nedelsky	266	59.38	12.89*
Ebel	100	22.32	

*p<.05

The value required for a significant difference at the .05 level in the *z*-test is 1.96. All *z*-values found as a result of comparing the methods' percentages in pairs were greater than 1.96. It was seen that the difference between the percentages of students who were considered successful according to the methods was significant. This result was obtained because the difference in cut-off scores affects the percentage of students who are considered successful according to the methods.

In the solution of the third sub-problem of the study, Cohen's Kappa test was performed to determine the fit in terms of classifying the students according to their success status according to the methods and the degree of this fit, if any, and the values found were interpreted. The results of the Cohen's Kappa test are given in Table 4.

Table 4. *Cohen's Kappa test results.*

Methods	Kappa coefficient (*k*)	Compliance Level
Angoff - Angoff Y/N	0.77	substantial fit
Angoff – Nedelsky	0.26	fair fit
Angoff Y/N- Nedelsky	0.17	slight fit
Angoff – Ebel	0.85	Almost perfect fit
Angoff Y/N – Ebel	0.64	Substantial fit
Nedelsky – Ebel	0.33	fair fit

As seen in Table 4, all *k* values are positive, which indicates that the methods were correctly understood by the experts and that the expert's decisions about the item were consistent. Considering the level of fit, the best fit was between Angoff and Ebel methods (Kappa=0.85, Kappa>0.75, almost perfect fit), and the lowest fit between Angoff Y/N and Nedelsky methods (Kappa=0.17, Kappa<0.20, slight fit). As the cut-off points of the methods get closer to each other, the fit value between them also increases. The results found between Angoff and Ebel also coincide with the results of previous studies. (Demir, 2014; Gündeğer, 2012).

In the solution of the fourth sub-problem of the study, the Spearman-Brown Rank Correlation Coefficient was calculated to examine the relationship between MPSs obtained from experts according to the methods. The Friedman Chi-Square test was used to check the existence of agreement between all methods in terms of the mean of MPSs. The Spearman-Brown Rank Differences Correlation Coefficient results are given in Table 5.

Table 5. *Spearman Brown rank differences correlation coefficients between MPSs.*

		Angoff	Angoff Y/N	Nedelsky	Ebel
Angoff	N	-			
	R	-			
	P	-			
Angoff Y/N	N	14	-		
	R	0.51	-		
	P	0.06	-		
Nedelsky	N	14	14	-	
	R	0.03	0.17	-	
	P	0.92	0.55	-	
Ebel	N	10	10	10	-
	R	0.86*	0.16	0.24	-
	P	0.00	0.67	0.51	-

A statistically significant relationship was found only between the experts' MPSs for the Angoff and Ebel methods. (*p*<.05). In addition, the correlation value between these two methods was positive and very high (*r*>.80, *p*<.05). As a result of the Friedman Chi-Square Test, it was observed that at least one of the MGP averages differed significantly from the others ($\chi^2=13.29$, *p*<.05). Wilcoxon Signed Ranks Test was used to check which mean of MGP of the methods was significant. The results of the Wilcoxon Signed Ranks Test are given in Table 6.

Table 6. *Wilcoxon signed-row test results.*

Methods	*N*	Z	*p*
Angoff Angoff Y/N	14	0.41	.68
Angoff Nedelsky	14	2.92*	.004
Angoff Ebel	10	2.81*	.005
Angoff Y/N Nedelsky	14	3.30*	.001
Angoff Y/N Ebel	10	0.66	.507
Nedelsky Ebel	10	2.80*	.005

**p*<.05

As can be seen in Table 6, the methods with a significant difference in terms of MPS averages are Angoff - Nedelsky, Angoff - Ebel, Angoff Y/N - Nedelsky and Nedelsky - Ebel methods. While there is a very high correlation between the MPSs of the Angoff and Ebel methods, the significant difference between the MPS averages indicates that the MPSs of the experts according to the two methods are in the same direction, but the MPS averages of one of the methods differ due to the lower MPSs of the other methods. While there is no relationship between the MPSs of Angoff Y/N – Ebel and Nedelsky - Ebel methods, the lack of a significant difference between the MPS averages shows that the experts' perception of ease-difficulty regarding the whole test for the two methods has changed. However, when the averages of these MPSs are averaged, the results are close to each other.

In the solution of the fifth sub-problem of the study, the difficulty levels of the items were calculated based on the answers of the students who participated in the exam. Then, the average of the item answer probability estimates made by the experts using the Angoff and Ebel methods were taken. Thus, the average response percentage of each item was found according to both methods. In Table 7, descriptive statistics based on real item difficulty with Angoff and Ebel methods are given:

Table 7. *Descriptive statistics for item difficulty and actual item difficulty based on Angoff and Ebel methods.*

	Estimated Item Difficulty Based on Angoff Method	Estimated Item Difficulty Based on Ebel Method	Real Item Difficulties
N	62	62	62
Minimum	0.54	0.51	0.13
Maksimum	0.72	0.89	0.89
Average	0.62	0.58	0.51
Standard deviation	0.04	0.04	0.20
Distortion	0.14	0.09	0.11
Kurtosis	0.33	0.56	0.72

When Table 7 is examined, it is seen that the difficulty levels estimated according to the Ebel and Angoff judgment method are easier than they actually are. Since the data showed a normal distribution, the relationship between the item difficulties according to the three conditions was examined by calculating the Pearson Product Moments Correlation Coefficient. The results are given in Table 8.

Table 8. *Correlation between Angoff and Ebel methods estimated item difficulties and actual item difficulties.*

		Real Item Difficulty	Angoff-Based Item Difficulty	Item Difficulty Based on Ebel
Real Item Difficulty	*r*	-		
	p	-		
Angoff-Based Item Difficulty	*r*	0.52^{*}	-	
	p	<.001	-	
Item Difficulty Based on Ebel	*r*	0.36^{*}	0.67^{*}	-
	p	0.004	<0.001	-

$^{*}p$<.05

It was observed that there was a positive and moderately significant correlation between the experts' average of the estimated item difficulties based on the Angoff method and the actual item difficulties (*r=0.52, p<.05, N=62*). This result coincides with the result of Çetin (2011)'s study. It was observed that there was a positive and weakly significant correlation between the experts' estimated item difficulties based on the Ebel method and the actual item difficulties (*r=0.36, p<.05, N=62*). It was observed that there was a positive and highly significant correlation between the experts' mean estimated item difficulties based on the Angoff and Ebel methods (*r=0.67, p<.05, N=62*).

The significant relationship between the average of the estimates made by the experts about the item difficulties according to the Angoff and Ebel method and the actual item difficulties indicate that the predictions made by the experts using the methods are valid. The weak correlation between the estimated item difficulty averages based on the Ebel method and the actual item difficulties may be because the percentage values given for cells in the Ebel method are considered on an item basis.

In the solution of the sixth sub-problem of the study, the harmony between the expert decisions was examined. Kendall's W coefficient of agreement was found to be .561 for the agreement between the estimates of 14 experts for 62 items in the Angoff method ($\chi^2=451.943$, *sd=13, p<.05)*. This value shows that the expert decisions are moderately compatible in the Angoff method. This harmony also coincides with the results of Kılıç (2013) study.

Cochran Q coefficient of agreement was checked for the consistency between the decisions made by 14 experts for 62 items in the Angoff Y/N method, and it was seen that the expert decisions were compatible (*Q=43.356, p<.05).* In the Nedelsky method, it is seen that the In-Class (Cluster) correlation coefficient of agreement between the decisions made by 14 experts for 62 items is 0.70. This value shows that the expert decisions are highly compatible with the Nedelsky method.

The Kendall W agreement coefficient for the agreement between the estimates of 10 experts for 62 items in the Ebel method was found to be .691 ($\chi^2=385.220$, *sd=9, p<.05*). This value shows that the expert decisions are strongly compatible in the Ebel method. The increase in the number of experts and the number of items in the test makes it difficult to achieve high agreement among experts.

In the solution of the seventh sub-problem of the study, 21.21% (95 students) of the students who took the exam according to the current cut-off score were successful. The significance of the difference between the current cut-off score and the percentages of students who were considered successful according to the cut-off scores obtained from the methods was examined with the *z*-test. The *z* test results are given in Table 9.

Table 9. *z-test results for the percentage of successful students according to the methods and current cut-off score.*

	N	%	*z*
Angoff Method	79	17.63	4*
Current Passing Score	95	21.21	
Angoff Y/N Method	53	11.83	6.48*
Current Passing Score	95	21.21	
Nedelsky Method	266	59.38	13.08*
Current Passing Score	95	21.21	
Ebel Method	100	22.32	2.23*
Current Passing Score	95	21.21	

*$p<.05$

When the current cut-off points and the methods were compared one by one in terms of the percentage of students who were considered successful, it was seen that all *z* values were significant. This shows that the current cut-off score and the cut-off score of the methods differ significantly from each other.

In the solution of the eighth sub-problem of the study, student scores were converted to T scores. In this evaluation, 50 T points were taken as a criterion. According to the 50 T score, 47.32% of the students (212 students) were successful. The significance of the difference between the 50 T score in terms of the percentage of students considered successful and those considered successful according to the cut-off points obtained from the methods was examined with the *z*-test. The *z*-test results are given in Table 10.

Table 10. *z-test results for the percentage of students deemed successful according to methods and 50 T-scores.*

	N	%	*z*
Angoff Method	79	17.63	11.53*
50 T Points	212	47.32	
Angoff Y/N Method	53	11.83	12.61*
50 T Points	212	47.32	
Nedelsky Method	266	59.38	7.35*
50 T Points	212	47.32	
Ebel Method	100	22.32	10.58*
50 T Points	212	47.32	

**p*<.05

Looking at Table 10, it was seen that all *z* values were significant. This indicates that the cut-off scores of standard-setting methods and the 50 T score, which is an assessment method based on norms, differ significantly. This result is similar to that of the study of Çukadar (2013) and Şahin (2019).

For the solution of the ninth sub-problem of the study, 50 T points and student scores considered successful according to the cut-off point of the methods were converted as 1, and student scores considered unsuccessful were converted to 0. Then, Cohen's Kappa Test was performed on these data. Statistical information about the test result is given in Table 11.

Table 11. *The results of the Cohen's Kappa test were performed with a T score of 50 and the level of agreement between the methods.*

Methods	Kappa coefficient (k)	Compliance Level
Angoff – 50 T	0.39	Fair fit
Angoff Y/N -50 T	0.26	Fair fit
Nedelsky – 50 T	0.76	Substantial fit
Ebel-50 T	0.49	Moderate fit

It was seen that Nedelsky method (k = 0.76, substantial fit) gave the best fit with a T score of 50, and Angoff Y/N method (k = 0.26, fair fit) gave the lowest fit, in terms of classifying students according to their achievement status. This is because the T score of 50 and the cut-off score of the Nedelsky method are close to each other.

4. DISCUSSION and CONCLUSION

In this study, the cut-off score of Gaziantep University foreign language B1 level exam was calculated using Angoff, Angoff Y/N, Nedelsky and Ebel standard-setting methods. These scores were then compared, in various aspects, with the existing cut-off score and the 50 T score, which is one of the norm-based evaluation methods. The results obtained and discussions based on these results are given below.

As evidenced by the findings, the cut-off scores of the methods in question exhibited notable discrepancies. These discrepancies can be attributed to the fact that the specific areas of focus for experts may vary depending on the method being employed. The result of the lowest cut-off point in this study belongs to the Nedelsky method, which is in line with the results of the previous studies, except for the study of Taşdelen (2009). This may be because the experts perceive the items as more difficult than they are since the Nedelsky method examines all the options one by one. The result of the Angoff Y/N method, which acts with only two judgments, has a very low cut-off score. This result is consistent with the results of the previous study. The Angoff Y/N method's ability to make values over only two sources from the fact that its results differ significantly from other methods. The cut-off score of the Ebel method is lower than the cut-off scores of the Angoff and Angoff Y/N methods. It has been shown that the more complex the understanding and application of the standard-setting method is, the lower the cut-off score is.

The results indicate that the percentages of students who are considered successful according to the cut-off scores differ significantly for all methods, and this finding showcases that even minor differences between the cut-off scores significantly impact the exam results. It has also been observed that there is an inverse proportion between the cut-off score and the percentage of students considered successful. In cases where the cut-off points of the methods were close to each other, it was seen that the results of the classification of the students according to their success were close to each other. The Nedelsky method gave lower coefficients in terms of compatibility with other methods because the cut-off score was much lower than the other cut-off scores. The perfect harmony between Angoff and Ebel methods stems from the common points in the way the methods are applied. The large difference between the percentages of students who are considered successful according to the standard-setting methods reveals the importance of making decisions by using more than one method in creating cut-off points for the exams.

The moderate relationship between the Angoff method and the Angoff Y/N method in terms of MPSs shows that the experts' perception of the difficulty of the exam is similar according to these two methods. The fact that these two methods do not differ significantly in terms of MPS averages shows that the MPS averages of the methods are close to each other. The fact that there is no relationship between Angoff-Nedelsky, Angoff Y/N- Nedelsky and Nedelsky in terms of MPSs and that there is a significant difference between the MPS averages of these methods shows that experts' ideas about the structure of the exam have changed while working with the Nedelsky method. The very high level of correlation between the MPS of the Angoff method and the MPS of the Ebel method may be because both methods involve estimating over 100 students at the pass-fail limit. Although there was a high level of correlation between the MPSs of these two methods, the differentiation in terms of MPS averages indicates that the experts perceived the items more easily in one of the methods. It was observed that experts made similar decisions using the Angoff method.

Although there is no relationship between the MPSs of the Ebel method and the MPS of Angoff Y/N and Nedelsky methods, the lack of difference between MPS averages indicates that the perceptions of the experts about the difficulty of the items in the test have changed. However, MPS averages of the methods are close to each other. Since there is a high level of agreement between the MPS of the Angoff and Ebel methods, only one of the methods can be used when the aim is to save time in determining the cut-off point.

The weak correlation between estimated item difficulties based on the Ebel method and actual item difficulties indicates that it is not a correct practice to consider the percentage values given by the experts for cells in the Ebel method on an item basis. A different study could examine whether the number of items in the test and the structure of the test have an impact on the relationship between actual item difficulties and experts' method-based item difficulty estimates. Angoff method is more appropriate to implement when estimating item difficulty in the test development process.

In order to see why the agreement between experts was at a medium level in the Angoff method, the expert forms were examined, and it was seen that one of the experts gave all probability values at a very high level. In cases where two judgments are used, such as the Angoff Y/N method, it has been found that it is more appropriate to check whether there is harmony between expert decisions. In cases where the Nedelsky method is used, the high level of agreement between expert decisions shows that the more detailed the experts examine the items, the greater the agreement between them. The higher agreement between experts in the Ebel method than in the Angoff method may be because fewer experts are employed in the Ebel method. The effect of the number of items on the harmony between experts can be examined by looking at the harmony between the experts' judgments in the first and last half of the test.

The divergence between norm-based assessment and standard-setting methods results is observed due to the fact that test-centered methods are not affected by student characteristics. Student-centered methods and norm-based assessment results are likely to yield similar results. As seen in the study, if a cut-off score is created without using the standard-setting method in exams that aim to recognize and place students, judging students' level of language skills, the results based on this cut-off score do not make accurate decisions about the students. In exams with high student participation, creating a cut-off score using at least one standard-setting method with a broad group of experts will increase the reliability and validity of the exam criteria.

In light of all these findings, it is seen that it is important to use various standard-setting methods together and keep the expert group-wide when determining the cut-off score in exams where absolute evaluation becomes important. In addition, the test items should be reviewed by looking at which items the expert judgments differ significantly on.

Acknowledgments

This study was conducted by Rümeysa KAYA at Gaziantep University, Institute of Educational Sciences, as a Master's Thesis. It is summarized from the master's thesis that was conducted under the supervision of Bayram ÇETİN.

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**: Gaziantep University, Social and Human Sciences Ethics Committee, 12.10.2022-246368.

Contribution of Authors

The authors contributed equally to all the stages of the study.

Orcid

Rümeysa Kaya https://orcid.org/0000-0003-3212-3032
Bayram Çetin https://orcid.org/0000-0001-5321-8028

REFERENCES

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), Educational measurement (2nd ed.). *American Council on Education.*

Berk, R.A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, *56*(1), 137-172. https://doi.org/10.2307/1170289

Boduroğlu, E. (2017). *Yükseköğretime geçiş sınavının sınıflama tutarlılığının farklı yöntemlerle elde edilen kesme puanlarına göre incelenmesi* [*The study of classification consistency of transition to higher education examination according to the cut-off scores obtained from different methods*] [Master's dissertation, Mersin University]. Higher Education Institution National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Buckhendal, W.C., Smith, W.R., Impara, C.J., & Plake, S.B. (2002). A comprasion of Angoff and Bookmark standard setting methods. *Journal of Educational Measurement, 39*(3), 253-263. https://doi.org/10.1111/j.1745-3984.2002.tb01177.x

Çetin, S. (2011*). İşaretleme ve angoff standart belirleme yöntemlerinin karşılaştırılması* [*Comparison of bookmark and angoff standard setting methods*] [Doctoral dissertation, Hacettepe University]. Higher Education Institution National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Çukadar, İ. (2013). *Norm ve ölçüt dayanaklı değerlendirmelerin karşılaştırılmasına ilişkin bir çalışma* [*A study upon comparison of norm and criterion referenced assessment*] [Master's dissertation, Hacettepe University]. Higher Education Institution National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Demir, O. (2014). *Angoff, nedelsky ve ebel standart belirleme yöntemleri ile belirlenen kesme puanlarının karşılaştırılması* [*A comparison of cutting points determined by angoff, nedelsky and ebel standard setting methods*] [Master's dissertation, Abant İzzet Baysal University]. Higher Education Institution National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Ebel, R.L. (1972). *Essentials of educational measurement*. Englewood Cliffs NJ: Prentice-Hall.

Gündeğer, C. (2012). *Angoff, Yes/No ve Ebel Standart Belirleme Yöntemlerinin karşılaştırılması* [*A comparison of Angoff, Yes/No and Ebel Standard Setting Methods*] [Master's dissertation, Hacettepe University]. Higher Education Institution National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Hambleton, R.K., Jaeger, R.M., Plake, B.S., & Mills, C. (2000). Setting Performance Standards on Complex Educational Assessments. *Applied Psychological Measurement*, *24* (4), 355–366. https://doi.org/10.1177/01466210022031804

Hambleton, R.K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89–116). Lawrence Erlbaum Associates Publishers.

Hurtz, G.M., & Hertz, N.R. (1999). How many raters should be used for establishing cutoff scores with the Angoff method? a Generalizability Theory Study. *Educational and Psychological Measurement, 59* (6), 885-897. https://doi.org/10.1177/00131649921970233

Impara, J.C., & Plake, B.S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*(4), 353-366. https://doi.org/10.1111/j.1745-3984.1997.tb00523.x

İlhan, M. (2022). Korelasyon [*Correlation*]. In Çetin B. (Ed.), *Eğitimde ölçme ve değerlendirme* [*Measurement and evaluation in education*]. (2nd ed., pp. 23-43). Anı Publishing.

Jaeger, R.M. (1989). Certification of student competence. In R.L. Linn (Ed.), *Educational measurement* (3rd ed. pp 485-514). Macmillan Publishing Co, Inc; American Council on Education.

Kılıç, A. (2018). *Angoff, yes/no ve sınır grup yöntemlerine göre kesme puanlarının karşılaştırılması [Comparison of cutting points by Angoff, yes / no and borderline group methods]* [Master's dissertation, Abant İzzet Baysal University]. Higher Education Institution National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Kaya, R. (2023). *İngilizce B1 seviye atlama sınavının kesme puanının farklı standart belirleme yöntemlerine göre incelenmesi* [*Examination of the cutting score of the English B1 leveling exam according to different standard determination methods*] [Master's dissertation, Gaziantep University]. Higher Education Institution National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Landis, J.R., & Koch, G.G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, *33*(2), 363-374. https://doi.org/10.2307/2529786

Livingston, S.A., & Zieky, M.J. (1983). A comparative study of standard-setting methods. *ETS Research Report Series*, *1983*(2), i-48. https://doi.org/10.1002/j.2330-8516.1983.tb00038.x

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, *14*(1), 3-19. https://doi.org/10.1177/001316445401400101

Norcini, J.J. (2003). Setting standards on educational tests. *Medical education, 37*(5), 464-469. https://doi.org/10.1046/j.1365-2923.2003.01495.x

Norcini, J.J., Lipner, R.S., Langdon, L.O., & Strecker, C.A. (1987). A Comparison of Three Variations on a Standard-Setting Method. *Journal of Educational Measurement*, *24*(1), 56–64. https://doi.org/10.1111/j.1745-3984.1987.tb00261.x

Ömür, S., & Selvi, H. (2010). Angoff, Ebel ve Nedelsky yöntemleriyle belirlenen kesme puanlarının sınıflama tutarlılıklarının karşılaştırılması. *Journal of Measurement and Evaluation in Education and Psychology*, *1*(2), 109-113. https://dergipark.org.tr/tr/download/article-file/65991

Rovai, A., Baker, J., & Ponton, M. (2014). *Social science research design and statistics: a practitioner's guide to research methods and ibm spss analysis.* Chesapeake, VA: Watertree Press LLC.

Şahin, T. (2019). *Nedelsky, sınır grup ve karşıt gruplar standart belirleme yöntemlerinin norma dayalı değerlendirmelerle karşılaştırılması* [*Comparison of nedelsky, borderline group and constrasting groups standard setting models with norm referenced assessment*] [Master's dissertation, Hacettepe University]. Higher Education Institution National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Tanrıverdi, S. (2006*). Standart belirleme yöntemlerinin geçme puanları üzerine etkisi* [*Impacts of standard setting methods over passing*] [Master's dissertation, Hacettepe University]. Higher Education Institution National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Taşdelen, G. (2009*). Nedelsky ve Angoff standart belirleme yöntemlerinin genellenebilirlik kuramı ile karşılaştırılmasına ilişkin bir araştırma* [*A comparison of Angoff and Nedelsky cutting score procedures using generalizability theory*] [Master's dissertation, Hacettepe University]. Higher Education Institution National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Taşdemir, F. (2013). *Angoff (1-0), Nedelsky ve sınır değerleri saptama yöntemleri ile bir testin sınıflama doğruluklarının incelenmesi* [*Angoff (1-0), Nedelsky and examination of classification accuracies of a test by determination methods of limit values*] [Doctoral dissertation, Ankara University]. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Wang, L., Pan, W., & Austin, J.T. (2003). *Standards – setting procedures in accountability research: Impacts of conceptual frameworks and mapping procedures on passing rates.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Yıldırım Kan, N. (2019). *İngilizce hazırlık atlama sınavı için kesme puanı belirlenmesinde standart belirleme yöntemlerinin karşılaştırılması* [*Comparing standard setting methods while determining cut point for English proficiency exam*] [Master's dissertation, Hacettepe University]. Higher Education Institution National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Zieky, M.J., & Livingston, S.A. (1977). *Basic skills assessment: Manual for setting standards on the Basic Skills Assessment tests*. Educational Testing Service.

International Journal of Assessment Tools in Education
2025, Vol. 12, No. 1, 93–112
https://doi.org/10.21449/ijate.1545220
journal homepage: https://dergipark.org.tr/en/pub/ijate *Research Article*

Turkish adaptation of the digital literacy scale: A rasch analysis

Hongwei Yang[1*], Müslim Alanoğlu[2], Songül Karabatak[2], Kelly D. Bradley[3]

[1]University of West Florida, United States
[2]Fırat University, Faculty of Education, Elazığ, Türkiye
[3]University of Kentucky, United States

ARTICLE HISTORY

Received: Sep. 07, 2024

Accepted: Nov. 12, 2024

Keywords:

Differential item functioning,

Digital literacy scale,

Scale validation,

Rasch analysis,

University students.

Abstract: The study took a Rasch measurement theory approach to validating the 10-item Digital Literacy Scale (DLS) using the unidimensional rating scale model (RSM). To that end, the study used the data from a sample of online Turkish university students. The study began the Rasch analysis with all 10 items in the scale and, to improve in the local independence assumption, identified and eliminated two items which did not adequately fit the RSM. Under the eight-item DLS, the assumptions of undimensionality and local independence were both satisfied and the fit of all individual items to the RSM was adequate. Next, the psychometric properties of the eight-item DLS were examined including rating scale effectiveness, relative endorsability of the items, differential item functioning (DIF) by each of three demographic variables: (a) gender, (b) connection device, and (c) grade level. Through the analysis, evidence of reliability and validity was identified which generally supports the use of the DLS instrument among the population of online Turkish university students from which the sample was obtained. The study also identified items which demonstrated either misfit to the model or DIF by the demographic variables, and recommends they be further reviewed and revised for future use.

1. INTRODUCTION

The term of digital literacy (DL) was first introduced and made known by Gilster (1997). This landmark book defined DL as the ability to comprehend and utilize information in multiple formats from various sources when the information is presented using computers. This definition, although it first emerged almost three decades ago, may still have relevance today cause it does not present any listing of specific digital skills or technologies which have evolved rapidly over the years. Instead, it approaches DL from a general and broad perspective to allow the interpretation and operationalization of the DL concept to easily develop as necessary (Ala-Mutka, 2011). In the research literature, digital literacy has had different definitions which could have substantial similarities and overlapping, could be based on different theoretical frameworks, and, with the emergence of new digital technologies, new tools, and new literacies, could evolve over

*CONTACT: Hongwei YANG ✉ pyang@uwf.edu University of West Florida, United States

e-ISSN: 2148-7456

time (Amin *et al*., 2022; Gillen & Barton, 2010; Olur & Ocak, 2021; Reddy *et al*., 2023; UNESCO, 2018).

1.1. A Multi-Literacy, Multi-Perspective Approach to Digital Literacy

Ng (2012a, 2012b) defined digital literacy as referring to the multitude of literacies related to the use of digital technologies which include both software and hardware employed by individuals for educational, social, and/or entertainment purposes both in schools and at home. Among such hardware and software are desktops, laptops, handheld devices (e.g., tablets), game consoles, smartphones, commercial and open-source programs, etc. Under this framework, digital literacy consists of cognitive, technical, and socio-emotional learning perspectives/dimensions overlapping between and among themselves, and involves the acquisition of skills under each of the three perspectives/dimensions in order to effectively engage with online/offline digital technologies.

In the education setting, academic digital literacy may be perceived as the ability and awareness to take advantage of digital technology as a learning tool and complete academic tasks in the right way, when also encompassing the cognitive, technical, and socio-emotional perspectives of the literacy (Anwar *et al*., 2023; Hwang *et al*., 2023). Academic digital literacy has an important role to play because it is viewed as the backbone of educational pedagogy (Anwar *et al*., 2023). Graduates with digital literacy competencies are likely to have substantially better job prospects because, with a vast majority of the jobs requiring digital literacy (Anthonysamy *et al*., 2020; Perera *et al*., 2016; Setiyowati & Razak, 2020), such competencies could well increase their productivity in the digital era.

1.2. Research Related to Digital Literacy

Over the years, particularly during the COVID-19 pandemic when there was a substantially increased exposure and use of digital platforms and technologies of all kinds in all walks of life, numerous studies have been conducted on a global scale which address digital literacy in a variety of contexts: (a) assessing the awareness and competencies of DL of individuals (e.g., students, teachers, etc.), (b) investigating how DL is related to other measures of interest (e.g., self-efficacy, self-esteem, professional competence, individuals' demographic variables, teachers' readiness to implement digital technologies), (c) examining the effectiveness of DL programs, (d) narrowing the DL skills gap, etc. (Aydınlar *et al*., 2024; Ceylan *et al*., 2023; Erol & Aydin, 2021; Garzon & Garzon, 2023; Liza & Andriyanti, 2020; Reddy *et al*., 2021; Reddy *et al*., 2023). A detailed review of DL-related studies is beyond the scope of the study. Readers are referred to related systematic reviews such as Nguyen and Habók (2024), Gutiérrez-Ángel *et al*. (2022), Wu *et al*. (2022), and Pangrazio *et al*. (2020).

Among the DL-related research are studies which address the development/adaptation and validation of the scales/assessment tools/instruments measuring digital literacy for various stakeholders, cultural contexts, etc. These studies use statistical and psychometric means to validate a multitude of instruments measuring DL.

Ng (2012a) presented one of the first instruments measuring digital literacy which is known as Digital Literacy Scale (DLS) in the literature (There are other DL instruments bearing the same name (e.g., that developed in Chandra *et al*. (2024)), but they are not discussed here in this study). Based on her DL framework, Ng developed this 10-item instrument and used it and several other instruments to investigate the learning of unfamiliar educational technologies among a group of Australian undergraduate students enrolled in an introduction course on eLearning. In her study, the DLS asked the students to evaluate their level of digital literacy using a 10-point Likert scale. Even though her study hardly investigated the psychometric properties of the DLS, many follow-up studies conducted psychometric validation of the instrument under various contexts (Anwar *et al*., 2023;

Hamutoğlu *et al*., 2017; Ustundag *et al*., 2017), or applied the DLS to content area research (Aydınlar *et al*., 2024; Durak & Seferoğlu, 2020; Erol & Aydin, 2021; Garzon & Garzon, 2023; Noorrizki *et al*., 2022; Tor *et al*., 2022).

1.3. Existing Validation Research of the DLS

The literature has witnessed multiple validation studies of the DLS instrument. A review of these studies is provided here in chronological order as the context justifying this new research.

Hamutoğlu *et al*. (2017) adapted a 17-item version of the DLS instrument into the Turkish context. Notably, Ng (2012a) presented DL as consisting of three dimensions / perspectives (i.e., cognitive, technical, and social-emotional) which were covered by 10 items in three subscales. Hamutoğlu *et al*. (2017) included those 10 items in their version of the DLS instrument and additionally treated the seven items measuring attitudes towards information and communications technology as the fourth dimension. This practice was not consistent with Ng (2012a) and several other studies, like Anwar *et al*. (2023), which were all conducted under a three-dimension structure for DL measured by 10 items. Based on their 17-item DLS, Hamutoğlu *et al*. (2017) first conducted an expoloratory factor analysis (EFA) using a sample of 185 students and next a confirmatory factor analysis (CFA) using a sample of 210 students. At the end of the analyses, they presented both validity (e.g., language validity) and reliability (e.g., Cronbach's α, test-retest reliability) evidence for the 17-item scale.

Ustundag *et al*. (2017) translated the 10-item version of the DLS instrument by Ng (2012a) into the Turkish context and administered the adapted instrument to a group of pre-service teachers studying science. Unlike Ng (2012a) who hardly investigated the psychometric properties of the original instrument, Ustundag *et al*. (2017) validated the adapted instrument using common statistical methods for scale validation. Among the analyses they conducted was an EFA which established that the DLS in Turkish was unidimensional and had relatively high internal consistency reliability.

Finally, Anwar *et al*. (2023) based their study on the digital literacy definition and the three- dimension DL model from Ng (2012a, 2012b). They adapted the 10-item DLS into the Indonesian context for university students to measure their academic digital literacy. In the validation of the adapted instrument, they primarily took the CFA approach using the data collected from a sample of 364 Indonesian students. Their final model included a second-order CFA model measuring academic digital literacy predicting the three dimensions of DL outlined by Ng (2012a, 2012b). Besides, they also reported several reliability statistics including Cronbach's α, composite reliability, and average variance extracted. Given the findings, they recommended the use of the adapted instrument among the Indonesian university students.

1.4. Research Gaps in the Existing DLS Instrument Validation Studies

Despite the multitude of existing DLS validation studies, in general, the validation of a scale should be a continuous process (Gocen & Sen, 2021; Nunnally, 1978). This process could require multiple validation iterations to continuously identify more evidence of an instrument's reliability and validity, and could also entail a broader variety of samples to further refine and validate the instrument under more research contexts. On the other hand, there is also room for improvement in the existing validation studies which warrants more research.

First, the existing studies primarily counted on the traditional EFA/CFA for continuous data without (any mention of) taking into consideration the typically ordinal, rating scale structure of the DLS item data. Even though treating ordinal data as continuous has been a long term debate (Frampton & Shepherd, 2011), the literature of multiple fields of studies (e.g., healthcare, nursing, etc.) has nevertheless indicated doing so could well run the risk of erroneous results and mis-inference (Adroher *et al*., 2018; Cape *et al*., 2010; Da Dalt *et al*., 2013, 2015; Hamilton & Chesworth, 2013; Miot, 2020).

Second, no existing validation studies have investigated whether the DLS instrument functioned equivalently across subgroups which may be of research interest (e.g., subgroups by participant demographic characteristics). Therefore, their findings did not address whether the DLS items were unfair to, for example, a particular gender subgroup. For instance, Erol and Aydin (2021) and Tor *et al*. (2022) each compared different gender (female vs. male) subgroups regarding the research participants' level of digital literacy measured by the DLS instrument.

Unfortunately, both studies did so without having first examined whether the DLS items were biased by gender, which left open the question on whether the statistically significant differences from the independent samples t tests they conducted regarding the measure of DL were artifacts of the characteristics of the biased items, if any, or due to variations of participants' digital literacy at the scale and the subscale levels. Besides gender, the literature has indicated that digital literacy could be impacted by multiple demographic factors which include, but are not limited to, age, education, family income, use of smartphones and the Internet, years of service in the profession, daily Internet usage time, technology usage level, social media usage in distance education (Erol & Aydin, 2021; Noorrizki *et al*., 2022; Tor *et al*., 2022; Urbancikova *et al*., 2017). In order to examine the difference in the DLS scores, if any, across the subgroups specified by a demographic variable, the DLS items should be first verified to function the same way across these subgroups. This topic has not been investigated in the existing studies validating the DLS instrument.

1.5. Rasch Analysis as an Instrument Validation Tool

Rasch Measurement Theory (RMT) is a latent modeling framework which is based on modern test theory. In Rasch analysis, the raw, ordinal data (e.g., responses to Likert type items like DLS items) of the instrument are transformed to interval/continuous measures of participant ability and item difficulty on a logit scale along which a side-by-side comparison of participants and items is made (Andrich & Marais, 2019; Bond & Fox, 2015). Many, but not all, Rasch models assume that Rasch measurement involves a single, underlying construct (i.e., assumption of unidimensionality) either increasing or decreasing monotonically along the interval logit scale. Under the RMT, to make valid comparisons across different subgroups regarding a latent construct (e.g., digital literacy), the items should function the same way across different subgroups of participant demographic characteristics (e.g., gender) (Hagquist *et al*., 2009; Hagquist, 2019). Otherwise, comparisons of scores across the subgroup participant characteristics (e.g., female vs. male) may be invalid. Such a violation of the requirement of invariance across subgroups is known as differential item functioning (DIF; Hagquist, 2019). In summary, RMT methods are designed to properly handle the ordinal categorical data. They can complement the traditional methods in psychometrics (e.g., proportion of correct responses as a measure of item difficulty) to provide additional evidence of reliability and validity of an instrument. Over the years, they have been widely used in studies (e.g., those validating a scale) in education including those of online education (e.g., Ningsih *et al*. (2021)), artificial intelligence in education (e.g., Capinding (2024)), among others.

1.6. Research Questions

Rasch analysis provides a detailed analysis of many aspects of an instrument when also being able to address the research gaps (e.g., taking into consideration the ordinal, rating scale structure of the DLS data, investigating item DIF, etc.) outlined above.

However, an extensive literature review indicates that there have not been any studies reporting the psychometric properties of the DLS instrument by means of RMT.

Given the discussions above, the study proposed three research questions (RQs) regarding the DLS instrument:

1. RQ1: Does the DLS instrument measure a unidimensional construct of digital literacy?
2. RQ2: What are the psychometric properties of the DLS instrument, after properly taking into account the rating scale structure of the DLS response data?
3. RQ3: Do the DLS items function equivalently across the subgroups specified by participants' demographic measures?

1.7. Organization of Research

The study is organized as follows. The study begins with an introduction of the research context, which is followed by a review of the existing DLS scale validation research and gaps in such research. Rasch analysis is introduced as a psychometric method addressing the gaps. Next come the research questions with regard to the DLS instrument which were formulated based on the literature review, outlined research gaps, and introduction of Rasch analysis. The study proceeds to a methodology section which examines the psychometric properties of the DLS under Rasch analysis. In the end, the study discusses the findings, implications, and limitations and future research before providing the final conclusions.

2. METHOD

2.1. DLS Instrument and Demographic Measures

This study used the 10-item (Table 1) version of the DLS instrument by Ng (2012a). Each DLS item is measured on a five-point Likert scale, ranging from 1 = *Strongly Disagree* to 5 = *Strongly Agree*. Note that, although Ng (2012a) developed the DLS on a 10-point Likert scale, many follow-up (scale validation or content area) studies (e.g., Ustundag *et al*. (2017), Garzon and Garzon (2023), among others) used a five-point Likert scale, instead, and this study followed the same practice. Finally, because all 10 DLS items are positively worded, a higher score on an individual DLS item, a subscale, and the scale as a whole corresponds to a higher level of digital literacy.

Table 1. *DLS items.*

Items	Item statements
DLS01	I know how to solve my own technical problems.
DLS02	I can learn new technologies easily.
DLS03	I keep up with important new technologies.
DLS04	I know about a lot of different technologies.
DLS05	I have the technical skills I need to use ICT[a] for learning and to create artifacts (e.g., presentations, digital stories, wikis, blogs) that demonstrate my understanding of what I have learned.
DLS06	I have good ICT[a] skills.
DLS07	I am confident with my search and evaluation skills in regards to obtaining information from the Web.
DLS08	I am familiar with issues related to web-based activities e.g., cyber safety, search issues, plagiarism.
DLS09	ICT enables me to collaborate better with my peers on project work and other learning activities.
DLS10	I frequently obtain help with my university work from my friends over the Internet e.g. through Skype, Facebook, Blogs.

Note. The sample size is consistently n = 404 across all 10 DLS items.
[a]ICT = Information and Communication Technology.

Regarding the demographic items, there were three dichotomously-coded ones: (a) gender, (b) connection device, and (c) grade level. Gender consists of the two categories of females and males, connection device the two categories of computers (desktop and laptop) and handheld devices (smart phone and tablet), and grade level the two categories of lower (first- and second- years) and higher (third- and fourth-years) grades of undergraduate students.

2.2. Participants and Data Collection

After securing the required approval from the research ethics committee of the research site of a Turkish university, the study proceeded to obtain a convenience sample. The data were collected in the university as part of a larger cross-sectional study among its undergraduate students of education taking online courses. After properly preparing the collected data, the final sample size of each item was consistently n = 404.

2.3. Rasch Analysis

The data were first summarized using descriptive statistics which were based on several breakdowns of the participants' demographic characteristics. Next, a Rasch analysis of the data was conducted using the Rasch Rating Scale Model (RSM) in Winsteps 5.6.4.0 (Linacre, 2023). An RSM is a type of Rasch model for polytomous data usually produced from a Likert scale.

The model requires every item should have the same number of response categories (e.g., the DLS instrument where all items have five response options). Besides, to each item, the model applies the same number of response thresholds, with which to progress from one response option to the next (e.g., from *Agree* to *Strongly Agree*); across all items, the relative distance between each pair of threholds remains the same, although each item is still allowed to have its own level of difficulty.

The RSM-based Rasch analysis began with all 10 items in the model and assessed the statistical assumptions (i.e., assumptions of unidimensionality and local independence) underlying the RSM and the fit of the data to the model. In the case of a problem (e.g., assumption violation, inadequate fit of the item data to the model, etc.), appropriate measures were taken to address it. After the assumptions were fully satisfied and the fit of the item data to the model was improved to an acceptable level, the Rasch analysis of the instrument was advanced to produce more evidence of reliability and validity.

3. RESULTS

As was shown in Table 1, the dataset contained 404 participants providing complete responses to all 10 DLS items. Therefore, the dataset led to a high participant-item ratio of about 40:1, satisfying the criterion that the sample size should be at least six times the number of items for stable results in factor analysis of which Rasch analysis is a special type for categorical data (Bartholomew *et al.*, 2008; Mundfrom *et al.*, 2005; Skrondal & Rabe-Hesketh, 2004).

3.1. Descriptive Statistics

Regarding the participant demographics, the sample of 404 participants ranged from 18 to 46 years old in age (M = 24.03, SD = 4.39) and consisted of 308 females and 96 males. They used different devices to connect to the Internet: (a) n = 21 using a desktop, (b) n = 156 using a laptop, (c) n = 216 using a smart phone, and (d) n = 8 using a tablet. Finally, they came from four different grades: (a) n = 31 from first-year, (b) n = 53 from second-year, (c) n = 40 from third-year, and (d) n = 280 from fourth-year.

Further, the mean response scores for individual items (computed by averaging all responses to each item across all participants who responded to the item) fall between *Agree* (= 4) and *Neither Agree nor Disagree* (= 3), ranging from 3.11 for DLS06 and to 3.89 for

DLS07. All items put together, the most frequently selected category is *Agree* (32.4%), which is immediately followed by *Neither Agree nor Disagree* (28.2%).

Finally, Table 2 documents the response frequencies of the categories of individual DLS items. According to the table, *Agree* is the most frequently selected category on five items (ranging from 27.7% for DLS10 to 40.6% for DLS07), and *Neither Agree nor Disagree* is most frequently selected on the other five items (ranging from 31.2% for DLS05 to 35.1% for DLS01). As a summary, the observations from descriptive statistics suggest the student participants mostly perceived neutrally to favorably of how well the items described their levels of digital literacy.

Table 2. *Summary of responses to all 10 DLS items.*

Items	Strongly Disagree (%)	Disagree (%)	Neither Agree nor Disagree (%)	Agree (%)	Strongly Agree (%)
DLS01	4.7	10.9	35.1	32.4	16.8
DLS02	3.2	6.9	19.8	39.9	30.2
DLS03	5.7	10.1	24.3	35.9	24.0
DLS04	5.0	16.8	31.9	28.5	17.8
DLS05	5.9	16.1	31.2	30.0	16.8
DLS06	10.4	18.8	33.7	23.5	13.6
DLS07	1.7	6.9	21.5	40.6	29.2
DLS08	5.2	12.4	31.4	29.0	22.0
DLS09	3.0	8.2	25.5	36.1	27.2
DLS10	8.4	14.6	27.5	27.7	21.8

3.2. Rasch Analysis

The study began with all 10 DLS items analyzed under the RSM and assessed whether the two statistical assumptions of the RSM were satisfied: unidimensionality and local independence (Bond & Fox, 2015).

3.2.1. *Analyzing 10-item DLS*

3.2.1.1. Assumption of Unidimensionality. This assessment of the unidimensionality assumption served to see if the DLS instrument, as a whole, measures a single underlying construct of digital literacy that the instrument was designed to measure. To that end, a principal component analysis (PCA) was used of the correlation matrix of standardized Rasch residuals (Bond & Fox, 2015; Linacre, 2023).

According to the Winsteps PCA output, the statistics of explained raw score variance in the observations/observed data by measures (i.e., items and persons) in the *Observed* column and those in the *Expected* column were about the same size (for persons, 46.6% under *Observed* vs. 46.7% under *Expected*; for items, 9.1% under *Observed* vs. 9.1% under *Expected*), indicating there was no problem in the model estimation and that the data provided an adequate fit to the Rasch model assuming unidimensionality (Linacre, n.d.; Linacre, 2018, September 2). Second, the contrasts were examined which were computed after the Rasch dimension was extracted from the data. Specifically, the first contrast (i.e., the first dimension beyond the Rasch dimension) had an eigenvalue of 1.9717, which was lower than 2, the size of an eigenvalue expected by chance. This evidence did not support the existence of a secondary dimension in the data (Linacre, 2023). Based on the multiple pieces of evidence from both the statistical analyses and the literature, the study concludes with the unidimensionality (i.e., Rasch dimension) of the 10-item DLS.

3.2.1.2. Assumption of Local Independence. Also assessed here was the local independence assumption which states that, after controlling for the underlying latent trait of digital literacy, the responses to one survey item do not covary with the responses to other items (Aryadoust *et al*., 2021; Borsboom, 2005). That is, in Rasch measurement, since DLS items are regressed on the latent variable of digital literacy, the local independence assumption requires that the unexplained variances in the DLS items should not correlate with each other. For the 10-item DLS, the local independence assumption was assessed using the correlations between the residuals of the DLS items (i.e., Q3 coefficients). (Fan & Bond, 2019; Lee, 2004; Wright, 1996; Yen, 1984). A Q3 coefficient larger than .30 in absolute value indicates a respectable degree of local dependence. Examining the Winsteps output of the largest standardized residual correlations of DLS items showed that the correlations in absolute value between the residuals of three pairs of items were higher than .30: (a) (-.32) between DLS03 and DLS10, (b) (-.31) between DLS02 and DLS06, and (c) (-.30) between DLS04 and DLS10. Therefore, there was a violation of the assumption of local independence among the three pairs of items. To find more evidence for addressing this assumption violation, individual item fit was next examined.

3.2.1.3. Individual Item Fit. Examining the item fit output containing the mean-square (MNSQ) infit and outfit statistics, one and only one item, DLS10, had an unusually large infit MNSQ (1.84) and outfit MNSQ (1.95) at the same time. Because these statistics were greater than 1.50, it indicates that, with this item, off-variable noise was markedly greater than useful information. As a result, even though these diagnostic statistics were (close to but) still not higher than the reshold of 2.00 indicative of degradation of measurement, the item may nonetheless need to be further scrutinized and revised to remedy its misfit to the model. Other than DLS10, the other nine items were all productive of measurement. None of them exhibited any substantial misfit to the Rasch model because their infit and outfit statistics were at most 1.28 (infit MNSQ) and 1.29 (outfit MNSQ) for DLS06 and at least 0.67 (infit MNSQ) and 0.73 (outfit MNSQ) for DLS04, which all fell into the range of 0.50 – 1.50 indicating productive of measurement. Finally, the point-polyserial correlations for all 10 items were high and positive where the lowest correlation was that for DLS10 at .60 and all other correlations were at least .71 (DLS06 and DLS07), indicating the orientation of the scoring on each DLS item was well aligned with the orientation of the latent variable measured by this instrument and that the items had adequate discriminatory power. The point-polyserial correlation for DLS10 was positive but was also markedly lower than the other nine correlations. Therefore, DLS10 probably did not have as much discriminatory power as any of the other nine items (Bond & Fox, 2015; Linacre, 2023).

3.2.2. *Analyzing eight-item DLS*

Based on the analyses above, DLS10 was identified as not having an adequate fit to the model and was among the items which led to a violation of the local independence assumption. Therefore, DLS10 was removed and the above analyses were repeated with the remaining nine items. This time, the unidimensionality assumption continued to be satisfied. But, there was still one pair of items, DLS02 and DLS06, with the Q3 coefficient being (-.34) whose absolute value was higher than .30. Therefore, with the remaining nine items, the local independence assumption was violated again. Next, the fit of individual items was examined. Among the remaining nine items, DLS06 had the highest outfit MNSQ (1.53) which did not fall into the range of 0.50-1.50, and its infit MNSQ (1.48), although also the highest, fell into the range. All other items had both infit and outfit MNSQ statistics in the range of 0.50-1.50. Given the information above, out of the only pair of items (DLS02 and DLS06) whose Q3 coefficient indicated a violation of the local independence assumption, DLS06 was removed from further consideration. There was a total of eight items left in the scale.

3.2.2.1. Overview of the Eight-Item Scale. Next, the eight-item scale was examined under the Rasch rating scale model. First, the assumption of unidimensionality was satisfied. The statistics of explained raw score variance in the *Observed* column and those in the *Expected* column were virtually identical (for persons, 53.7% under *Observed* vs. 53.6% under *Expected*; for items, 8.1% under *Observed* vs. 8.1% under *Expected*), indicating there was no problem in model estimation and the data provided an adequate fit to the unidimensional Rasch model. The first contrast beyond the Rasch dimension had an eigenvalue of 1.7880, which was lower than 2 and thus did not support the existence of a secondary dimension in the data. Second, the local independence assumption was satisfied under the eight-item scale because, among the correlations between the residuals of the eight DLS items, the highest was .29 in absolute value which was lower than .30. Third, regarding the fit of individual items to the model, all infit and outfit MNSQ statistics fell into the range of .50-1.50 (highest and lowest infit MNSQ statistics were, respectively, 1.20 for DLS08 and 0.78 for DLS04; highest and lowest outfit MNSQ statistics were, respectively, 1.20 for DLS08 and 0.76 for DLS02). Therefore, given the statistics above, the eight-item scale met the assumptions of undimensionality and local independence and provided an adequate fit at both the overall and individual item levels. It was therefore further examined and interpreted under the Rasch model.

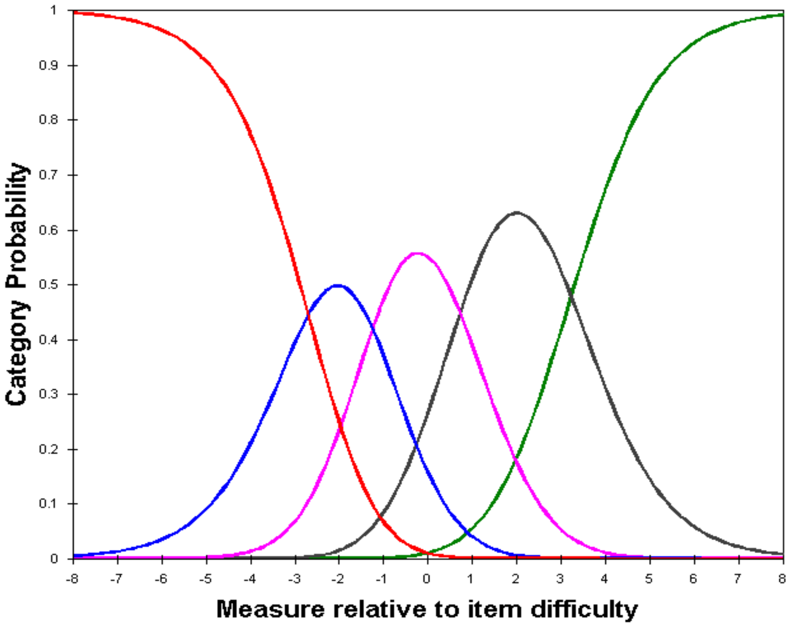
3.2.2.2. Separation and Reliability. In the eight-item DLS, person and item separation statistics were, respectively, as high as 2.71 and 6.16. The high person separation statistic indicated the DLS instrument was adequately sensitive to distinguish between individual participants with higher and lower levels of digital literacy, and the high item separation statistic indicated the sample was large enough to confirm item difficulty/endorsability/agreeability hierarchy. Regarding the reliability statistics, person reliability was .88 (i.e., the DLS instrument discriminated the participants into adequate levels of digital literacy), and item reliability was also very high at .97 (i.e., the sample was large enough to precisely locate the items on the underlying latent difficulty/endorsability/agreeability continuum) (Bond & Fox, 2015; Linacre, 2023).

3.2.2.3. Rating Scale Effectiveness for DLS. The study also examined the rating scale effectiveness of the eight items in DLS. First, according to the response category probability curves shared by all eight items in the scale (Figure 1), each category had a distinctive peak indicating it was a meaningful endorsement choice for the participants at a certain level of ability as measured in DLS. Stated differently, the Turkish student participants were capable of adequately separating one response category from another in the eight DLS items, which served as evidence of validity (Bond & Fox, 2015; Linacre, 2023).

Second, regarding the quality of the rating scale categories, none of the outfit MNSQ statistics on the five categories was greater than 2. The infit MNSQ statistics ranged from 0.86 for *Agree* (= 4) to 1.14 for *Strongly Agree* (= 5) and the outfit MNSQ statistics from .85 for *Agree* (= 4) to 1.14 for *Disagree* (= 2), indicating that none of the categories was introducing more noise than meaning into the measurement process and thus warranted further empirical investigation (e.g., considered as a candidate for collapsing with adjacent categories) (Bond & Fox, 2015). Third, the measure of Andrich threshold advanced in a stepwise manner (the four threshold statistics (-2.70 < -1.25 < 0.72 < 3.23) ascended monotonically in value up the rating scale) as anticipated, indicating that the lower threshold was always smaller than the higher threshold in an adjacent pair of categories. Stated differently, there was no disordering of thresholds (Bond & Fox, 2015; Linacre, 2023).

As a summary, the findings here support the rating scale structure of the DLS instrument functioned in the intended way, and that the response categories were correctly and consistently interpreted by the student participants as the sequence of most likely outcomes.

Figure 1. *Response category probability curves shared by all eight items in the DLS instrument.*



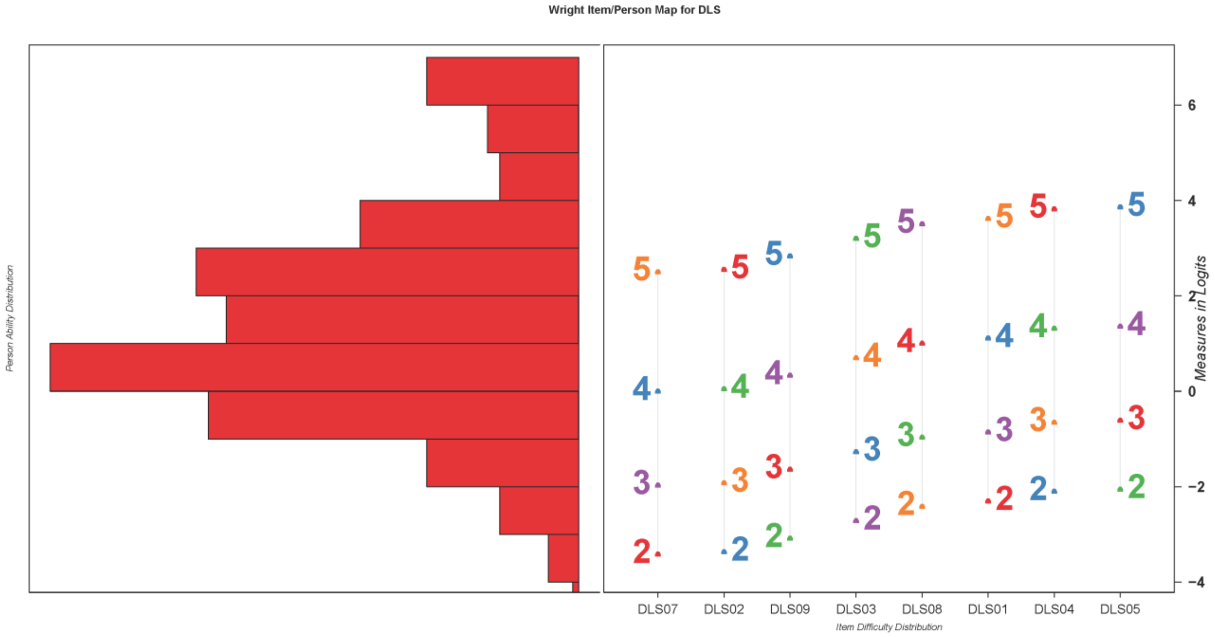
Note. Curve peaks for response categories (from left to right): 1 = *Strongly Disagree*, 2 = *Disagree*, 3 = *Neither Agree nor Disagree*, 4 = *Agree*, 5 = *Strongly Agree*.

3.2.2.4. Wright Item/Person Map for DLS. The Wright, item/person map in Figure 2 visually demonstrates and rank-orders the relative difficulty/endorsability/agreeability of the eight DLS items and students' level of digital literacy. In the right portion of the panel, from left to right, items are ranked from the most favorite item (i.e., easiest to endorse) to the least favorite item (i.e., hardest to endorse) and the four Andrich thresholds (i.e., step values) of the RSM for each individual DLS item with five response categories are indicated vertically and in ascending order by numeric values of 2, 3, 4, and 5 above that item; in the left portion of the panel, from bottom to top, students are ranked from those who had the lowest level of digital literacy to those who had the highest level of digital literacy (Linacre, 2023).

Based on Figure 2, the student participants most easily endorsed DLS07, "*I am confident with my search and evaluation skills in regards to obtaining information from the Web.*" and DLS02, "*I can learn new technologies easily.*". Next, in an ascending order of difficulty, the students almost equally easily endorsed DLS09, "*ICT enables me to collaborate better with my peers on project work and other learning activities.*". However, when it came to DLS03, "*I keep up with important new technologies.*", the item was more difficult to endorse by the students than the previous items. Next, at a higher level of difficulty was DLS08, "*I am familiar with issues related to web-based activities e.g., cyber safety, search issues, plagiarism.*". Even more difficult to endorse was DLS01, "*I know how to solve my own technical problems.*". Finally, the two most difficult-to-endorse items were DLS04, "*I know about a lot of different technologies.*", and, subsequently, DLS05, "*I have the technical skills I need to use ICT for learning and to create artifacts (e.g., presentations, digital stories, wikis, blogs) that demonstrate my understanding of what I have learned.*".

The results indicated that, overall, the student participants willingly demonstrated their confidence in the level of digital literacy. However, that confidence might not have easily translated into the participants' actual digital literacy. Therefore, it was not surprising to see that they were hesitant to acknowledge that they actually had the knowledge, technologies, or skills.

Figure 2. *Wright item/person map for validating DLS.*



Note. In the right portion of the panel, the four Andrich thresholds (i.e., step values) of the RSM for each individual DLS item with five response categories are indicated vertically and in ascending order by numeric values of 2, 3, 4, and 5 above that item.

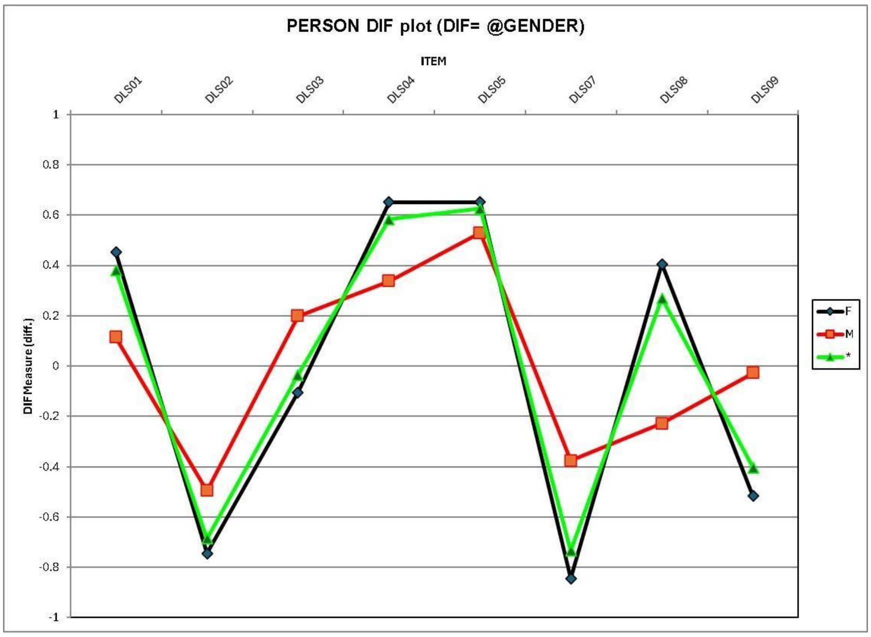
3.2.2.5. Differential Item Functioning Analysis of DLS. A pairwise differential item functioning analysis of the items in DLS by each of three dichotomously-coded demographic items (i.e., gender, connection device, and grade level) was conducted where the null hypothesis was set up that each DLS item had the same level of difficulty for the two subgroups specified by each demographic variable. Both statistical significance and substantive significance were assessed using, respectively, (a) the Rasch-Welch t and the Mantel χ^2 tests and (b) the cumulative log-odds ratio approximating the DIF size for polytomous data (Linacre, 2023). The results of the three DIF analyses are outlined in Table 3.

Table 3. *Results of three DIF analyses.*

Items	Female Minus male			Computer Minus handheld device			Higher grade Minus lower grade		
	Rasch-Welch *t* test	Mantel χ^2 test	DIF size	Rasch-Welch *t* test	Mantel χ^2 test	DIF size	Rasch-Welch *t* test	Mantel χ^2 test	DIF size
DLS01	.0769	.0898	.48	.5284	.7322	.08	.9100	.9278	-.02
DLS02	.2090	.0821	-.53	.4600	.1165	.41	.9007	.9874	.00
DLS03	.1120	.1179	-.45	.4778	.1850	.34	1.0000	.6525	.13
DLS04	.0965	.1176	.43	1.0000	.9408	-.02	.4327	.2969	-.30
DLS05	.5103	.4180	.21	1.0000	.8805	-.03	1.0000	.6130	-.13
DLS07	.0189	.0401	-.62	1.0000	.6507	.11	.9089	.9262	.03
DLS08	.0013	.0133	.75	.6728	.3403	-.23	.1289	.2401	.33
DLS09	.0123	.1333	-.41	.1582	.0528	-.45	.4965	.7500	-.09

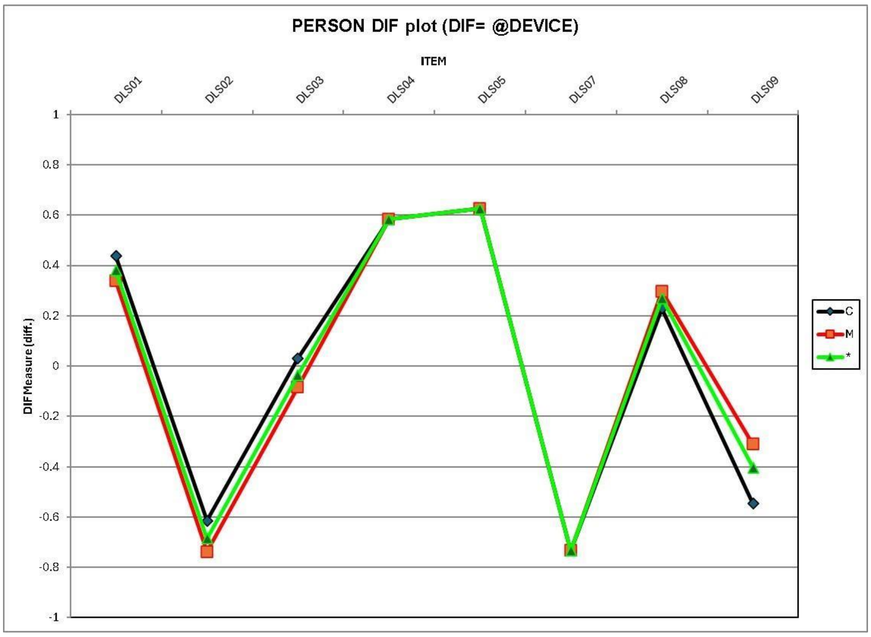
3.2.2.5.1. DIF analysis by Gender. Per the measure of DIF contrast for each item computed as the difficulty estimate of the item for females minus that for males, two items were statistically significant at the .05 level of significance on both the Rasch-Welch t and the Mantel χ^2 tests: DLS07 and DLS08. DLS07 had a negative DIF contrast and therefore was easier for the female subgroup than for the male subgroup. In comparison, since DLS08 demonstrated a positive DIF contrast, this item was the other way around (i.e., more difficult for the female subgroup than for the male subgroup). Next, both items demonstrated a moderate to large level of DIF (i.e., Level *C* DIF): (a) DLS07 with a DIF size of (-0.62) and (b) DLS08 with a DIF size of 0.75. Finally, DLS09 was significant on the Rasch-Welch t test, p = .0123, but not significant on the Mantel χ^2 test, p = .1333. It had a negligible DIF size of (-0.41) (i.e., Level *A* DIF) (Linacre, 2023; Zwick, 2012; Zwick *et al*., 1999). Finally, the DIF analysis by gender is presented graphically in Figure 3.

Figure 3. *DIF analysis by gender.*



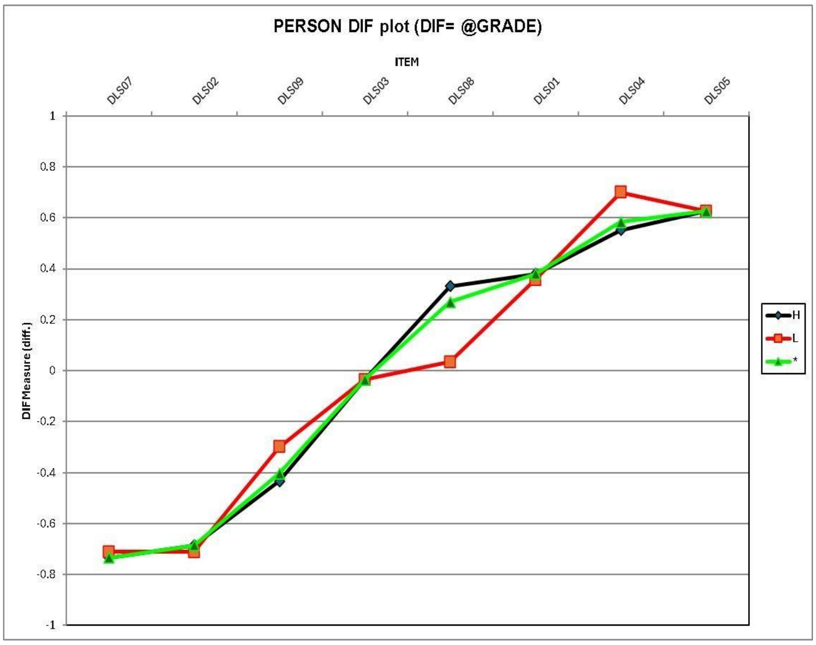
3.2.2.5.2. DIF analysis by Connection Device. Per the measure of DIF contrast for each item computed as the difficulty estimate of the item for computers minus that for handheld devices, none of the eight items was statistically significant at the .05 level of significance on any of the Rasch-Welch t and the Mantel χ^2 tests. Next, all but one item demonstrated a negligible level of DIF (Level *A* DIF) and DLS09 demonstrated a slight to moderate level of DIF of (-.45) (Level *B* DIF) (Linacre, 2023; Zwick, 2012; Zwick *et al*., 1999). Finally, the DIF analysis by connection device is presented graphically in Figure 4.

Figure 4. *DIF analysis by connection device.*



3.2.2.5.3. DIF analysis by Grade Level. Per the measure of DIF contrast for each item computed as the difficulty estimate of the item for higher grade students minus that for lower grade students, none of the eight items was statistically significant at the .05 level of significance on any of the Rasch-Welch t and the Mantel χ^2 tests. Next, all items demonstrated a negligible level of DIF (Level A DIF) (Linacre, 2023; Zwick, 2012; Zwick *et al*., 1999). Finally, the DIF analysis by grade level is presented graphically in Figure 5.

Figure 5. *DIF analysis by grade level.*



4. DISCUSSION

In the digital era, digital literacy is constantly referred to and its importance is evidenced by the numerous efforts at different levels (e.g., regional, national, etc.) to develop and implement DL frameworks and strategic plans to support and improve their citizens' digital literacy (UNESCO, 2018). In an effort to contribute to the proper measurement of digital literacy, the study adapted to the Turkish context the Digital Literacy Scale through a Rasch measurement theory perspective. The study identified evidence of undimensionality of the eight-item DLS instrument, which is very close to the conclusion of Ustundag *et al*. (2017) stating that all ten DLS items constituted a unidimensional measure of digital literacy. By contrast, the study may differ from other DLS adaptation research, such as Anwar *et al*. (2023) and Hamutoğlu *et al*. (2017), in terms of the conclusion on scale dimensionality and an attempt is made at a later point in this study to address such a discrepancy. Besides, new evidence of reliability and validity was found in the study which provided more insights into the psychometric properties of the DLS items. Three research questions were proposed and addressed.

4.1. Addressing Research Questions

Regarding RQ1, the study found that, with all 10 items in the scale, the fundamental unidimensionality assumption of the RSM was satisfied. However, the 10-item scale led to a violation of the local independence assumption. After identifying and removing two items, DLS10 and DLS06, the unidimensionality and local independence assumptions were both satisfied under the eight-item DLS. Finally, because DLS10 and DLS06 exhibited a misfit to the model, DLS10 in particular, they merit further review and revision to prevent them from degrading the measurement of digital literacy.

Regarding RQ2, the study conducted a Rasch analysis under the eight-item DLS to investigate item/person separation and reliability, rating scale effectiveness, and relative endorsability of items. In the analysis, measures of item/person separation and reliability were all high. The high level of person separation indicated the DLS instrument was able to distinguish between participants with higher and lower levels of digital literacy, and the high level of item separation indicated the sample was adequately large to confirm item endorsability hierarchy. That the item and person reliability measures were high suggested the item difficulty and participant ability measures would be highly reproducible, should the same test be administered to the same group of student participants repeatedly. Next, a diagnostic analysis of the rating scale effectiveness in the eight-item scale indicated its response categories functioned as intended, and that the participants were able to adequately separate one response category from another and correctly and consistently interpret the response categories. Finally, in the Wright, item/person map, the item hierarchy measuring relative endorsability was demonstrated. Overall, the student participants easily agreed they were confident in their level of digital literacy, but that confidence did not easily translate into the actual digital literacy skills they would acknowledge they had.

Regarding RQ3, the study conducted a DIF analysis under the eight-item DLS to see if any items were endorsed to different extents by the two subgroups specified by each of the three demographic variables: (a) gender, (b) connection device, and (c) grade level. First, under gender, two items, DLS07 and DLS08, demonstrated statistical significance as measured by both the Rasch-Welch t and the Mantel χ^2 tests. DLS07 was easier for females to endorse than for males, whereas the DIF of DLS08 was in the opposite direction. Both DLS07 and DLS08 demonstrated a Level C DIF. Besides DLS07 and DLS08, DLS09 was significant on the Rasch-Welch t test only and demonstrated a Level A (i.e., negligible) DIF. Second, under connection device, none of the eight items was significant on any of the Rasch-Welch t and the Mantel χ^2 tests. DLS09 was the only item which demonstrated a slight to moderate level of (i.e., Level B) DIF. Third, under grade level, none of the eight items was significant on any of the Rasch-Welch t and the Mantel χ^2 tests, neither was there any item demonstrating a level of DIF beyond negligible.

Because several items were flagged as having gender-related DIF in this study, it is reasonable to be wondering if the gender-based comparisons presented in studies like Erol and Aydin (2021) would have led to different results. Therefore, such studies should probably have begun with an assessment of whether gender-related DIF existed on any items before comparing the two gender subgroups on digital literacy at the scale and subscale levels. This assessment is necessary because differences in DLS scores between the gender subgroups could reflect the characteristics of DLS items instead of variations in the participants' level of digital literacy that the study intended to assess. In the long run, it is important to be aware of any bias coming from item DIF, particularly if thresholds are to be applied to the DLS scores to inform decisions on diagnosis and subsequent interventions or treatments. When DIF exists, the associated bias could lead to under- or over-intervention or treatment for certain subgroups, depending on the direction of the bias. Accordingly, it is important for the DLS instrument to be assessed for DIF and the extent to which it exists should be taken into consideration when interpreting the DLS scores (Cameron *et al*., 2014).

4.2. Implications

The DLS instrument, together with its adaptation using the Rasch measurement theory in this study, has implications for assisting researchers, policymakers, instructional designers, and online instructors. This instrument is well-suited for gaining insights into the specific digital literacy requirements of Turkish university students as they engage with digital technologies. Furthermore, it may also serve as a catalyst for targeted interventions and programs aimed to improve the digital literacy skills of Turkish university students. By properly measuring the digital literacy of university students, this assessment tool likely has the potential to improve efficiency, effectiveness, and success in the adoption of ICT-based online education practices.

4.3. Limitations and Future Research

This study has its limitations which may serve as grounds of future research. First, this study is limited to examining the effects of three demographic variables as potential sources/covariates of DIF and the findings already cast doubts on the results from the existing literature (e.g., Erol and Aydin (2021)). Future research might investigate other possible sources/covariates (e.g., race/ethnicity) of DIF which might be of interest to content area researchers. Second, in this study, the DLS survey was not completed over time and no consideration was given to the ability of the DLS instrument to identify changes in digital literacy longitudinally. Future research might focus on the longitudinal measurement invariance aspect of the psychometric properties of the DLS instrument to see whether the DLS items assess the same digital literacy construct invariantly across time (Horn & McArdle, 1992; Liu *et al*., 2017; Meredith, 1993). For example, yearly, as in Lazonder *et al*. (2020).

Third/finally, the current study is limited in that it did not evaluate the bifactor model as an alternative structural representation (e.g., dimensionality) of the DLS instrument. Although this study presented evidence of undimensionality and this conclusion is largely consistent with that from certain previous research (e.g., Ustundag *et al*. (2017)), there are nonetheless other DLS validation studies (e.g., Ng (2012a) and Anwar *et al*. (2023)) which demonstrate the DLS instrument is multi-dimensional. A tentative explanation for this discrepancy might be that neither conclusion adequately explains the true dimensionality of the DLS instrument. Instead, a combination of the two solutions in the form of a bifactor model (Chen *et al*., 2012; Gignac & Kretzschmar, 2017; Reise, 2012; Reise *et al*., 2007; Rodriguez *et al*., 2016a, 2016b) might provide a fuller representation of the underlying structure of the DLS instrument. Under a bifactor model, previous research has indicated that an instrument consisting of multiple dimensions/subscales could be consistent with both a unidimensional and a multi-dimensional model but may be alternatively and likely better represented by the bifactor structure (Reise *et al*., 2007). For example, the bifactor structure might be able to more effectively handle the violation of the local independence assumption due to item clustering demonstrated earlier in the study. Besides, the DL framework proposed in Figure 1 from Ng (2012a) features three separate circles (e.g., representing the three dimensions of DL: cognitive, technical, and socio-emotional learning) overlapping in pairs and in an intersection of all three circles. The bifactor model can not only include the overlap of each pair of DL dimensions but also incorporate the intersection of all three DL dimensions into a general DL measure underlying all DLS items, thus suggesting the bifactor structure is likely more aligned with the DL framework on which the DLS instrument is based. In summary, given the unique features of the bifactor model, this alternative structure might be another direction of future research.

5. CONCLUSION

As a summary, the study largely reconfirmed the unidimensional structure of the DLS instrument as was previously reported in the literature (e.g., Ustundag *et al*. (2017)). From the perspective of Rasch measurement theory, the study identified new evidence of reliability and validity to show the DLS instrument is mostly psychometrically sound and therefore is able to produce high quality data measuring digital literacy, which largely supports the findings of the literature that the DLS instrument has a special potential in the research of digital literacy among the Turkish university students. Items demonstrating misfit or DIF were identified which should be further examined and revised using both statistical and nonstatistical criteria through an iterative process.

Acknowledgments

The authors acknowledge the work of the editorial team of the International Journal of Assessment Tools in 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**: Firat University Social Sciences and Humanities Ethics Committee, 23.05.2023-16088.

Contribution of Authors

Hongwei Yang: Conducted the Rasch analysis in the Winsteps software. **Müslim Alanoğlu**: Took responsibility in data collection and related matters. **Songül Karabatak**: Took responsibility in data collection and related matters. **Kelly D. Bradley**: Provided methodological support on Rasch analysis. Finally, all four authors contributed to the writing and finalization of the manuscript.

Orcid

Hongwei Yang https://orcid.org/0000-0003-0225-3792
Müslim Alanoğlu https://orcid.org/0000-0003-1828-4593
Songül Karabatak https://orcid.org/0000-0002-1303-2429
Kelly D. Bradley https://orcid.org/0000-0002-4682-8212

REFERENCES

Adroher, N.D., Prodinger, B., Fellinghauer, C.S., & Tennant, A. (2018) All metrics are equal, but some metrics are more equal than others: A systematic search and review on the use of the term metric'. *PloS ONE, 13*(3), e0193861. https://doi.org/10.1371/journal.pone.0193861

Ala-Mutka, K. (2011). *Mapping digital competence: Towards a conceptual understanding* (Technical Note: JRC 67075). https://doi.org/10.13140/RG.2.2.18046.00322

Amin, H., Malik, M.A., & Akkaya, B. (2022). Development and validation of Digital Literacy Scale (DLS) and its Implication for higher education. *International Journal of Distance Education and E-Learning, 7*(1), 24-43. https://doi.org/10.36261/ijdeel.v7i1.2224

Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory: Measuring in the educational, social and health sciences*. Springer.

Anthonysamy, L., Koo, A.C., & Hew, S.H. (2020). Self-regulated learning strategies in higher education: Fostering digital literacy for sustainable lifelong learning. *Education and Information Technologies, 25,* 2393-2414. https://doi.org/10.1007/s10639-020-10201-8

Anwar, Z., Hanurawan, F., Chusniyah, T., & Setiyowati, N. (2023). Adaptation of the Academic Digital Literacy Scale for college students: A validity and reliability study. *Psychological Science and Education, 28*(4), 98-111. https://doi.org/10.17759/pse.2023280406

Aryadoust, V., Ng, L.Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing, 38*(1), 6-40. https://doi.org/10.1177/0265532220927487

Aydınlar, A., Mavi, A., Kütükçü, E., Kırımlı, E.E., Alış, D., Akın, A., & Altıntaş, L. (2024). Awareness and level of digital literacy among students receiving health-based education. *BMC Medical Education,* 24-38. https://doi.org/10.1186/s12909-024-05025-w

Bartholomew, D.J., Steele, F., Moustaki, I., & Galbraith, J. (2008). *Analysis of multivariate social science data* (2nd ed.). Routledge.

Bond, T.G., & Fox, C.M. (2015). *Applying the Rasch model* (3rd ed.). Routledge.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.

Cameron, I.M., Scott, N.W., Adler, M., & Reid, I.C. (2014). A comparison of three methods of assessing Differential Item Functioning (DIF) in the hospital anxiety depression scale: Ordinal logistic regression, Rasch analysis and the Mantel chi-square procedure. *Quality of Life Research*, *23*(10), 2883-2888. https://doi.org/10.1007/s11136-014-0719-3

Cape, J., Whittington, C., Buszewicz, M., & Underwood, L. (2010). Brief psychological therapies for anxiety and depression in primary care: Meta-analysis and meta-regression. *BMC Medicine*, *8*, 38. https://doi.org/10.1186/1741-7015-8-38

Capinding, A.T. (2024). Development and validation of instruments for assessing the impact of artificial intelligence on students in higher education. *International Journal of Educational Methodology*, *10*(2), 997 – 1011. https://doi.org/10.12973/ijem.10.2.997

Ceylan, G., Eken, M.O., Yuruk, S., & Emir, F. (2023). Examining the influence of self-esteem and digital literacy on professional competence factors in dental education: A cross-sectional study. *Applied Sciences*, *13*(16), 9411. https://doi.org/10.3390/app13169411

Chandra, S., Ghadei, K., Chennamadhava, M., & Ali, W. (2024). Development and validation of a farmer's focused Digital Literacy Scale. *Indian Journal of Extension Education*, *60*(1), 111-115. https://doi.org/10.48165/IJEE.2024.601RT1

Chen, F.F., Hayes, A., Carver, C.S., Laurenceau, J.P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality*, *80*(1), 219–251. https://doi.org/10.1111/j.1467-6494.2011.00739.x

Da Dalt, L., Anselmi, P., Bressan, S., Carraro, S., Baraldi, E., Robusto, E. & Perilongo, G. (2013). A short questionnaire to assess pediatric resident's competencies: the validation process. *Italian Journal of Pediatrics*, *39*(41), 1-8. https://doi.org/10.1186/1824-7288-39-41

Da Dalt, L., Anselmi, P., Furlan, S., Carraro, S., Baraldi, E., Robusto, E., & Perilongo, G. (2015). Validating a set of tools designed to assess the perceived quality of training of pediatric residency programs. *Italian Journal of Pediatrics*, *41*(2), 1-7. https://doi.org/10.1186/s13052-014-0106-2

Durak, H.Y., & Seferoğlu, S.S. (2020). Antecedents of social media usage status: Examination of predictiveness of digital literacy, academic performance, and fear of missing out variables. *Social Science Quarterly*, *101*(3), 1056-1074. https://doi.org/10.1111/ssqu.12790

Erol, S., & Aydin, E. (2021). Digital literacy status of Turkish teachers. *International Online Journal of Educational Sciences*, *13*(2), 620-633. https://doi.org/10.15345/iojes.2021.02.020

Fan, J., & Bond T. (2019). Applying Rasch measurement in language assessment: Unidimensionality and local independence. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment*, Vol. I: Fundamental techniques (pp. 83–102). Routledge. https://doi.org/10.4324/9781315187815

Frampton, G.K., & Shepherd, J. (2011). Patient-reported outcomes in clinical trials of inhaled asthma medications: Systematic review and research needs. *Quality of Life Research*, *20*, 343–357. https://doi.org/10.1007/s11136-010-9750-1

Garzon, J., & Garzon, J. (2023). Teachers' digital literacy and self-efficacy in blended learning. *International Journal of Multidisciplinary Educational Research and Innovation*. *1*(4), 162-174. https://doi.org/10.17613/cmjv-1386

Gignac, G.E., & Kretzschmar, A. (2017). Evaluating dimensional distinctness with correlated-factor models: Limitations and suggestions. *Intelligence, 62,* 138-147. https://doi.org/10.1016/j.intell.2017.04.001

Gillen, J., & Barton, D. (2010). *Digital literacies: A research briefing by the technology enhanced learning phase of the Teaching and Learning Research Programme*. London Knowledge Lab, Institute of Education.

Gilster, P. (1997). *Digital literacy*. John Wiley & Sons.

Gocen, A., & Sen, S. (2021). A validation of Servant Leadership Scale on multinational sample. *Psychological Reports, 124*(2), 752-770. https://doi.org/10.1177/0033294120957246

Gutiérrez-Ángel, N., Sánchez-García, J.-N., Mercader-Rubio, I., García-Martín, J., & Brito-Costa, S. (2022). Digital literacy in the university setting: A literature review of empirical studies between 2010 and 2021. *Frontiers in Psychology, 13,* 896800. https://doi.org/10.3389/fpsyg.2022.896800

Hagquist, C. (2019). Explaining differential item functioning focusing on the crucial role of external information–an example from the measurement of adolescent mental health. *BMC Medical Research Methodology, 19*(185), 1-9. https://doi.org/10.1186/s12874-019-0828-3

Hagquist, C., Bruce, M., & Gustavsson, J.P. (2009). Using the Rasch model in nursing research: An introduction and illustrative example. *International Journal of Nursing Studies, 46*(3), 380-393. https://doi.org/10.1016/j.ijnurstu.2008.10.007

Hamilton, C.B., & Chesworth, B.M. (2013). A Rasch-validated version of the upper extremity functional index for interval-level measurement of upper extremity function. *Physical Therapy, 93*(11), 1507–1519. https://doi.org/10.2522/ptj.20130041

Hamutoğlu, N.B., Canan Güngören, Ö., Kaya Uyanık, G., & Gür Erdoğan, D. (2017). Adapting digital literacy scale into Turkish. *Ege Journal of Education, 18*(1), 408–29. https://dergipark.org.tr/tr/download/article-file/326304

Horn, J.L., & McArdle, J.J. (1992). A practical and theoretical to measurement invariance in Aging research. *Experimental Aging Research, 18,* 117-144. https://doi.org/10.1080/03610739208253916

Hwang, H.S., Zhu, L.C., & Cui, Q. (2023). Development and validation of a Digital Literacy Scale in the artificial intelligence era for college students. *KSII Transactions on Internet and Information Systems, 17*(8), 2241-2258. https://doi.org/10.3837/tiis.2023.08.016

Lazonder, A.W., Walraven, A., Gijlers, H., & Janssen, N. (2020). Longitudinal assessment of digital literacy in children: Findings from a large Dutch single-school study. *Computers & Education, 143,* 103681. https://doi.org/10.1016/j.compedu.2019.103681

Lee, Y.W. (2004). Examining passage-related local item dependence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test. *Language Testing, 21*(1), 74–100. https://doi.org/10.1191/0265532204lt260oa

Linacre, J.M. (n.d.). *Table 23.0 Variance components for items*. https://www.winsteps.com/winman/table23_0.htm

Linacre, J.M. (2018, September 2). *Detecting multidimensionality in Rasch data using Winsteps Table 23* [Video]. Youtube. https://www.youtube.com/watch?v=sna19QemE50

Linacre, J.M. (2023). *Winsteps® Rasch measurement computer program user's guide*. Version 5.6.4. Winsteps.com

Liu, Y., Millsap, R.E., West, S.G., Tein, J.-Y., Tanaka, R., & Grimm, K.J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods, 22*(3), 486–506. https://doi.org/10.1037/met0000075

Liza, K., & Andriyanti, E. (2020). Digital literacy scale of English pre-service teachers and their perceived readiness toward the application of digital technologies. *Journal of Education and Learning, 14*(1), 74-79. https://doi.org/10.11591/edulearn.v14i1.13925

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invaria nce. *Psychometrika, 58*, 525-543. https://doi.org/10.1007/BF02294825

Miot, H.A. (2020). Analysis of ordinal data in clinical and experimental studies. *Jornal Vascular Brasileiro, 19*, e20200185. https://doi.org/10.1590/1677-5449.200185

Mundfrom, D.J., Shaw, D.G., & Ke, T.L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing, 5*(2), 159–168. https://doi.org/10.1207/s15327574ijt0502_4

Ng, W. (2012a). Can we teach digital natives digital literacy?. *Computers & Education, 59*(3), 6881065-1078. https://doi.org/10.1016/j.compedu.2012.04.016

Ng, W. (2012b). *Empowering scientific literacy through digital literacy and multiliteracies*. Nova Science Publishers, Inc.

Nguyen, L.A.T., & Habók, A. (2024). Tools for assessing teacher digital literacy: A review. *Journals of Computers in Education, 11*, 305-346). https://doi.org/10.1007/s40692-022-00257-5

Ningsih, S.K., Mulyono, H., Rahmah, R.A., & Fitriani, N.A. (2021). A Rasch-based validation of EFL teachers' received online social support scale. *Cogent Education, 8*(1), 1-13. https://doi.org/10.1080/2331186X.2021.1957529

Noorrizki, R.D., Abadi, D., Siwi, N.S.W., Sa'id, M., Mantara, A.Y., & Ramadhani, F. (2022). Factors affecting digital literacy in young adults. *Proceedings of the International Conference of Psychology 2022 (ICoPsy 2022), 308-315*, KnE Social Sciences. https://doi.org/10.18502/kss.v7i18.12396

Nunnally, J.C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Olur, B., & Ocak, G. (2021). Digital literacy self-efficacy scale: A scale development study. *African Educational Research Journal, 9*(2), 581-590. https://doi.org/10.30918/AERJ.92.21.074

Pangrazio, L., Godhe, A.-L., & Ledesma, A.G.L. (2020). What is digital literacy? A comparative review of publications across three language contexts. *E-Learning and Digital Media, 17*(6), 442-459. https://doi.org/10.1177/2042753020946291

Perera, M.U., Gardner, L., & Peiris, A. (2016). *Investigating the Interrelationship between undergraduates' digital literacy and self-regulated learning skills*. Proceedings of the 2016 International Conference on Information Systems (ICIS), Dublin, Ireland.

Reddy, P., Chaudhary, K., Sharma, B., & Chand, D. (2021). Contextualized game-based intervention for digital literacy for the Pacific Islands. *Education and Information Technologies, 26*, 5535–5562. https://doi.org/10.1007/s10639-021-10534-y

Reddy, P., Chaudhary, K., Sharma, B., & Hussein, S. (2023). Essaying the design, development and validation processes of a new digital literacy scale. *Online Information Review, 47*(2), 371-397. https://doi.org/10.1108/oir-10-2021-0532

Reddy, P., Chaudhary, K., & Hussein, S. (2023). A digital literacy model to narrow the digital literacy skills gap. *Heliyon, 9*, e14878. https://doi.org/10.1016/j.heliyon.2023.e14878

Reise, S.P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*(5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Reise, S.P., Morizot, J., & Hays, R.D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16*(1), 19–31. https://doi.org/10.1007/s11136-007-9183-7

Rodriguez, A., Reise, S.P., & Haviland, M.G. (2016a). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment, 98*(3), 223–237. https://doi.org/10.1080/00223891.2015.1089249

Rodriguez, A., Reise, S.P., & Haviland, M.G. (2016b). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137–150. https://doi.org/10.1037/met0000045

Setiyowati N., & Razak, A.Z.A. (2020). Perceived leadership styles and academicians' job performance: Teaching, research, and community services in Indonesia. *Malaysia Online Journal of Psychology and Counseling*, *7*(1), 11-26. https://jupidi.um.edu.my/index.php/MOJC/article/view/24179/11654

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman Hall/CRC.

Tor, D.L., Başaran, S.D., & Arık, E. (2022). Examining the digital Literacy level of teacher candidates. *Journal of Kırşehir Education Faculty, 23*(2), 2027-2064. https://doi.org/10.29299/kefad.1047590

UNESCO. (2018). *A global framework of reference on digital literacy skills for indicator 4.4.2*. https://uis.unesco.org/sites/default/files/documents/ip51-global-framework-reference-digital-literacy-skills-2018-en.pdf

Urbancikova, N., Manakova, N., & Ganna, B. (2017). Socio-economic and regional factors of digital literacy related to prosperity. *Quality Innovation Prosperity*, *21*(2), 124-141. https://doi.org/10.12776/qip.v21i2.942

Üstündağ, M.T., Güneş, E., & Bahçivan, E. (2017). Turkish adaptation of Digital Literacy Scale and investigating pre-service science teachers' digital literacy. *Journal of Education and Future, 12,* 19-29. https://dergipark.org.tr/tr/download/article-file/332115

Wright, B.D. (1996). Local dependency, correlations and principal components. *Rasch Measurement Transactions*, *10*(3), 509–511. https://www.rasch.org/rmt/rmt103b.htm

Wu, M., DeWitt1, D., Alias, N., & Nazry, N.N.M. (2022). *Digital literacy from 2018-2021: A scientometric study of the literature*. https://doi.org/10.21203/rs.3.rs-1848300/v1

Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8,* 125-145. https://doi.org/10.1177/014662168400800201

Zwick, R., Thayer, D.T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement*, *36*(1), 1-28. https://doi.org/10.1111/j.1745-3984.1999.tb00543.x

Zwick, R. (2012). A review of ETS Differential Item Functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series*, i-30. https://doi.org/10.1002/j.2333-8504.2012.tb02290.x

International Journal of Assessment Tools in Education
2025, Vol. 12, No. 1, 113–130

https://doi.org/10.21449/ijate.1481201

journal homepage: https://dergipark.org.tr/en/pub/ijate **_Research Article_**

Factor extraction in exploratory factor analysis for ordinal indicators: Is principal component analysis the best option?

Tugay Kaçak[1*], **Abdullah Faruk Kılıç**[1]

[1]Trakya University, Faculty of Education, Department of Educational Sciences, Edirne, Türkiye

ARTICLE HISTORY

Received: May 09, 2024

Accepted: Nov. 04, 2024

Keywords:

PCA,

Factor extraction,

Monte Carlo simulation.

Abstract: Researchers continue to choose PCA in scale development and adaptation studies because it is the default setting and overestimates measurement quality. When PCA is utilized in investigations, the explained variance and factor loadings can be exaggerated. PCA, in contrast to the models given in the literature, should be investigated in categorical/ordered, severely skewed data, and multidimensional structures. The purpose of this study is to compare the relative bias and percent correct estimation of PCA, PAF, and MINRES techniques with Monte Carlo simulations. In Monte Carlo simulations sample size, level of skewness, number of categories, average factor loadings, number of factors, level of inter-factor correlation and test length were manipulated. The results show that PCA overestimates most models with lower average factor loadings, but PAF and MINRES provide unbiased results even with low factor loadings. PAF and MINRES produce more accurate and impartial results, and it is projected that PCA will lead researchers to believe that the items in scale development or adaptation studies are of "high quality."

1. INTRODUCTION

Factor analysis is frequently used as evidence of construct validity in scale development and adaptation studies. Several studies in the literature have examined how often researchers who develop or adapt scales use Principal Component Analysis [PCA] (Ford *et al.*, 1986; Gaskin & Happell, 2014; Goretzko *et al.*, 2019; Henson & Roberts, 2006; Koyuncu & Kılıç, 2019). The result of all these review studies is that PCA is frequently used in research. Despite the popularity of PCA, it is not recommended for the factor extraction step in EFA (Fabrigar *et al.*, 1999; MacCallum & Tucker, 1991). Although there are many methods for factor extraction, the attention given to PCA in scale development and adaptation studies makes investigating its usage particularly important. One of the main focuses of this study is to explore how PCA interacts with different data characteristics and to compare it with other widely recognized and robust methods. In addition to determining the performance of methods, it is also necessary to demonstrate their implications for empirical studies. In this study, as we have anticipated, we hope that studies examining the performance of methods can provide valuable insights for method

*CONTACT: Tugay KAÇAK ✉ tugaykacak@trakya.edu.tr 🖃 Trakya University, Faculty of Education, Department of Educational Sciences, Edirne, Türkiye

e-ISSN: 2148-7456

selection in empirical research. We reviewed studies indexed in two different databases and analyzed the stages of scale development/adaptation studies, compiling the characteristics of the data. With this line, we aimed to highlight the critical steps of EFA and data characteristics, focusing on factor extraction methods in empirical studies. The study includes a systematic review followed by a series of Monte Carlo simulations designed to critique the findings derived from this review.

We first examined the use of exploratory factor analysis (EFA) in journals indexed in TRIndex (Türkiye) and the Web of Science (journals in the Q1, Q2, and Q3 quartiles of the SSCI) in terms of the methods used as factor extraction in EFAs. We have compared the studies indexed in WoS and TRIndex to provide Turkish researchers with a perspective. Then, we compared the estimation performance of PCA with minimum residual (MINRES) and principal axis factoring (PAF) in a simulation study. To determine the use of factor extraction methods in studies published in Türkiye and internationally, we examined the studies searched in TRIndex and WoS (Q1, Q2, and Q3) between 2015 and 2023. Koyuncu & Kılıç (2019) focused on the studies about social sciences were published between 2006-2016. We aimed to reveal the current usage of EFA, specifically factor extraction methods. Thus, the inclusion criterion was specified as being published between 2015 and 2023. In addition, we focused on scale development and scale adaptation studies published in the field of education.

We searched with the keywords "scale development, exploratory factor analysis, factor analysis, validity.". We used “published in journals indexed in TRIndex”, “Published in the field of “education.”, and “Published in journals (Q1, Q2 or Q3) indexed in WoS” as inclusion criteria, “Studies related to nursing, engineering, and training sciences were not included in the study to show similarities with the fields of the studies in the TRIndex.”, and “Studies in journals indexed in both TRIndex and WoS are considered in the WoS category and are not included in the TRIndex category.” as exclusion criteria.

As a result of the searches with keywords, 675 studies in journals indexed in TRIndex and 819 studies in journals indexed in WoS (Q1, Q2, and Q3) were found. For each search group, 100 studies were randomly selected and reviewed. The findings of the studies reviewed within the scope of the research are presented in Table 1.

Table 1. *Review of articles which use EFA.*

Number of Categories	TRIndex (n = 100)	WoS (n=100)	Factor Extraction Methods	TRIndex (n = 100)	WoS (n=100)
2	0%	1%	PCA	67%	39%
3	4%	4%	PAF	4%	29%
4	3%	8%	ML	3%	15%
5	89%	57%	ULS	0%	3%
6	0%	7%	MINRES	0%	2%
7	1%	18%	IMAGE	0%	2%
Not specified	2%	0%	WLSMV	1%	1%
Others	1%	5%	Not specified	25%	8%
			FIML	0%	1%
Sample Size	**TRIndex**	**WoS**	**Factor Rotation Methods**	**TRIndex**	**WoS**
0-99	1%	3%	Varimax	59%	30%
100-199	8%	20%	Promax	4%	22%
200-299	20%	23%	Direct Oblimin	6%	25%
300-399	33%	17%	Oblique(?)	1%	11%
400-499	15%	13%	Geomin	1%	3%
≥500	23%	24%	Promin	1%	1%
Mean	418	569	Equamax	1%	0%
			Not specified/ Unrotated	27%	8%

Number of Factors (p)	TRIndex	WoS	Mean of Factor Loadings	TRIndex	WoS
1	14%	8%	$0.4 \leq \lambda < 0.6$	22%	10%
2-3	30%	31%	$0.6 \leq \lambda < 0.8$	76%	85%
$p \geq 4$	56%	61%	$\lambda \geq 0.8$	0%	3%
Item(k)/Factor(p) Ratio	TRIndex	WoS	Others	2%	2%
k/p≤3:1	0%	0%	Mean	0.652	0.683
3:1 < k/p ≤ 5:1	16%	46%			
5:1< k/p ≤10:1	67%	46%			
k/p>10:1	17%	8%			
Number of Variables	TRIndex	WoS	Inter-factor Correlations	TRIndex	WoS
$k \leq 10$	3%	8%	$\varphi > \lvert 0.30 \rvert$	3%	26%
11-20	26%	35%	Including $\varphi < \lvert 0.30 \rvert$ correlations	4%	21%
21-30	36%	32%	Not reported (Oblique)	10%	22%
$k \geq 31$	35%	25%	Uncorrelated factors or unidimensional structure	83%	31%

1.1. Number of Categories

For the studies published in journals indexed in both TRIndex and WoS, it is seen that the majority of them were developed in 5-point Likert type (73%). 2 studies indexed in TRIndex determined that only Likert-type scales were used. However, no information was provided about the number of categories. Since it will also change the type of correlation matrix to be created according to the number of categories of the data, it affects the analysis processes (Holgado–Tello *et al.*, 2010).

1.2. Factor Extraction Methods: PCA vs the Others

PCA was the most frequently used factor extraction method in the studies searched in TRIndex and WoS (53%). It was determined that 25 studies in TRIndex and 8 in WoS (17%) did not report factor extraction methods. Factor extraction methods must be reported due to their assumptions and performance under various conditions (Goretzko *et al.*, 2019). There are studies in literature that compare factor extraction methods under various conditions. Although it is the most frequently used factor extraction method, studies indicate that PCA is not a factor analysis method (Fabrigar *et al.*, 1999; Harman, 1970; MacCallum & Tucker, 1991; Mulaik, 1990). In addition, Matsunaga (2010) states in his study that PCA can not be used instead of exploratory factor analysis methods because it determines the components by taking the diagonal elements in the correlation matrix with a value of 1.00 - that is, perfect reliability - without including the error variance.

In contrast to these views, studies argue that PCA is preferable (Arrindell & Van Der Ende, 1985; Costello & Osborne, 2005). Although there is no consensus on factor extraction methods, the current literature recommends using factor extraction methods that separate the error variance. Therefore, examining the performance of factor extraction methods will enlighten practical applications.

1.3. Sample Size

Most sample sizes of randomly selected and reviewed studies are between 300-399 sample size range. For the studies searched in TRIndex, the average sample size is 418, the minimum sample size is 46, the maximum sample size is 2083, and the median is 351. For the studies indexed in WoS, the average sample size is 569, the minimum sample size is 55, the maximum sample size is 9231, and the median is 314.5. It is seen that the sample size of 84% (n=168) of the reviewed studies is larger than 200, which is the minimum sample size required for EFA, as stated in Fabrigar *et al.* (1999).

1.4. Inter-factor Correlations and Factor Rotation Methods: Oblique or Orthogonal Rotation?

It is seen that most of the studies consisted of at least four dimensions (56% for TRIndex, 61% for WoS). With these findings, Varimax rotation (makes the factors as uncorrelated) is the most popular rotation method. Although a large of number of multidimensional constructs, still orthogonal rotations were preferred. This two findings conflict with the literature about the construct of interest in social sciences which commonly are correlated and multidimensional.

Total score analyses should not be performed with multidimensional scales (ntotal=25) that are multidimensional and have correlations less than |.30|. Although there is no certainty that the correlation between factors will be significant and above .30 when oblique rotation is preferred, it is theoretically possible that the factors may be unrelated after oblique rotation. In this case, it will be necessary to examine the scale structure regarding reproducibility for the studies in which oblique rotation was preferred and did not report the correlation between factors (ntotal = 33). In addition, it was found that the Varimax rotation method was preferred for one-factor structures in 2 studies in TRIndex and 1 study in WoS, and rotation methods for single-factor structures are not theoretically appropriate (Osborne, 2015). Direct Oblimin (16%) was frequently used in the studies. Thirty-four studies (17%) did not report the rotation method. Since oblique rotation methods allow for all levels of correlation between factors, it is suggested to be used for related and uncorrelated constructs (Comrey & Lee, 1992; Costello & Osborne, 2005; Nunnally & Bernstein, 1994).

1.5. Number of Variables, Factors and Items per Factor Ratio

None of the reviewed studies had a lower than 3:1 item/factor ratio recommended by Brown (2006) and Downing & Haladyna (2006) as the minimum ratio. In terms of the ideal ratio of items per factor, 5:1 (Gorsuch, 2015), 10:1 (Nunnally & Bernstein, 1994) has been suggested for EFA. In contrast to all of these, MacCallum *et al.* (2001) reject just one ratio criterion; they suggest focusing on the quality of items (factor loadings). The studies in TRIndex are mostly clustered in the 5:1 and 10:1 range, and in WoS, they are located mostly in the range of 3:1 between 5:1 and 5:1 between 10:1.

1.6. Factor Loadings

Numerous cut points about the factor loadings of variables can be found in the literature. In Hinkin's (1995) study 0.40, Costello and Osborne (2005) 0.30, Tabachnick and Fidell (2013) suggested that a loading of 0.32 would be significant. These cut points are towards the loadings of the variables on the primary factors. We evaluated the studies for average factor loadings as 0.40 low, 0.60 medium, and 0.80 high (Comrey & Lee, 1992). The average factor loadings are above 0.60 for studies in both groups. In the "Others" group, there are studies that did not publish factor loadings, reported factor loadings above a factor loading value, or published average factor loadings. Factor loadings should be reported as they provide information about the items' quality and the measurements' quality.

1.7. Current Study

PCA extracts the principal factors/components from the correlation matrix with diagonal elements of 1.00, and each extracted factor aims to explain the maximum amount of the correlation matrix that can be obtained. Since the diagonal elements do not change, this method tries to determine the entire variance for a variable. Unlike PCA, MINRES (equivalent to ULS according to Jöreskog (2003)) aims to maximally reproduce the off-diagonal elements in the correlation matrix using a least squares approach. This causes the operations performed on the correlation matrix to differ according to the methods. Therefore, the results obtained vary according to the methods. Although PCA practically takes the diagonal elements in the correlation matrix as 1, PAF differs from PCA in focusing on common variance (Mabel & Olayemi, 2020). Methods such as maximum likelihood (ML), alpha factoring, image factoring, and GLS, which follow different assumptions and procedures for factor extraction, are also available in the literature. Specifically, ML assumes multivariate normality (Garson, 2023) which is often violated by ordinal/categorical datasets (Kaplan, 2004). Fabrigar & Wegener (2012) discussed ML with ordinal/categorical datasets. Thus, it is clear that performance of ML is limitless with non-continuous datasets, and we decided to exclude ML. Watkins (2018) recommends PAF to deal with non-normal datasets. With this recommendation, PAF was the other method that we chose to analyze. Third method, MINRES, have no distributional assumptions (Jöreskog, 2003), so we decided to examine performance MINRES in this study. In sum, this study focuses on PAF, PCA, and MINRES for listed reasons. Previous studies have examined the PCA method, but its application to simulation studies typically involves continuous data sets. Therefore, in the current study, we performed analyses for the 5-point Likert type data set, a commonly used data set. Additionally, we examined binary data. Unlike other studies in literature, this study examined how biased the average factor loadings were. Therefore, it was possible to observe the practical outcomes of using PCA. Table 1 demonstrates the frequent use of PCA, despite its examination in previous studies. Therefore, this study stands out from others in literature and holds significant importance. Detailed information about other factor extraction methods will not be given. In addition, it can be said that PCA is frequently used among factor extraction methods because it overestimates factor loadings, explains total variance, and is set as default in most statistical software. Simulative studies examining the performance of the focused methods are given in Table 2.

The studies in Table 2 show that factor extraction methods have been examined under many conditions. These studies were mainly conducted with normally distributed continuous data sets. However, as accepted in educational research, the assumption that psychological characteristics are normally distributed is often not met due to the characteristics of the samples (Ho & Yu, 2015). In addition, indicators are mostly ordinal. Considering all these reasons, more work needs to be done for ordinal data with skewed distributions. Unlike earlier studies, this study focused on ordinal variables followed normal and non-normal distribution that are mostly encountered in educational and psychological structures.

Table 2. *Simulation studies in the literature*.

Studies	Factor Extraction Methods	Data Type	Distribution	Sample Size	Test Length	Number of Factors	Factor Loading / Communalities	Inter-factor correlation and rotation method
Widaman (1993)	PCA, PAF	Continuous	Normal	200	9, 18, 36, 24, 48, 96	3	0.40 0.60 0.80	None Varimax 0.50 Harris-Kaiser Orthoblique
Snook & Gorsuch (1989).	PCA, PAF	Continuous	Normal	200	9, 18, 36	3	0.40 0.60 0.80	None Varimax
De Winter & Dodou (2016).	PCA, PAF, ML	Continuous	Normal	50 5000	10, 50, 100	2,3,4,5	0.30 0.60	Varimax, Direct Quartimin, Procrustes Rotation
Coughlin (2013)	PAF, OLS, ML	Mixed (5%, 25%, 50%, 75%, 95%) Dichotomous and continuous	Normal	100, 200, 300,1000	20, 40, 60	2, 4, 8	High – 0.6, 0.7, 0.8 Wide – 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 Low - 0.2, 0.3, 0.4	Varimax
This study	MINRES, PCA, PAF	Ordinal (2 and 5)	Right-skewed, normal, left-skewed	200, 500, 1000	5 – 10 items per factor	1, 2, 3	0.40 0.70	0.00 0.30 0.60 Varimax and Promax

1.8. Importance

In this study, the performance of PCA in predicting factor loadings and inter-factor correlations is compared with MINRES and PAF. It can be said that this comparison will contribute to the literature in the following four aspects: 1) examining how accurate PCA, which is frequently used in scale development studies, gives accurate results will shed light on practitioners in practice; 2) the performance of PCA on categorical data can be examined in areas where categorical data are frequently used, 3) the performance of MINRES and PAF, which are recommended to be used on skewed data, can be examined on skewed data and their performance can be compared with PCA, 4) unlike other studies in the literature, the effects of categorical EFA on factor extraction can also be examined since it is studied with categorical data. Therefore, this study is important in providing information about the dominant use of PCA in the literature and the results obtained from the scales developed with PCA. In this direction, the study aims to investigate:

1. What is the average factor loading bias?
2. What is the percentage of correct estimation of average factor loading?

2. METHOD

A Monte Carlo simulation study examined which factor extraction method gives more accurate results for the examined models. The focus of the study was principal component analysis. Monte Carlo simulations are studies where data is produced according to a certain distribution, the produced data is analyzed with different statistical methods, and the results are compared (Sigal & Chalmers, 2016). We examined principal component analysis, principal axis, and MINRES methods.

2.1. Simulation Conditions and Data Generation

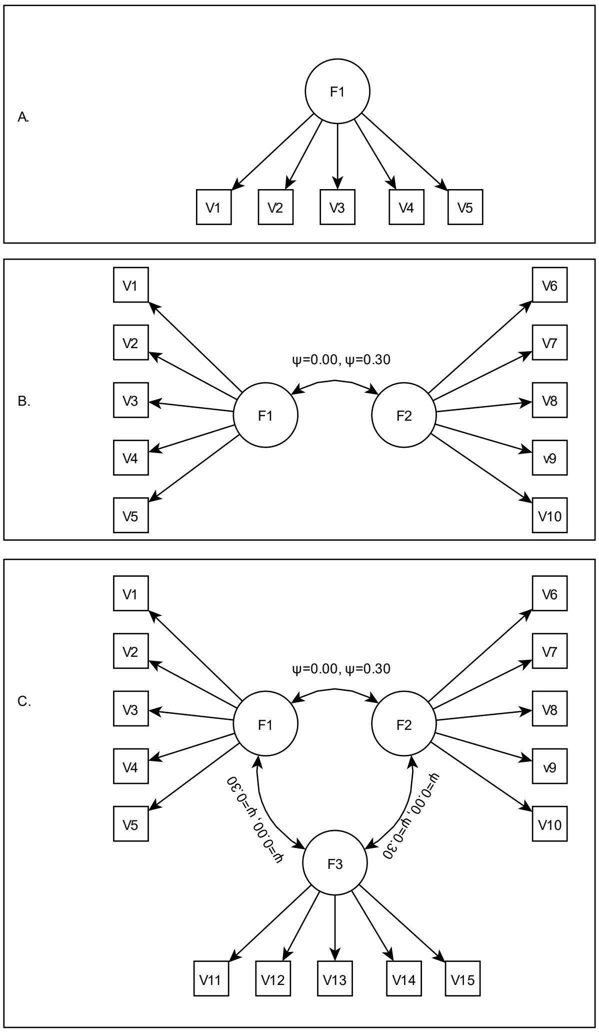
The current study examined the factor extraction method’s performance; we determined the simulation factors as the number of categories, measurement model, items per factor, average factor loadings, distribution of variables, and sample size. Table 3 presents the simulation conditions.

Table 3. *Simulation conditions.*

Simulation Factors	Simulation Conditions	The number of conditions
Measurement Models (Figure 1)	Unidimensional, 2 factors (ψ=0.00), 2 factors (ψ=0.30), 3 factors (ψ=0.00) and 3 factors (ψ=0.30)	5
The number of categories	2 and 5	2
Items per factor	5 and 10	2
Average factor loadings	0.40 and 0.70	2
Distribution of variables	Left skewed, normal, and right skewed	3
Sample size	200, 500, and 1000	3
	Total	2x5x2x2x3x3=360 with 1000 replication

In the review study conducted by Koyuncu & Kılıç (2019), it was reported that more than half (55.5%) of the scale development studies in the field of social sciences had 1-3 dimensions. Therefore, the number of dimensions in the current study was determined to be 1, 2, and 3. Inter-factor correlations were determined to examine factor extraction methods' performance in unrelated and moderately related constructs. The number of categories of variables was manipulated as 2 and 5. Variables with two categories can be in achievement tests or checklists. Otherwise, measurement results in different fields, such as health, can also be categorical. For this reason, a 2-category condition was added to the study. On the other hand, since the most frequently used category number in Likert-type scales is 5 (Lozano *et al.*, 2008), it was added to the study.

Figure 1. *Measurement models.*



We manipulated items per factor as 5 and 10 items. Since it is known that there should be at least three items in a factor for a factor to be formed (Brown, 2006), the minimum number of items was considered to be 5 in this study. Considering the 3-dimensional structures, when items per factor are 10, the upper limit of the number of items is set as 10 since 30-item measurement tools will be formed. The average factor loading was manipulated to be 0.40 and 0.70. Although there are different suggestions for the minimum factor loading to be obtained as a result of EFA, it can be said that it will generally be around 0.30 (Costello & Osborne, 2005; Howard, 2016; Tabachnick & Fidell, 2013). Therefore, in this study, the average factor loading condition was determined to be 0.40. At the same time, it can be said that the factor loadings of the items will be higher in stronger scales. For this reason, the condition of 0.70 was added to the study to include scales with better items.

The skewness of the variables is also an issue that needs to be studied. Costello & Osborne (2005) states that the PAF method can be used when the variables are skewed. On the other hand, Zygmont and Smith (2014) stated that the MINRES method can be used in skewed variables. Therefore, in order to evaluate the performance of these methods on skewed data, both left and right-skewed conditions were added to the study. Normal conditions were also added to the study to evaluate the changes that may occur in the performance of the methods under conditions with normal distribution. The skewness coefficient of the variables was set as -2.5 for left-skewed variables and 2.5 for right-skewed variables. According to the literature of skewness, ±2 skewness can be justified as an acceptable limit for normality (Hair, 2014). Thus, ±2.5 skewness may be justified as non-normal, or skewed distributions.

At last, the sample size was manipulated to be 200, 500, and 1000. In studies in the literature, these sample sizes are often considered small, medium, and large. They are also frequently used in simulation studies (Beauducel & Herzberg, 2006; Kılıç & Doğan, 2021; Li, 2016; Oranje, 2003; West *et al.*, 1995) For this reason, sample sizes were handled in this way in this study.

2.2. Evaluation Criteria

We used relative bias (RB) and percent correct (PC) as evaluation criteria in the study. Relative bias is calculated as follows;

$$Relative\ Bias = \frac{\bar{\psi} - \psi_{True}}{\psi_{True}} \quad 1$$

Where $\bar{\psi}$ is the average of the estimated factor loadings across 1000 replications, and ψ_{True} is the true average factor loading. |RB| > 0.10 means substantial bias (Flora & Curran, 2004; Forero *et al.*, 2009; Rhemtulla *et al.*, 2012).

We calculate the ±5% of the true factor loadings for percent correct. Then, for 1000 replications, we examined what proportion of the average factor loadings estimated by the models fell between this range (±5%). We used 90% PC value as “acceptable” in this study. (Collins *et al.*, 2001).

2.3. Data Analysis

We used a uniform distribution to determine factor loadings. First, we determined the factor loadings for the population, yielding an average factor loading of 0.40 and 0.70. Second, we generated continuous data followed by a multivariate normal distribution. Lastly, we categorized the dataset using predetermined thresholds. We used thresholds in Appendix 1.

We used the “lavaan” package (Rosseel, 2012) to generate data. The generated data sets were analyzed with the "psych" package (Revelle, 2024). Polychoric correlation matrices were used in the analysis. In multidimensional structures, Promax was used in conditions with an inter-factor correlation of 0.30, and Varimax was used in conditions with an inter-factor correlation of 0.00.

3. RESULTS

3.1. Relative Bias of Factor Loadings

One-way ANOVA was conducted to determine the simulation conditions influencing the RB values. ANOVA results indicated that all of the simulation conditions have an effect on the RB values. Partial eta squares of each simulation condition are represented in Table 4.

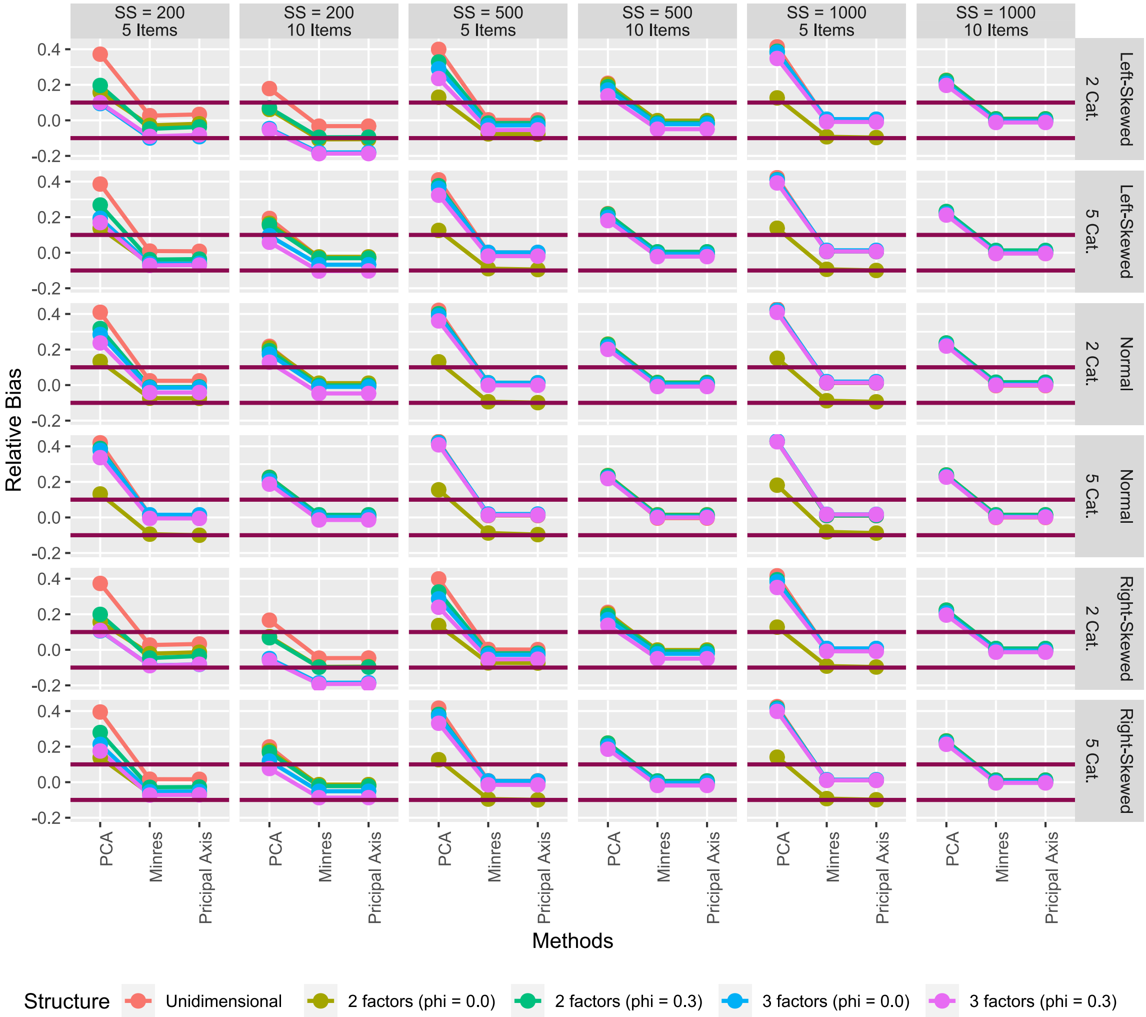
Table 4. *The ANOVA results for each simulation factors on RB values*.

Simulation Factors	The ANOVA Results
Measurement models	[F(4, 1066) = 11.61, p < 0.01, η^2 = 0.04]
The number of categories	[F(1, 1066) = 11.86, p <0.01, η^2 =0.01]
Items per factor	[F(1, 1066) = 41.38, p < 0.01, η^2 = 0.04]
Average factor loadings	[F(1, 1066) = 122.86, p < 0.01, η^2 = 0.10]
Distribution of variables	[F(2, 1066) = 11.99, p < 0.01, η^2 =0.02]
Sample size	[F(2, 1066) = 30.19, p < 0.01, η^2 = 0.05]
Factor extraction method	[F(2, 1066) = 803.29, p < 0.01, η^2 = 0.60]

All simulation factors and factor extraction method have statistically significant effect on RB values. Measurement models [$F(4, 1066) = 11.61$, $p < 0.01$, $\eta^2 = 0.04$], the number of categories of variables [$F(1, 1066) = 11.86$, $p <0.01$, $\eta^2 =0.01$], items per factor [$F(1, 1066) = 41.38$, $p < 0.01$, $\eta^2 = 0.04$], average factor loadings [$F(1, 1066) = 122.86$, $p < 0.01$, $\eta^2 = 0.10$], distribution of variables [$F(2, 1066) = 11.99$, $p < 0.01$, $\eta^2 =0.02$], sample size [$F(2, 1066) = 30.19$, $p < 0.01$, $\eta^2 = 0.05$], and factor extraction method [F(2, 1066) = 803.29, $p < 0.01$, $\eta^2 = 0.60$]. The mean and median of RB of the methods for all conditions were 0.16 and 0.10 for PCA, -0.01 and 0.00 for MINRES, and -0.01 and 0.00 for PAF, respectively.

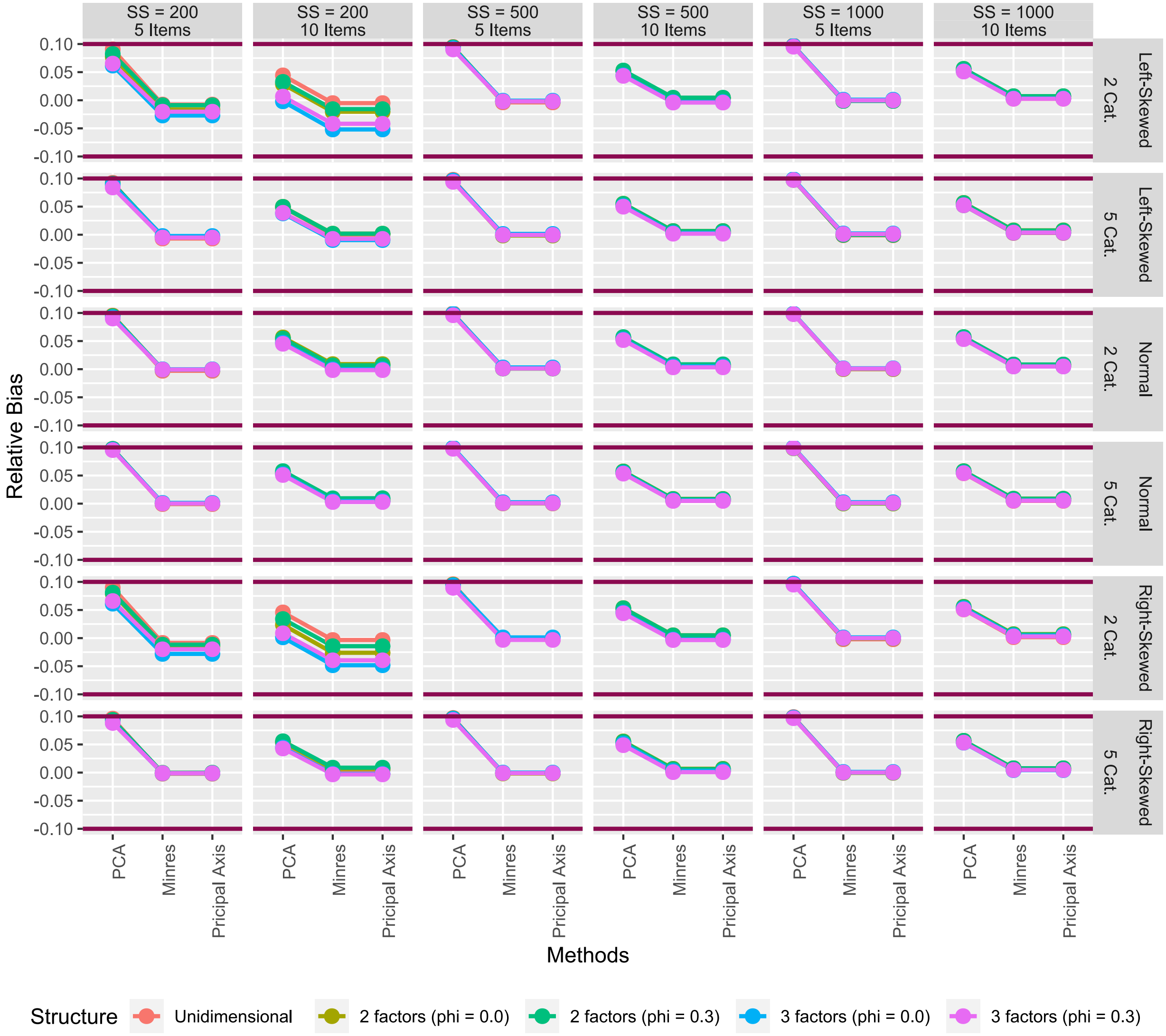
The graphs of the RB values are presented in Figures 2 and 3 with average factor loadings of 0.40 and 0.70, respectively. The relative bias values are within the acceptable range for PCA, MINRES and PA in all conditions where the average factor loading is 0.70 (see Figure 3).

Figure 2. *RB values of methods for average factor loading is 0.40.*



PCA is biased within the acceptable range in 10 (5.6%) conditions with an average factor loading of 0.40. When these ten conditions are analyzed, it can be said that the distribution is skewed, the sample is small, the item per factor is high, the structure is 3-dimensional, and the number of categories is low. In these conditions, while MINRES and PAF are under factoring, PCA is biased within the acceptable range. In other words, it can be said that there is a structure suitable for the general pattern. PCA overestimated the factor loading in all other conditions except for these conditions. MINRES underestimated in 7 (3.9%) conditions where the average factor loading was 0.40. These conditions were observed in cases where the number of items was high, variables were skewed, multifactor structures, and dichotomous variables. PAF underestimated factor loadings in 9 (5%) conditions where the average factor loading was 0.40.

Figure 3. *RB values of methods for average factor loading is 0.70.*



3.2. Percent Correct of Factor Loadings

One-way ANOVA was conducted to determine the simulation conditions influencing the PC values. ANOVA results indicated that all of the simulation conditions have an effect on the PC values. The mean of PC values statistically significantly differed from each other in terms of the number of categories of variables [$F(1, 1066) = 6.23$, p <0.05, η^2 =0.01], items per factor [$F(1, 1066) = 197.410$, $p < 0.01$, $\eta^2 = 0.16$], average factor loading [$F(1, 1066) = 967.06$, $p < 0.01$, $\eta^2 = 0.48$], structure [$F(4, 1066) = 8.27$, $p < 0.01$, $\eta^2 = 0.03$], distribution of variables [$F(2, 1066) = 20.98$, $p < 0.01$, η^2 =0.04], sample size [$F(2, 1066) = 96.79$, $p < 0.01$, $\eta^2 = 0.15$], and method [$F(2, 1066) = 1597.47$, $p < 0.01$, $\eta^2 = 0.75$]. The mean and median of the methods in terms of the PC values for all conditions were 14.70% and 2.9% for PCA, 72.82% and 80.45% for MINRES, and 72.79% and 80.45% for PAF, respectively. The graphs are presented in Figures 4 and 5 with average factor loadings of 0.40 and 0.70, respectively.

PC values are lower than 90% for most conditions of average factor loading, which is 0.40. In 15 conditions (8.33%), MINRES and PAF have PC values above 90%, while in the other conditions where the average factor loading is 0.40, they have PC values below 90%. When the 15 conditions with adequate performance are examined, it can be said that the sample is mostly 1000, the variable distribution is normal, the number of items is 10, the number of categories is five, and the number of dimensions is 3. In the same conditions, PCA could not reach 90%.

PC values increased for MINRES and PAF in conditions where the average factor loading was 0.70. PCA has PC values above 90% only in two conditions. These conditions were observed in data sets with a sample size of 200, 2 categories, 10 items, three factors, and skewed distribution. In conditions where the average factor loading was 0.70, MINRES and PAF failed to perform adequately in 51 conditions (28.33%). When these conditions were examined, it was observed that the sample was small, the variables were skewed, and they were in multidimensional structures. The number of items and categories does not affect the performance of the methods.

Figure 4. *PC values of methods for average factor loading is 0.40.*

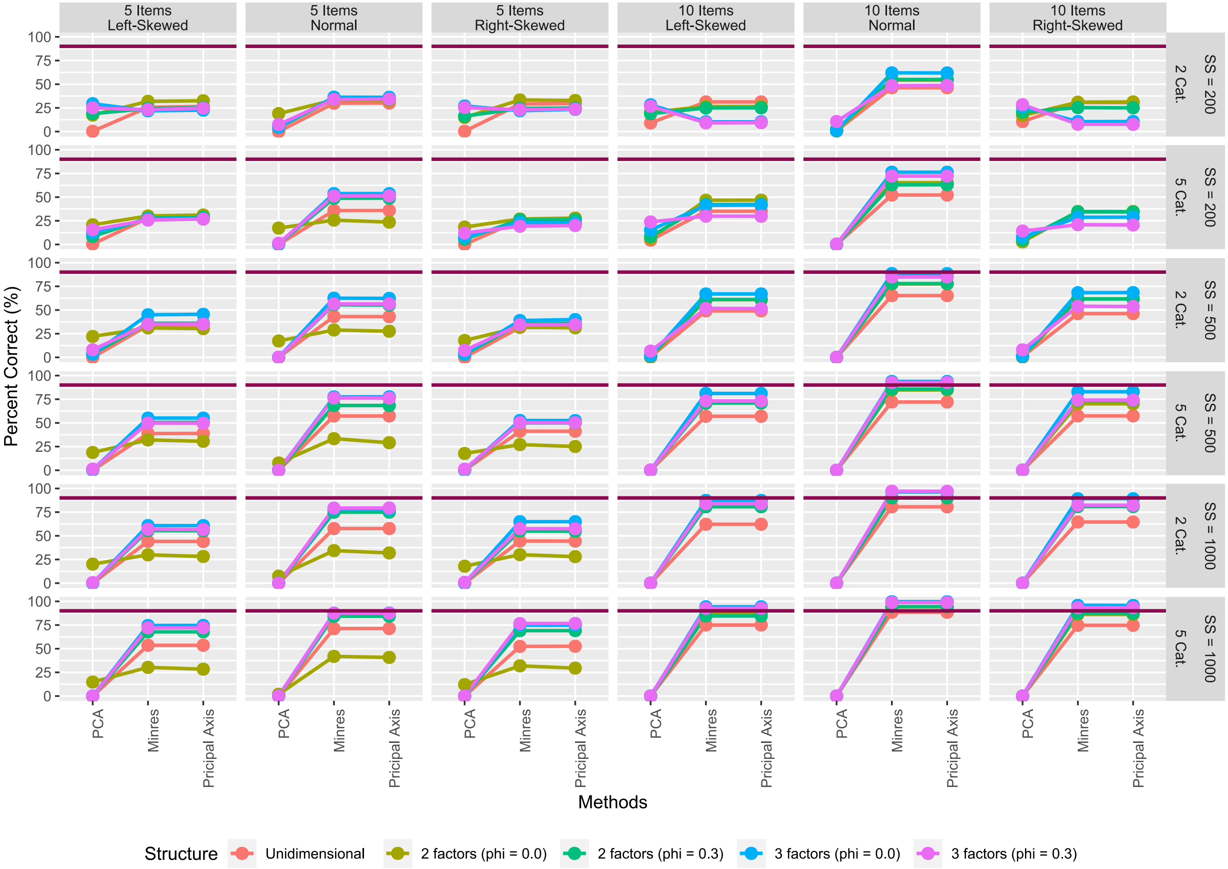
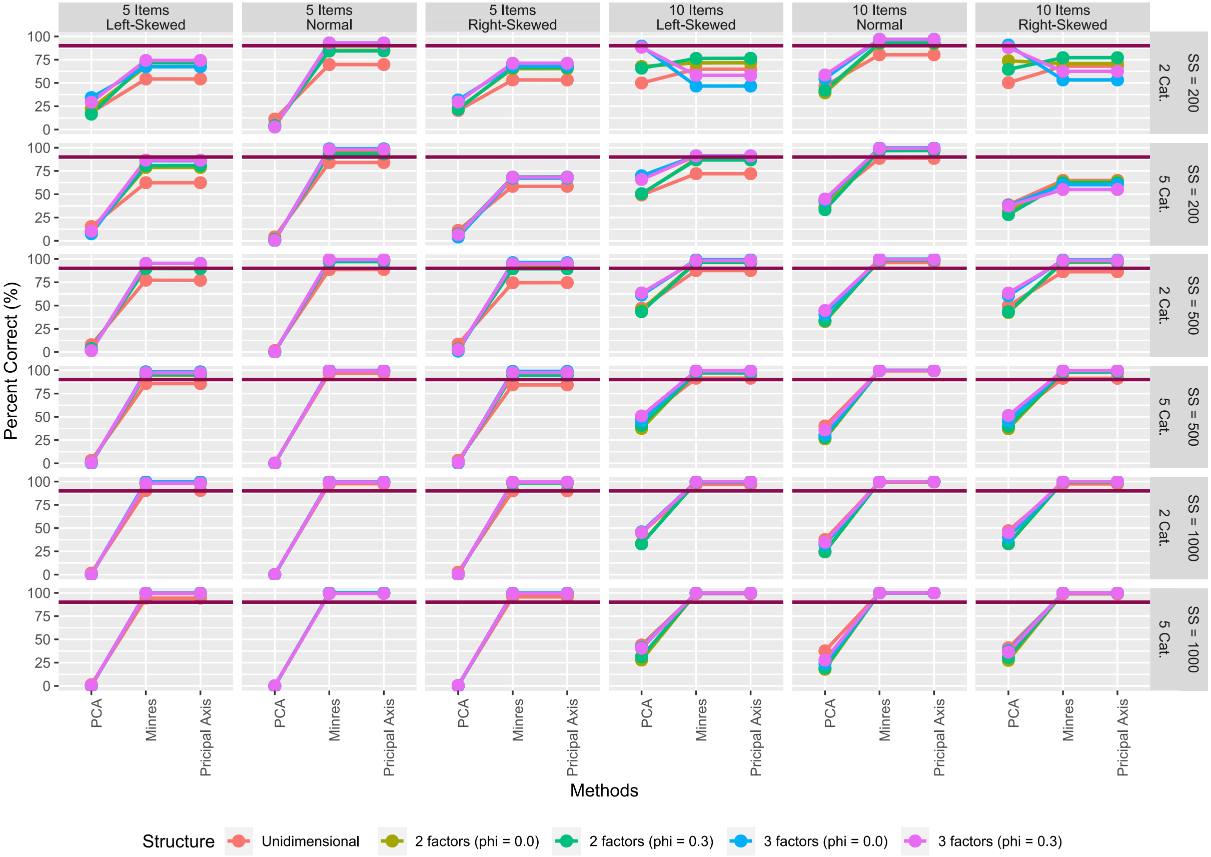


Figure 5. *PC values of methods for average factor loading is 0.70.*



4. DISCUSSION and CONCLUSION

We compared PCA, MINRES, and PAF methods regarding bias and percent correct values. As a result of the study, all methods gave unbiased results in conditions with high average factor loading. However, when evaluated in terms of percent correct, PCA performed adequately in none of the conditions, and the other methods performed adequately in about 30%. Although this result may seem contradictory, it is related to the calculation of the RB and PC values. For example, in the simulation condition where the average factor loading is 0.40, the average factor loading is estimated to be 0.43 in all 1000 replications. In this case, the average of 1000 replications will be obtained as 0.43. The RB value will be calculated as 0.08 ($\frac{0.43-0.40}{0.40} = 0.075$) and will be considered biased within the acceptable range. However, in terms of PC, since not all 1000 replicates are in the range of 0.38-0.42 (the average factor loading is 0.43 in all replicates), the PC value will be 0. This indicates that even if unbiased estimates are made, there are inaccurate estimates in terms of accuracy. Therefore, a decision can be made by evaluating the methods in terms of both bias and accuracy.

The methods were overestimated in almost all PCA conditions when the average factor loading was low. This result is consistent with De Winter and Dodou (2016). However, MINRES and PAF gave unbiased results in almost all conditions where the average factor loading was low, and underestimation was observed in a small part of the conditions. According to this, the fact that the variables are skewed, the number of categories or the measured construct is multidimensional does not affect the performance of MINRES and PAF much. Since there is an underestimate in the already biased results, it can be considered as the lower limit of the factor loadings obtained when MINRES or PAF is used in EFA. For this reason, it can be said that similar quality results will be obtained in similar samples due to its use in scale development studies.

According to studies that compare PCA and PAF like Snook and Gorsuch (1989) and Widaman (1993), PAF outperformed PCA in most of the conditions, especially for shorter tests. PCA overestimated loadings across all factors. Differences between estimated and population loadings have decreased if loadings get higher. Our study is consistent with Snook and Gorsuch (1989) and Widaman's (1993) study with this line. We found that if the factor loadings get higher, RB values get lower.

From this point of view, the preference of PCA in scale development studies, especially in cases where the average factor loading is low, may cause the scale to appear of higher quality than it is. When this situation affects reproducibility, it may cause the scale to give different results from the results in the development study, even if it is used in similar samples. For this reason, attention should be paid to whether PCA is used in scale development studies, and the results should be evaluated with this sensitivity.

When the results are evaluated in terms of PC values, in simulation conditions with low factor loadings, PCA did not perform adequately in any condition. In contrast, MINRES and PAF performed adequately in approximately 10% of the conditions. In conditions with low factor loadings, skewness of distribution, the number of categories, and items per factor affect the performance of these methods. However, PC values can be considered a more conservative statistic since they show what percentage of all replications are within ±5% of the actual factor loading.

When the simulation conditions affecting the RB and PC values are analyzed, it is observed that the method used (η^2 = 0.60) and average factor loading (η^2 = 0.10) affect the RB values more. It can be said that items per factor (η^2 = 0.16), average factor loadings (η^2 = 0.48), sample size (η^2 = 0.15), and methods (η^2 = 0.75) are effective on PC values. The mean and median of RB and PC values for all conditions are similar (~80%).

4.1. Recommendations

As a result of this study, researchers who develop or adapt scales may be advised not to use PCA when using EFA as a factor extraction method. If they use PCA, factor loadings should be taken into consideration, as they are mostly overestimated. In the current literature review, PCA is still the most commonly used factor extraction method (see Table 1). However, it should be considered that the factor loadings obtained from these scales are overestimated. Researchers who will use the developed scales should consider this when selecting scales. It can be said that the reported factor loadings can be considered as the upper limit of the actual factor loadings. In addition, this situation will create problems in terms of both reliability and reproducibility. Therefore, for skewed, two- or five-category data, it may be recommended that practitioners use the MINRES or PAF method regardless of the number of dimensions and correlations between dimensions.

Researchers may conduct similar simulation studies on variables with different numbers of categories or mixed-format data. In future studies, comparing PCA with other methods in terms of inter-factor correlation may be considered.

4.2. Limitations

In this study, smoothing was performed when calculating the tetrachoric correlation matrix, especially in the case of skewed distribution of two-category data sets. Therefore, the results obtained should be evaluated within the framework of smoothing bias. However, considering that smoothing will also be required in real data sets with skewed distributions, it can be said that the results will be similar to the real situations. In addition, in this study, categorical data was handled with only 2 and 5 categories. It should be taken into account that in real situations, different numbers of categories (such as 3, 4, 6, or 7) may be encountered. It would not be appropriate to generalize these results to all categorical data. In addition, in the simulation study, thresholds were used to categorize the variables. This causes each variable to have a different skewness coefficient. Although the average skewness coefficient is ±2.5, it should be taken into consideration that not all variables have this value but have values close to it. The study, the k/p ratio is considered as the items per factor ratio, 5/1 and 10/1. There is a need for studies with higher items per factor.

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.

Contribution of Authors

Tugay Kaçak: Investigation, Systematic review, Writing-original draft, and Formal analysis. **Abdullah Faruk Kılıç:** Methodology, Supervision, Coding for simulations, and Writing-original draft.

Orcid

Tugay Kaçak https://orcid.org/0000-0002-5319-7148
Abdullah Faruk Kılıç https://orcid.org/0000-0003-3129-1763

REFERENCES

Beauducel, A., & Herzberg, P.Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(2), 186-203. https://doi.org/10.1207/s15328007sem1302_2

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

Collins, L.M., Schafer, J.L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351. https://doi.org/10.1037/1082-989X.6.4.330

Comrey, A.L., & Lee, H.B. (1992). *A first course in factor analysis* (2nd ed.). L. Erlbaum Associates.

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*(1), 1–9. https://doi.org/10.7275/JYJ1-4868

Coughlin, K. (2013). An analysis of factor extraction strategies: A comparison of the relative strengths of principal axis, ordinary least squares, and maximum likelihood in research contexts that include both categorical and continuous variables. https://www.semanticscholar.org/paper/7bbfc6050a24dce49454a56d9af2ad2e6fc40ad2

De Winter, J.C.F., & Dodou, D. (2016). Common factor analysis versus principal component analysis: A comparison of loadings by means of simulations. *Communications in Statistics - Simulation and Computation*, *45*(1), 299-321. https://doi.org/10.1080/03610918.2013.862274

Fabrigar, L.R., & Wegener, D.T. (2012). *Exploratory factor analysis*. Oxford University Press. https://doi.org/10.1093/acprof:osobl/9780199734177.001.0001

Fabrigar, L.R., Wegener, D.T., Maccallum, R., & Strahan, E. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Flora, D.B., & Curran, P.J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466–491. https://doi.org/10.1037/1082-989X.9.4.466

Forero, C.G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 625-641. https://doi.org/10.1080/10705510903203573

Garson, G.D. (2023). *Factor analysis and dimension reduction in R: A social scientist's toolkit*. Routledge, Taylor & Francis Group.

Ho, A.D., & Yu, C.C. (2015). Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. *Educational and Psychological Measurement*, *75*(3), 365–388. https://doi.org/10.1177/0013164414548576

Howard, M.C. (2016). A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interaction, 32*, 51-62. https://doi.org/10.1080/10447318.2015.1087664

Jöreskog, K.G. (2003). *Factor analysis by MINRES*. https://www.ssicentral.com/wp-content/uploads/2021/04/lis_minres.pdf

Kaplan, D. (Ed.). (2004). *The SAGE handbook of quantitative methodology for the social sciences*. SAGE.

Kılıç, A.F., & Doğan, N. (2021). Comparison of confirmatory factor analysis estimation methods on mixed-format data. *International Journal of Assessment Tools in Education, 8*(1), 21–37. https://doi.org/10.21449/ijate.782351

Koyuncu, İ., & Kılıç, A.F. (2019). The use of exploratory and confirmatory factor analyses: A document analysis. *Eğitim ve Bilim, 44*(198), 361-388. https://doi.org/10.15390/EB.2019.7665

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

Lozano, L.M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology, 4*(2), 73–79. https://doi.org/10.1027/1614-2241.4.2.73

Mabel, O.A., & Olayemi, O.S. (2020). A comparison of principal component analysis, maximum likelihood and the principal axis in factor analysis. *American Journal of Mathematics and Statistics, 2*(10), 44–54.

MacCallum, R.C., & Tucker, L.R. (1991). Representing sources of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin, 109*(3), Article 3. https://doi.org/10.1037/0033-2909.109.3.502

Oranje, A. (2003, April 21). Comparison of estimation methods in factor analysis with categorized variables: Applications to NEAP data [Paper presentation]. *Annual Meeting of the National Council on Measurement in Education*.

Revelle, W. (2024). *psych: Procedures for psychological, psychometric, and personality research* (Version 2.4.3) [Computer software]. Northwestern University. https://cran.r-project.org/package=psych

Rhemtulla, M., Brosseau-Liard, P.É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354–373. https://doi.org/10.1037/a0029315

Sigal, M.J., & Chalmers, R.P. (2016). Play it again: Teaching statistics with Monte Carlo simulation. *Journal of Statistics Education, 24*(3), 136-156. https://doi.org/10.1080/10691898.2016.1246953

Snook, S., & Gorsuch, R. (1989). Component analysis versus common factor analysis: A Monte Carlo study. *Psychological Bulletin, 106*, 148-154. https://doi.org/10.1037/0033-2909.106.1.148

Tabachnick, B.G., & Fidell, L.S. (2013). *Using multivariate statistics* (6th ed., international ed.). Pearson.

Watkins, M.W. (2018). Exploratory factor analysis: A guide to best practice. *Journal of Black Psychology, 44*(3), 219–246. https://doi.org/10.1177/0095798418771807

West, S.G., Finch, J.F., & Curran, P.J. (1995). Structural equation models with non-normal variables: Problems and remedies. In R. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56–75). Sage.

Widaman, K. (1993). Common factor analysis versus principal component analysis: Differential bias in representing model parameters? *Multivariate Behavioral Research, 28*(3), 263–311. https://doi.org/10.1207/s15327906mbr2803_1

Zygmont, C., & Smith,M.R. (2014). Robust factor analysis in the presence of normality violations, missing data, and outliers: Empirical questions and possible solutions. *The Quantitative Methods for Psychology*, *10*(1), 40-55. https://doi.org/10.20982/tqmp.10.1.p040

APPENDIX

Appendix 1. *Thresholds.*

Categories	Right Skewed (S.C. = 2.5)	Normally Distributed	Left Skewed (S.C. = -2.5)
2	$Y = \begin{cases} 0, y_i \leq 1.178 \\ 1, y_i > 1.178 \end{cases}$	$Y = \begin{cases} 0, y_i \leq 0.00 \\ 1, y_i > 0.00 \end{cases}$	$Y = \begin{cases} 0, y_i \leq -1.178 \\ 1, y_i > -1.178 \end{cases}$
5	$Y = \begin{cases} 0, & y_i \leq 1 \\ 1, & 1 < y_i \leq 1.189 \\ 2, & 1.189 < y_i \leq 1.5 \\ 3, & 1.5 < y_i \leq 2.1 \\ 4, & y_i > 2.1 \end{cases}$	$Y = \begin{cases} 0, & y_i \leq 1{,}5 \\ 1, & -1{,}5 < y_i \leq -0{,}5 \\ 2, & -0.5 < y_i \leq 0.5 \\ 3, & 0{,}5 < y_i \leq 1{,}5 \\ 4, & y_i > 15 \end{cases}$	$Y = \begin{cases} 0, & y_i \leq -2 \\ 1, & -2 < y_i \leq -1.7 \\ 2, & -1.7 < y_i \leq -1.2 \\ 3, & -1.2 < y_i \leq -0.99 \\ 4, & y_i > -0.99 \end{cases}$

International Journal of Assessment Tools in Education
2025, Vol. 12, No. 1, 131–146

https://doi.org/10.21449/ijate.1468887

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

Research Article

Digital leadership in educational organizations: A scale adaptation study

Mehmet Emin Ören[1*], **Servet Atik**[2]

[1]Abdülkerim Kuzu Special Education Kindergarten, Siirt, Türkiye
[2]İnönü University, Faculty of Education, Department of Educational Sciences, Malatya, Türkiye

ARTICLE HISTORY

Received: Apr. 15, 2024

Accepted: Sep. 17, 2024

Keywords:

Digital leadership,
Leadership,
Scale adaptation.

Abstract: In this study, it was aimed to adapt the DigiFuehr 2.0 Scale developed by Claassen *et al.* (2023) to Turkish and to conduct validity and reliability studies on three groups of participants consisting of teachers. In the study, exploratory and confirmatory factor analyses were performed in line with translation study, linguistic application, and validity and reliability studies. The findings indicate that the scale is a valid and reliable assessment tool for Turkish education leaders. In particular, the dimensions of support and self-organization play an important role in evaluating the digital leadership skills of leaders. In addition, this scale provides a powerful tool for evaluating and developing the digital leadership skills of educational leaders. Therefore, it will allow a more in-depth examination of the effects of digital leadership skills in studies to be carried out in educational organizations.

1. INTRODUCTION

Leadership, a phenomenon that has attracted attention throughout human history, is an important concept in the context of the growth, development and struggle for the survival of organizations. Looking at leadership from a broad perspective, Yukl (2009) pointed out that the individual motivation, abilities, and power relations of the group members affect the perception of leadership, as well as interpreting leadership as the group's reactions to internal and external influences. In the twenty-first century, a dramatic change was observed in the relations between school principals and teachers within the framework of leadership and management. According to Tanniru and Peral (2021), the main reason for this change is the political, social, economic and technological changes that have occurred in educational organizations in the twenty-first century and have significantly affected these organizations. These changes have led to a change in school management and understanding of leadership. According to Figus (2021), new technologies that can dynamically change and transform society and schools have paved the way for the transformation of educational policies and the understanding of individuals, groups and leaders in the organization.

When we consider educational organizations, it is important for students, teachers and all organizational employees to gain personality and individualize, to keep up with change and

*CONTACT: Mehmet Emin ÖREN ✉ mehmeteminoren@gmail.com Abdülkerim Kuzu Special Education Kindergarten, Siirt, Türkiye

e-ISSN: 2148-7456

transformation, to understand the modern world and to keep up with the modern world. In other words, technology and digitalization have ceased to be a choice or choice for 21st century managers, learners and teachers and have become a necessity (Ceylan, 2019). According to Zhong (2017), while school structures that try to keep up with the transformation started to progress physically, they also tried to get rid of the classical school understanding by improving their technological infrastructures. During this time, schools managed in line with the bureaucratic structure have led scientists to research and develop alternative forms of management in order to keep up with digital transformation and catch up with technology (Richardson *et al.*, 2012). Thus, changing living conditions in the globalizing world have begun to develop school management styles and leadership styles. Individuals who can keep up with change in school organizations, provide all kinds of technological development for the school and have the knowledge and experience to both develop and support themselves and all employees of the organization exhibit an impressive leadership example (Antonopoulou *et al.*, 2020; Tanniru *et al.*, 2018).

In the new century, organizations need to integrate common knowledge, experience, judgment, values and beliefs and transform at the group and organization level with digital leaders in order to benefit from the information stacks according to the needs of the age (Rooney & McKenna, 2007). Based on this context, it can be said that it is not possible for educational organizations not to be affected by digital education technologies. However, it can be stated that administrators, teachers and students use technological devices effectively in their lives outside of school. Therefore, it is impossible to keep educational organizations away from digital media and tools. For this reason, digital tools are expected to be actively used in other processes of management as well as teaching activities.

The rapid changes of the digital age are radically transforming educational organizations and taking the understanding of leadership to a new dimension. Traditional education models, together with the rise of digital technologies, affect learning processes and corporate governance. In this context, the concept of digital leadership in educational organizations has evolved into a broad perspective that not only emphasizes technological skills, but also includes features such as managing change, encouraging innovation, and strengthening the learning culture (Arham *et al.*, 2023). In short, it is an inevitable reality that digital tools are used more and more effectively every day in educational organizations. In summary, in order for educational organizations to reach organizational wisdom, it is expected that leaders who will succeed in using digital technologies for the benefit of the organization will play an important role in addition to phenomena such as information management and digital leadership.

Digital leadership in educational organizations has a critical role in managing modern learning environments. Ridho *et al.* (2023) express this leadership style as the ability to effectively integrate digital technologies, increase student success, and reshape educational processes. However, it can be said that digital leaders have a vision to use technology strategically in educational institutions. However, digital leaders enrich the learning experience by interacting with students, teachers, and even all stakeholders through advanced learning and teaching systems, online platforms, and other digital tools (Yusof *et al.*, 2019). According to Highton (2022), these leaders also provide support to teachers in developing digital skills. It encourages innovative practices in education and leads teachers to use the potential provided by technology more effectively.

It can also be said that school administrators with digital leadership characteristics are skilled in obtaining scientific outputs in educational organizations. Thus, digital leaders can use data analysis, output evaluation and monitoring processes to achieve the studies conducted in schools and the targeted level of success (Karaköse *et al.*, 2021). The outputs obtained in line with these processes are used for the analysis of data on student achievement, the improvement of teaching processes and the creation of individualized learning strategies. As digital leaders communicate effectively, they can enable all stakeholders of the organization to participate more in decision-making processes and gain self-confidence (Tigre *et al.*, 2023). Thus, it is assumed that digital leaders in education can also increase student-parent communication. Based on this assumption, digital communication tools allow parents to provide instant information about their students' progress and success, which more effectively engages families in the learning processes.

Despite these benefits of digital transformation and leadership in educational organizations, it can be observed that there are some negative situations encountered. Lack of technological infrastructure in educational organizations, inadequate access to technology; school administrators' lack of general digital skills or feeling inadequate in digital skills may create obstacles to implementing digital leadership and effectively transferring digital skills to teachers and students (Sousa *et al.*, 2017). Similarly, insufficient training and support for teachers to perform digital leadership tasks may cause a lack of skills in this regard. In addition, the commitment of administrators and teachers to traditional teaching methods may cause them to resist adopting technology (Keleş *et al.*, 2020). This may prevent the effective fulfillment of digital leadership roles. In addition, the general financial problems of schools can create difficulties in investing in new technological solutions and providing financial support for the digital development of educational staff. In our world of rapid technological transformation, the inability to integrate new technological tools and applications into the learning environment may reduce the motivation of all stakeholders of educational organizations. In this context, it is thought that it is important for teachers to be able to evaluate their basic digital skills, communication and cooperation skills, adaptation skills to change, and innovation and creativity capacities. In addition, it can be stated that the fact that teachers have a say in the determination of school management and education policies and have the opportunity to work with leaders who can support digital learning, change and transformation is a critical point not only for educational organizations but also for the digital development of society. In this context, it is hoped that the outputs of the digital leadership scale, which has been adapted, will be a valuable tool for school leaders to make strategic decisions for teachers to understand and develop their digital skills. In this way, it is thought that educational institutions can be directed to a more effective digital transformation process and offer stronger digital learning experiences to all stakeholders.

In this study, it was aimed to adapt the DigiFuehr 2.0 Scale, originally developed by Claassen *et al.* (2023) to evaluate the digital leadership level of individuals and their managers, into Turkish and to conduct validity and reliability studies of the Turkish form on a group of teachers. The DigiFuehr 2.0 Scale allows for the analysis of not only individual leadership skills but also both horizontal and vertical leadership approaches, providing a broader understanding of digital leadership culture (Claassen *et al.*, 2023). Such scales assess the roles and competencies of individual leaders in digital transformation processes while also measuring how leaders participate in collaboration and decision-making processes within the organization (Petry, 2018). These features are particularly important for educational institutions, as digital leadership is not limited to the use of technology by administrators; it also encompasses the development of digital skills among teachers, students, and other stakeholders (Highton, 2022).

One of the reasons for adapting the DigiFuehr 2.0 scale to the Turkish context is the increasing digitalization initiatives in Turkey's education system in recent years. In particular, the digital education infrastructures accelerated by the pandemic have made it imperative for teachers and administrators to develop their digital leadership skills (Karaköse *et al.*, 2021). Measuring the digital leadership levels of educational administrators will provide a strategic perspective on Turkey's digital transformation processes in education. Therefore, the DigiFuehr 2.0 scale is a suitable and effective tool for assessing the digital competencies of educational leaders in Turkey.

Developed by Claassen *et al.* (2023), the DigiFuehr 2.0 scale evaluates the contributions of not only a single administrator but also all members of the organization to the role of digital leadership, emphasizing both horizontal and vertical leadership aspects of digital leadership culture. Particularly with the widespread digitalization of educational institutions, such an assessment tool allows for a comprehensive evaluation of leaders' digital skills (Rooney & McKenna, 2007). In this respect, it is expected that the adaptation of the scale will make a significant scientific contribution, particularly for education systems like Turkey's, which are in the process of digital transformation.

2. METHOD

2.1. Research Method

The Digital Leadership Scale (DigiFuehr 2.0) was developed by Claassen *et al.* (2023) to evaluate his and his manager's level of digital leadership. Before starting the studies, the authors who developed the scale were asked for permission to adapt the scale. In the process of adapting the scale to Turkish, (1) a translation study and (2) a validity and reliability study were conducted. The Digital Leadership Scale was applied to the workstation employees of municipalities in Germany when it was first developed. In contrast, the scale was adapted by applying a different sampling (teachers). Confirmatory factor analysis and exploratory factor analysis methods were used during the validity study of the scale. The exploratory factor analysis method was used because the researcher did not have an idea about the factor structures during the development or adaptation of the scale, and the scale adaptation needed scientific evidence (Finch & West, 1997). SPSS Statistics 22 and Amos 24 software were used for the study.

2.2. Translation Study

During the translation study, the support of expert linguists was obtained. The items used in the scale were translated into Turkish by a lecturer who speaks Turkish and English well, two faculty members and three doctoral students. Then, translation options were evaluated by a faculty member and four experts from the field of educational administration and different translations were decided. The decided scale items were examined by three Turkish language and literature experts in terms of meaning and fiction integrity and Turkish spelling check before the pilot application. The recommendations given by these experts were applied on the scale. Later, the back translation process of the scale items translated into Turkish was carried out by two associate professors and a doctor faculty member, who were different from the experts who made the first translation process and had a command of both languages. Finally, the scale, which was translated into English, was compared by two experts and the differences that may occur were resolved. The last edited Digital Leadership Scale was applied to a group of 50 educators. As a result of the pilot study, it was understood that the scale items were understandable and clear to the participants.

2.3. Participant Groups

During the adaptation of the Digital Leadership Scale to Turkish, three separate sample groups consisting of teachers working actively in public, and private schools and institutions participated in the study.

2.3.1. *First group*

This group was the one from which data on Exploratory Factor Analysis (EFA) were obtained. This group consisted of 310 participants including 181 female (58.4%), 129 male (41.6%), 248 undergraduate (80%), and 62 (20%) graduate-doctorate teachers.

2.3.2. *Second group*

This second group was the one from which data on Confirmatory Factor Analysis (CFA) were obtained. This group consisted of 183 participants including 60 female (32.8%), 123 male (67.2%), 131 undergraduate (71.6%), and 52 (28.4%) graduate-doctorate teachers.

2.3.3. *Third group*

It is the group where the data related to the Test-Retest were obtained. In this group, it was aimed to test the consistency of the scale against time. The scale was applied to this group twice with an interval of 15 days. This group consisted of 63 participants including 24 female (38.1%), 39 male (61.9%), 44 undergraduate (69.8%), and 19 (30.2%) graduate-doctorate teachers.

2.4. Data Collection Tool

Digital Leadership Scale: It is a scale originally called DigiFuehr 2.0 developed by Claassen *et al.* (2023), which aims to measure the digital leadership culture at the team level, including horizontal leadership, instead of evaluating the competence of a single leader or non-leader. The concept of digital leadership is defined as a process of development and transformation (Petry, 2018). Therefore, in addition to digital demands and support for managers, the assumption that the entire organization has responsibilities in this regard and that every employee can be considered as a digital leader over time is a prevailing opinion (Ahlemann *et al.*, 2021). DigiFuehr 2.0, developed in this context, is a four-point Likert-type scale consisting of nine items and two sub-dimensions. The support sub-dimension consists of items (2, 3, 4, 5, 6, 7) that measure how much the individual is encouraged and supported for his/her digital development. The self-organization sub-dimension consists of the items expressing the participation of the person in the intra-organizational decisions related to him/herself (item 1), the ability of the person to make his/her own decisions within the organization (item 8) and his/her involvement in the intra-organizational coordination (item 9). The internal consistency of the scale was found to be α=0.88 throughout the scale. In the adaptation study, the scale was adapted as a five-point Likert type and it was determined that the highest score obtained from the scale would be 45 and the lowest score would be 9. A high score indicates a high level of digital leadership skills and perception, while a low score indicates a low level of digital leadership skills and perception. The finalization of the Turkish version of the scale for implementation (see the Turkish version of the DLS in the Appendix).

2.5. Data Analysis

In the study, SPSS for Windows 22.0 and Amos 21.0 package software were used for statistical analysis of the data obtained from the scale. For the internal consistency of the scale, the Cronbach Alpha coefficient was calculated for both sample groups in general and separately for each dimension. For the content validity of the scale, the opinions of experts in the field were consulted, while exploratory factor analysis was applied to the first sample group for construct validity, and confirmatory factor analysis was applied to the second sample group. Kaiser-Meyer-Olkin (KMO) and Bartlett (Bartlett's Test of Sphericity) tests were performed before the Exploratory Factor Analysis was performed. As a result of the KMO (= .923) and Bartlett (= 2338.354, p = .000) tests, exploratory factor analysis was deemed appropriate because the KMO value higher than .60 showed that the data were suitable for factor analysis (Büyüköztürk, 2014).

While conducting Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA), the normality assumption of the data set was based on the assumptions that skewness and kurtosis values should be between +1 and -1 and Z scores should be between +3 and -3 (Çokluk *et al.*, 2012). As a result of the analyses, it was found that the data sets in both EFA and CFA studies were normally distributed. The reliability of the scale was tested using Cronbach's alpha and composite reliability (CR) over the data collected for EFA. An α value above 0.7 is considered acceptable (Büyüköztürk, 2014). The discrimination power of the items was examined by comparing the upper 27% and lower 27% of the data with the corrected item-total correlations (Can, 2018). SPSS 22.0 software was used for composite reliability, Cronbach's alpha, construct validity and item analysis of the Digital Leadership Scale.

Lastly, measurement invariance was examined using multi-group confirmatory factor analysis (MG-CFA) (Cheung & Lau, 2012; Horn & McArdle, 1992). Measurement invariance provides information on the psychometric equivalence of a construct across groups or over time (Putnick & Bornstein, 2016). In this study, measurement invariance was tested at the configural, followed by the metric, and finally the scalar levels of measurement invariance (Cheung & Rensvold, 2002; Vandenberg & Lance, 2000).

3. RESULTS

As a result of the exploratory factor analysis, the support dimension explains 64.24%of the total variance and the self-organization dimension explains 13.28% of the total variance. Cronbach Alpha internal consistency coefficients of the Digital Leadership Scale were calculated as .957 in the support dimension, .738 in the self-organization dimension, and .929 in the entire scale. In light of these data, it can be said that the self-organization dimension is reliable, and the support dimension and the entire scale are highly reliable (Yang & Green, 2009). The results obtained as a result of the exploratory factor analysis are shown in Table 1.

Table 1. *Exploratory factor analysis results.*

		Factor I	Factor II	Factor	Corrected
Dimension	Articles	Support	Self-Organization	Common Variance	Item-Test Correlation
Support	2. My school principal supports me to improve my digital literacy.	.854		.817	.851
	3. When I have problems with digitalization, I get support from my school principal.	.857		.781	.805
	4. I regularly receive feedback from my school principal on the quality of my digital work.	.865		.800	.823
	5. My school principal supports me in accessing the information I need to do my digital work.	.878		.850	.867
	6. My school principal supports me in understanding and using digital applications better.	.896		.859	.858
	7. My principal promotes digital ways of working at school.	.863		.835	.861
Self-Organization	1. I am involved in decisions that affect my work and digital work environment.		.631	.529	.546
	8. I can determine what working methods, processes, and solution approaches I will use to accomplish my tasks.		.794	.718	.569
	9. I perform my duties in cooperation with my colleagues.		.881	.788	.443
	Eigenvalue	5.782	1.196		
	Total Variance Explained (77.525)	*64.241*	*13.284*		
	Cronbach's Alpha	*.957*	*.738*		
	Cronbach Alpha (for the full scale)	*.929*			

Cronbach's alpha coefficients for each dimension of the scale are .96 for the support dimension (6 items), .74 for the self-organization dimension (3 items) and .93 for the total scale. The CR coefficients for the support and self-organization dimensions are .94, .82 and .86 for the total scale. In general, reliability coefficients of .70 and above are presented as evidence that measurement tools can be accepted as reliable (Fraenkel *et al.*, 2012). Composite Reliability (CR) and AVE were used to test the convergent validity of the scale. All CR values related to the scale are expected to be greater than the AVE values and the AVE value is expected to be greater than 0.5 (Hair *et al.*, 2019). In this respect, the AVE values of the scale were found to be at an acceptable level.

Cronbach's alpha (α) and composite reliability (CR), AVE coefficients for each dimension of Dijital Ledaership Scale were used to assess the reliability of the Digital Leadership Turkish version. The results are shown in Table 2.

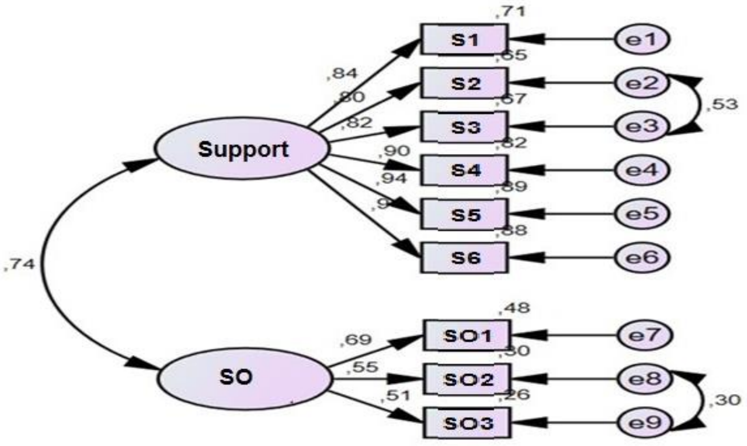
Table 2. *Digital Ledaership Scale Cronbach's alpha (α), CR and AVE values.*

Dimension	Cronbach's α	Composite Reliability (CR)	AVE
Support (6 item)	.96	.94	.76
Self-Organization (3 item)	.74	.82	.60

After the exploratory factor analysis results, data were collected from the second sampling group (*n*=183). According to Fidell and Tabachnick (2003), criteria such as missing data, univariate and multivariate normality, linearity, outlier observations, and multiple linear connection problems should be examined in the data set before applying statistical analysis. As a result of the analysis, it was determined that there was no lost or missing data. The number of data collected for confirmatory factor analysis was determined as 183. In order to perform the confirmatory factor analysis, it is stated that 100-200 people are required according to Boomsma (1985), and the sample size should be larger than 100 according to Anderson and Gerbing (1984). Some researchers state that the sample size needed depends on the number of items. According to Cohen and Cohen (1983), a minimum of 10 participants are recommended for each item, and according to Stevens (2002), between 5-20 participants are recommended for each item on the scale. In light of this information, it can be said that the sample size of the research is sufficient for analysis. However, skewness and kurtosis values were calculated for each item, Variance Increase Factor (VIF), Tolerance (T), and Status Index (CI) values were examined and whether there was normality and multicollinearity in the data set was tested. In line with the findings obtained, it was observed that CI values were less than 30, VIF values were less than 10, and T values were different from zero, so multicollinearity assumption was provided (Black and Babin, 2019) and confirmatory factor analysis was performed.

While performing confirmatory factor analysis, many fit data were checked. According to these data, it can be said that CFI, NNFI and GFI values mean a perfect fit greater than .95, good fit between .95 and .90 (Bentler & Bonett, 1980); RMSEA, RMR and SRMR values below .05 mean a good fit level, acceptable level up to .08 (Browne & Cudeck, 1992), AGFI value is a good fit if a value greater than ".95", values greater than ".85" mean acceptable fit (Yılmaz & Çelik, 2009). Firstly, model fit statistics were determined without any limitation in the model created. According to the analyzes, χ^2 = 114.29, *sd* = 260, χ^2/sd = .439 (*p* = .00) , RMSEA = .137, NNFI = .91, CFI = .93, GFI = .87, AGFI = .78, RMR = .07 and SRMR = .048. In the light of these data, the modification indices were examined in order for the model to fit better, the items S2 and S3 and SO2 and SO3 were reviewed, it was determined that these items were meaningfully close to each other and measured similar properties, and necessary arrangements were made. As a result of repeated analysis, new data; χ^2 = 51.59, *sd* = 24, χ^2/sd = 2.15 (*p* = .01) , RMSEA = .07, NNFI = .96, CFI = .98, GFI = .94, AGFI = .88, RMR = .04 and SRMR = .032. In this way, it can be said that the data fit the model better. Cronbach Alpha internal consistency coefficients of the Digital Leadership Scale were calculated as .957 in the support dimension, .738 in the self-organization dimension, and .929 in the entire scale. In light of these data, it can be said that the self-organization dimension is reliable, and the support dimension and the entire scale are highly reliable (Yang & Green, 2009). The analysis diagram showing the data obtained for the confirmatory factor analysis is shown in Figure 1 together with the standard coefficients.

Figure 1. *Confirmatory factor analysis.*



The scale, which reached its final form as a result of the analyzes, was applied to a group of 63 teachers every 15 days. As a result of the application of the Digital Leadership Scale to the same sample group of teachers at 15-day intervals, the correlation between the sub-dimensions of the scale and the scores obtained from the scale total was obtained as .918 in the self-organization dimension, .852 in the support dimension, and .887 in the scale total. In light of these data, it can be said that the test-retest reliability of the scale is high. The data regarding the test-retest application of the Digital Leadership Scale are given in Table 3.

Table 3. *Digital Leadership Scale test-retest application.*

		2. Application (Cronbach Alpha= .923)		
		Self-Organization	Support	Scale Total
1. Application (Cronbach Alpha= .915)	Self-Organization	.918		
	Support		.852	
	Scale Total			.887

Looking at the item analysis results, when 27 was taken as the cut-off value (lower and upper groups), the results showed that the t-values for the difference between the upper 27% and lower 27% of the participants ranged between 3.56 and 7.46 for the self-organization dimension and between 4.24 and 7.14 for the support dimension. The t-test values are significant for all items according to the comparison between the lower 27% and the upper 27% of the participants. Significant t-values in the comparisons between the lower and upper groups of the participants were accepted as evidence of the discriminative power of the items (Büyüköztürk, 2014). Table 4 also shows that item-total correlations ranged between .44 and .87. When the results obtained are evaluated together, it is concluded that each item of the Digital Leadership Scale is discriminative.

The item analyses of the scale were conducted with Item-Test Correlation methods and Sub-Upper Group Analysis techniques (Büyüköztürk, 2014). A t-test was used to determine whether there was a significant difference between the upper 27% and the lower 27% of the Turkish version of the Digital Leadership Scale. The results of the item analysis of the scale are given in Table 4.

Table 4. *Item analysis results of Digital Leadership Scale*.

Dimension	Items	Corrected Item Total Correlations (r)	*Upper (%27)* $\bar{X}$	Lower (%27) $\bar{X}$	Lower-Upper 27% *t*-Test	*p*
Support	2. My school principal supports me in developing my digital literacy.	.851	2.28	1.58	4.24	.00
	3. Whenever I have problems with digitalization, I get support from my school principal.	.805	2.38	1.42	5.31	.00
	4. I regularly receive feedback from my school principal about the quality of my digital work.	.823	2.42	1.38	5.76	.00
	5. My school principal supports me in accessing the information I need to do my digital work.	.867	2.59	1.66	6.30	.00
	6. My school principal supports me in understanding and using digital applications better.	.858	2.40	1.52	5.85	.00
	7. My school principal encourages digital working methods at school.	.861	2.85	1.47	7.14	.00
Self-Organization	1. I am involved in decisions that affect my job and my digital work environment.	.546	3.25	2.45	3.83	.00
	8. I can determine which work methods, processes and solution approaches I will use to accomplish my tasks.	.569	2.67	1.67	7.46	.00
	9. I fulfill my duties in cooperation with my colleagues.	.443	3.88	3.02	3.56	.00

Before conducting the measurement invariance analyses, the model fit indices of the original factor structure of the Digital Leadership Scale by gender and subject area are presented in Table 5. When Table 5 is examined, it is evident that the fit indices of the measurement model of the Digital Leadership Scale for the gender and subject area variables fall within the widely accepted ranges used to evaluate model fit in the literature. In this context, the two-factor structure of the Digital Leadership Scale demonstrates compatibility with the data obtained from all subgroups. In other words, the original factor structure has been confirmed for each subgroup, providing evidence that construct validity is established within each subgroup.

Table 5. *Fit indices of the subgroups for the Digital Leadership Scale.*

Groups		χ^2	*sd*	RMSEA (%90 CI)	SRMR	CFI	TLI
Gender	Female	111.55	251	.068 (.065 - .071)	.040	.96	.95
	Male	114.32	251	.069 (.065 - .072)	.038	.95	.94
Branch	Primary School Teacher	135.68	2240	.059 (.039 - .078)	.038	.97	94
	Specialist Teacher	158.95	224	.079 (.060 - .098)	.051	.95	.92

Note: df = degrees of freedom, and the 90% confidence intervals for the RMSEA values are provided in parentheses.

Whether the Digital Leadership Scale holds measurement invariance across gender and subject area variables was examined using multi-group confirmatory factor analyses. In this context, configural, metric, and scalar invariance models were tested for each variable. The findings related to measurement invariance are presented in Table 6.

Table 6. *Multi-group CFA results for the Digital Leadership Scale.*

Variable		χ^2	*sd*	RMSEA	CFI	SRMR	$\Delta\chi^2$	Δ*sd*	*p*	ΔCFI	ΔRMSEA	ΔSRMR
Gender	Formal	264	98	.071	.964	.036						
	Metric	274.73	106	.067	.959	.043	15.23	9	.08	.003	.002	.008
	Scalar		117	.066	.958	.045	14.59	9	.13	.002	.002	.002
Branch	Formal	267.41	98	.071	.957	.037						
	Metric	276.95	106	.066	.954	.042	8.48	9	.48	.002	.001	.002
	Scalar	290.98	117	.066	.954	.043	13.95	9	.13	.000	.000	.001

When Table 6 is examined, it can be said that the fit indices used to evaluate model fit for the configural invariance stage are within acceptable limits for all groups (RMSEA < .08, CFI > .90, NFI > .90, NNFI > .90, IFI > .90). Since the factor loadings, inter-factor correlations, and error variances of the model are freely estimated across subgroups in the configural invariance stage, it can be stated that the structure of the measurement model for the Digital Leadership Scale is the same across gender and subject area subgroups. As configural invariance was achieved, the next stage, metric invariance, was tested.

In the metric invariance stage, the factor loadings were constrained to be equal across subgroups. The fit indices obtained were examined, and the model was found to exhibit good fit with the data. To test metric invariance, the differences between the CFI and RMSEA values from the configural and metric invariance stages were examined, and it was observed that the ΔCFI and ΔRMSEA values were within acceptable limits for metric invariance ($\Delta\text{CFI} \leq .01$; $\Delta\text{RMSEA} \leq .015$). This finding indicates that the factor loadings of the variables in the model did not change across gender and subject area subgroups. After metric invariance was established, the final stage of scalar invariance was tested by constraining the factor structures, factor loadings, and item intercepts to be equal across groups.

The fit indices for scalar invariance indicated that the model exhibited adequate fit. Scalar and metric invariance models were compared, and it was determined that the obtained values remained within the criteria recommended by Chen (2007). The findings from the model comparisons demonstrate that the Digital Leadership Scale achieved configural, metric, and scalar measurement invariance across both male and female teachers, as well as between primary school teachers and subject teachers.

4. DISCUSSION and CONCLUSION

Digital leadership in education is a critical factor for the sustainable success of educational institutions in today's rapidly changing technological environment. Digital leadership in education contributes to the professional development of teachers and school administrators, playing a key role in adapting them to the rapidly changing digital age. Today, technological advances have profound effects on the success of educational institutions, and at this point, digital leadership guides teachers, students, and parents in using digital tools effectively (Levin & Schrum, 2013; Robiah & Nurdin, 2021; Sheninger, 2019). Studies emphasize that teachers' digital leadership skills play an important role in supporting technology integration in the classroom, increasing student motivation, and strengthening teaching strategies (Levin & Schrum, 2013; Robiah & Nurdin, 2021; Sheninger, 2019).

In this context, research tools such as the Digital Leadership Scale are an important resource for evaluating and developing the digital leadership skills of teachers and school administrators (AlAjmi, 2022). Digital leadership provides a framework that guides school administrators, teachers, and other stakeholders in the process of integrating educational technologies (Sheninger, 2019). This leadership approach allows teachers to create student-centered learning environments in their classrooms and use digital tools effectively (Ertmer *et al.*, 2006).

With this important role in education, digital leadership helps students develop their digital skills and prepare them for the future digital world (Bersin, 2018). Therefore, it is emphasized by many studies that digital leadership enriches the interaction and learning experience in education by supporting the professional development of teachers and school administrators (Levin & Schrum, 2013; Robiah & Nurdin, 2021; Sheninger, 2019). For this reason, the Digital Leadership Scale, which can be used for educational organizations, has been adapted because it is worth further examination in terms of the role of digital leadership in education, student success, teaching strategies and its impact on technology integration.

This study focused on evaluating the validity and reliability of the digital leadership scale on Turkish education leaders. The findings show that the scale is compatible with the Turkish education system and can be a reliable tool for evaluating digital leadership skills of leaders. The high factor loads obtained in the support dimension of the study emphasize the effective role of leaders in increasing the level of digital literacy and improving the quality of digital studies (AlAjmi, 2022). In addition, in the research on the role of digital leadership in education in the literature, Arham *et al.* (2022) found that digital leadership has a positive effect on teacher and student success. From this point of view, it is thought that future studies that examine the effect of leader behaviors in the support dimension on student success in more detail will contribute to educational organizations.

The capacity of leaders to effectively support students and teachers in digitalization can positively affect the digital transformation process in education (Hakansson *et al.*, 2019). AlAjmi (2022) emphasizes that self-organization skills are important for the effective management of digital learning environments. Similarly, studies by Cvetković *et al.* (2023) examining the effects of digital leadership on increasing student achievement show that this scale can guide educational leaders in developing digital leadership skills. In the light of this information, it is seen that the items in the self-organization dimension evaluate the participation of leaders in digital decisions and the self-regulation skills required to effectively fulfill their duties. This emphasizes the importance of leaders focusing on their personal and professional development in order to fulfill their digital leadership roles more effectively (Gierlich-Joas *et al.*, 2020).

In summary, the study of adapting the Digital Leadership Scale to Turkish confirmed the validity and reliability of the scale on Turkish education leaders. This supports that the digital leadership scale can be used as an effective assessment tool for educational leaders. The findings indicate that the internal consistency and factor structure of the scale are strong. The high factor loads of the items in the support dimension emphasize the effective role of leaders in improving digital literacy and increasing the quality of digital studies. Items in the self-organization dimension, on the other hand, revealed that leaders can measure the self-regulation abilities necessary for them to be included in digital decisions that affect their work and to perform their duties effectively.

As a result, it is thought that the adaptation of the digital leadership scale developed in Germany to Turkish can be an effective tool in evaluating the digital leadership skills of educational leaders. Therefore, it is hoped that the adapted digital leadership scale, which is a powerful tool that can be used to evaluate and develop the digital leadership skills of educational leaders, will make significant contributions to the literature. Future research may allow us to better understand and develop the digital leadership skills of educational leaders, as the effects of the scale on leaders at different educational levels are examined in more depth and the individual demonstrates both his/her self-organization skills, digital leadership, and the digital leadership development of his/her manager.

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**: İnönü University, E.383601.

Contribution of Authors

The authors contributed equally to all the stages of the study.

Orcid

Mehmet Emin Ören https://orcid.org/0000-0002-2227-7145
Servet Atik https://orcid.org/0000-0003-2841-6182

REFERENCES

Ahlemann, F., Schütte, R., & Stieglitz, S. (2021). *Innovation Through Information Systems*. Springer International Publishing.

AlAjmi, M.K. (2022). The impact of digital leadership on teachers' technology integration during the COVID-19 pandemic in Kuwait. *International Journal of Educational Research, 112*, 101928. https://doi.org/10.1016/j.ijer.2022.101928

Anderson, J.C., & Gerbing, D.W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49*(2), 155-173. https://link.springer.com/article/10.1007/bf02294170

Antonopoulou, H., Halkiopoulos, C., Barlou, O., & Beligiannis, G.N. (2020). Leadership types and digital leadership in higher education: Behavioural data analysis from University of Patras in Greece. *International Journal of Learning, Teaching and Educational Research, 19*(4), 110-129. https://doi.org/10.26803/ijlter.19.4.8

Arham, A.F., Norizan, N.S., Arham, A.F., Hasbullah, N.N., Malan, I.N.B., & Alwi, S. (2022). Initializing the need for digital leadership: A meta-analysis review on leadership styles in educational sector. *Journal of Positive School Psychology*, *6*(8), 2755-2773. https://journalppw.com/index.php/jpsp/article/view/10280/6661

Balcı, A. (2000). İkibinli yıllarda Türk milli eğitim sisteminin örgütlenmesi [Organization of the Turkish national education system in the twentieth century]. *Kuram ve Uygulamada Eğitim Yönetimi, 24*(24), 495-508. https://dergipark.org.tr/en/download/article-file/108514

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

Bersin, J. (2018). The rise of the social enterprise: A new paradigm for business. Forbes April, 3, 2018.

Black, W., & Babin, B.J. (2019). Multivariate data analysis: Its approach, evolution, and impact. *In The great facilitator: Reflections on the contributions of Joseph F. Hair, Jr. to marketing and business research* (pp. 121-130). Cham: Springer International Publishing.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika*, *50*(2), 229-242.

Browne, M.W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological methods & research*, *21*(2), 230-258. https://doi.org/10.1177/0049124192021002005

Büyüköztürk, Ş. (2014). *Sosyal bilimler için veri analizi el kitabı* [*Data analysis handbook for social sciences*]. Pegema Yayıncılık.

Can, A. (2018). *SPSS ile bilimsel araştırma sürecinde nicel veri analizi* [*Quantitative Data Analysis in Scientific Research Process with SPSS*]. Pegem Akademi.

Ceylan, M. (2019). *21. yüzyıl becerileri bağlamında okul yöneticilerinin değişen rollerinin öğretmen görüşlerine göre incelenmesi* [*Examining the changing roles of the school principles in the context of the 21st century skills from teachers' point of view*] [Unpublished master's thesis]. Trakya Üniversitesi.

Chen, F.F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*(3), 464-504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Lau, R. S. (2012). A direct comparison approach for testing measurement invariance. *Organizational Research Methods*, *15*(2), 167-198. https://doi.org/10.1177/1094428111421987

Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5

Claassen, K., Dos Anjos, D.R., Kettschau, J.P., Wrede, S.J.S., & Broding, H.C. (2023). DigiFuehr 2.0: Novel insights for digital leadership. *Journal of Occupational Health*, *65*(1), e12383. https://doi.org/10.1002/1348-9585.12383

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. L. NJ Eribaum.

Cvetković, B.N., Stošić, A.S., & Mitić, I.T. (2023). *Leadership in education in the digital age*. Facta Universitatis, Series: Teaching, Learning and Teacher Education, 189-199. https://doi.org/10.22190/FUTLTE221115019N

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik* [*Multivariate statistics for social sciences*]. Pegem Akademi.

Ertmer, P.A., Ottenbreit-Leftwich, A., & York, C.S. (2006). Exemplary technology-using teachers: Perceptions of factors influencing success. *Journal of computing in teacher education*, *23*(2), 55-61. https://doi.org/10.1080/10402454.2006.10784561

Fidell, L.S., & Tabachnick, B.G. (2003). Preparatory data analysis. *Handbook of Psychology: Research Methods in Psychology*, *2*, 115-141. https://doi.org/10.1002/0471264385.wei0205

Figus, A. (2021). Information Society and Digital Leadership in the Globalized Educational System: Political Approach. *European Proceedings of Social and Behavioural Sciences*. https://doi.org/10.15405/epsbs.2021.07.02.3

Finch, J.F., & West, S.G. (1997). The investigation of personality structure: Statistical models. *Journal of Research in Personality*. *31*(4). 439-485. https://doi.org/10.1006/jrpe.1997.2194

Fraenkel, J.R., Wallen, N.E., & Hyun, H.H. (2012). *How to design and evaluate research in education* (8th ed.). McGraw-Hill.

Gierlich-Joas, M., Hess, T., & Neuburger, R. (2020). More self-organization, more control-or even both? Inverse transparency as a digital leadership concept. *Business Research, 13*, 921-947. https://doi.org/10.1007/s40685-020-00130-0

Hair, J., Black, W., Babin, B., & Anderson, R. (2019). *Multivariate data analysis* (8th ed.). Cengage Learning.

Håkansson, L.M., & Pettersson, F. (2019). Digitalization and school leadership: on the complexity of leading for digitalization in school. *The International Journal of Information and Learning Technology*, *36*(3), 218-230. https://doi.org/10.1108/IJILT-11-2018-0126

Highton, M. (2022). The importance of diversity and digital leadership in education: a feminist perspective from higher education. *Handbook of Digital Higher Education*, 351-362. https://doi.org/10.4337/9781800888494.00039

Karakose, T., Polat, H., & Papadakis, S. (2021). Examining teachers' perspectives on school principals' digital leadership roles and technology capabilities during the COVID-19 pandemic. *Sustainability*, *13*(23), 13448. https://doi.org/10.3390/su132313448

Keleş, H.N., Atay, D., & Karanfil, F. (2020). Instructional leadership behaviors of school principals during the COVID 19 pandemic process. *Milli Egitim*, *49*(1), 155-174. https://doi.org/10.37669/milliegitim.787255

Levin, B.B., & Schrum, L. (2013). Using systems thinking to leverage technology for school improvement: Lessons learned from award-winning secondary schools/districts. *Journal of Research on Technology in Education*, *46*(1), 29-51. https://doi.org/10.1080/15391523.2013.10782612

Petry, T. (2018). *Knowledge management in digital change*. In: North K, Maier R, Haas O, eds. Digital leadership. Springer.

Putnick, D.L., & Bornstein, M.H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, 41, 71-90. https://doi.org/10.1016/j.dr.2016.06.004

Richardson, J.W., Bathon, J., Flora, K.L., & Lewis, W.D. (2012). NETS• A scholarship: A review of published literature. *Journal of Research on Technology in Education*, *45*(2), 131-151. https://doi.org/10.1080/15391523.2012.10782600

Ridho, M.R., Lesmana, I., Safitri, H.D.A., Meirani, R.K., & Prestiadi, D. (2023, February). Digital Leadership in the Scope of Education. In *International Conference on Educational Management and Technology (ICEMT 2022)* (pp. 52-61). Atlantis Press.

Robiah, P.S., & Nurdin, D. (2021). Implementation of Digital Leadership in developing student learning at SMP Manggala Kab. Bandung. *In Proceeding of International Conference on Research of Educational Administration and Management (ICREAM)* (Vol. 5, No. 1, pp. 23-27).

Rooney, D., & McKenna, B. (2007). Wisdom in organizations: Whence and whither. *Social Epistemology*, *21*(2), 113-138. http://dx.doi.org/10.1080/02691720701393434

Sheninger, E. (2019). *Digital leadership: Changing paradigms for changing times*. Corwin Press.

Sousa, M.J., Cruz, R., & Martins, J.M. (2017). Digital learning methodologies and tools–a literature review. *Edulearn17 Proceedings*, 5185-5192. https://doi.org/10.21125/edulearn.2017.2158

Tanniru, M., & Peral, J. (2021). Digital Leadership in Education. *In Effective Leadership for Overcoming ICT Challenges in Higher Education: What Faculty, Staff and Administrators Can Do to Thrive Amidst the Chaos* (pp. 73-91). Emerald Publishing Limited. https://doi.org/10.1108/978-1-83982-306-020211008

Tanniru, M., Khuntia, J., & Weiner, J. (2018). Hospital leadership in support of digital transformation. *Pacific Asia Journal of the Association for Information Systems*, *10*(3), 1. https://aisel.aisnet.org/pajais/vol10/iss3/1/

Tigre, F.B., Curado, C., & Henriques, P.L. (2023). Digital leadership: A bibliometric analysis. *Journal of Leadership & Organizational Studies*, *30*(1), 40-70. https://doi.org/10.1177/15480518221123132

Vandenberg, R.J., & Lance, C.E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4-69. https://doi.org/10.1177/109442810031002

Yang, Y., & Green, S.B. (2011). Coefficient alpha: a reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment*, *29*(4) 377-392. https://doi.org/10.1177/0734282911406668

Yılmaz, V., & Çelik, H.E. (2009). *Yapısal eşitlik modellemesi*-*I*, (1. baskı) [*Structural equation modeling*-*I*], (1st ed.). Pegem Akademi Yayınları.

Yukl, G. (2009). Leading organizational learning: Reflections on theory and research. *The Leadership Quarterly, 20*(1), 49-53. https://doi.org/10.1016/j.leaqua.2008.11.006

Yusof, M., Yaakob, M., & Ibrahim, M. (2019). Measurement model of teaching competency of secondary school teachers in Malaysia. *International Journal of Emerging Technologies in Learning (iJET), 14*(20), 157-164. https://doi.org/10.3991/ijet.v14i20.11465

Zhong, L. (2017). Indicators of digital leadership in the context of K-12 education. *Journal of Educational Technology Development and Exchange (JETDE)*, *10*(1), https://doi.org/10.18785/jetde.1001.03

APPENDIX

Dijital Leadership Scale (Turkish Version)

Aşağıdaki maddeler öğretmenlerin dijital liderlik özellikleri ve algılarını ortaya çıkarmak amacıyla hazırlanmıştır. Lütfen maddeleri görevli olduğunuz okulu dikkate alarak değerlendirip, uygun olan kutucuğa **X** işareti koyunuz.	**SIKLIK DERECESİ**				
	① Hiçbir Zaman	**② Nadiren**	**③ Bazen**	**④ Çoğu Zaman**	**⑤Her Zaman**
Dijital Liderlik Ölçeği					
1. İşimi ve dijital çalışma ortamımı etkileyen kararlara dahil edilirim.	①	②	③	④	⑤
2. Okul müdürüm dijital okuryazarlığımı geliştirmem için beni destekler.	①	②	③	④	⑤
3. Dijitalleşmeyle ilgili sorunlarım olduğunda okul müdürümden destek alırım.	①	②	③	④	⑤
4. Dijital çalışmalarımın niteliği hakkında düzenli olarak okul müdürümden geri bildirim alırım.	①	②	③	④	⑤
5. Okul müdürüm dijital işlerimi yapmak için ihtiyacım olan bilgilere ulaşmamı destekler.	①	②	③	④	⑤
6. Dijital uygulamaları daha iyi anlamam ve kullanmam konusunda okul müdürüm beni destekler.	①	②	③	④	⑤
7. Okul müdürüm, okulda dijital çalışma yöntemlerini teşvik eder.	①	②	③	④	⑤
8. Görevlerimi yerine getirmek için hangi çalışma yöntemlerini, süreçleri ve çözüm yaklaşımlarını kullanacağımı belirleyebilirim.	①	②	③	④	⑤
9. Meslektaşlarımla işbirliği içinde görevlerimi yerine getiririm.	①	②	③	④	⑤

International Journal of Assessment Tools in Education
2025, Vol. 12, No. 1, 147–164

https://doi.org/10.21449/ijate.1403112

journal homepage: https://dergipark.org.tr/en/pub/ijate **Research Article**

Validity and reliability study of metacognitive listening strategies teaching self-efficacy scale for teachers

Murat Ermiş[1*], Şafak Uluçınar Sağır[2]

[1]Amasya University, Institute of Social Sciences, Amasya, Türkiye
[2]Amasya University, Faculty of Education, Amasya, Türkiye

ARTICLE HISTORY

Received: Dec. 11, 2023

Accepted: Oct. 13, 2024

Keywords:
Metacognition,
Metacognitive strategies,
Listening,
Self-efficacy,
Teacher training.

Abstract: In this study, an attempt was made to develop a valid and reliable measurement tool to determine teachers' self-efficacy levels for teaching metacognitive listening strategies. The study group consisted of 205 teachers for EFA and 248 teachers for CFA. As a result of the analyzes, a scale consisting of 16 items with 4 factors was developed. It was determined that the scale explained 74.10 of the total variances. For EFA, Kaiser-Mayer-Olkin (KMO) test, Barlett test, total variance, item-total correlation, common factor variance, factor loadings, χ^2/df RMSEA, SRMR, RMR, NFI, NNFI, CFI, GFI, AGFI, IFI, RFI, CR, AVE, MSV and ASV values and Cronbach Alpha statistics were performed. The KMO value of the scale is .915, the result of Barlett's Test of Sphericity is significant (p=.000) and the Cronbach Alpha value is .932. Five of the fit indices showed excellent fit and six of them showed acceptable fit. The CR, AVE, MSV and ASV values showed that it provided divergent and convergent validity. After the analysis, it was concluded that the Self-Efficacy Scale for Teaching Metacognitive Listening Strategies for Teachers is valid and reliable.

1. INTRODUCTION

Flavell introduced metacognition as a concept that refers to the forecasting, planning, monitoring and evaluation of one's own cognition. Metacognition includes awareness and control of cognitive strategies as well as knowledge of the person, task and strategy variables that affect an individual's learning and problem-solving. Listening is an important basic skill in education as well as in many areas of daily life. People can develop various listening strategies according to their own cognitive processes and make their listening more efficient. Metacognitive listening strategies are thought to positively affect students' listening skills. Therefore, it is important for teachers to teach these strategies and their use to their students. In this study, a measurement tool was tried to be developed in order to determine teachers' self-efficacy in teaching these strategies.

*CONTACT: Murat ERMİŞ ✉ ermismurat05@gmail.com Amasya University, Institute of Social Sciences, Amasya, Türkiye

e-ISSN: 2148-7456

1.1. Metacognition

The concept of metacognition was based on Flavell's meta-memory concept in the 1970s and first appeared in Flavell's work. According to Flavell (1976), metacognition is a concept that includes monitoring and regulation. Since metacognition is an abstract concept, there are many definitions of metacognition. Metacognition is the individual's self-knowledge about his own learning and knowledge about his cognition (Flavell, 1979). According to Brown (1978), metacognition is students' awareness and regulation of their own thinking processes in previously planned learning activities and problem situations. According to McCormick *et al.* (1989), it refers to the knowledge that individuals have about their own thinking processes and strategies, as well as their monitoring and regulation abilities in these learning processes (Melanlıoğlu, 2011). Doğanay and Kara (1995) state that the individual's awareness of his own thinking; Taylor (1999) stated that the individual's evaluation of what he knows is metacognition. Hacker and Dunlosky (2003) defined metacognition as the awareness of the mental activities carried out in the human brain and the ability to control them. Although there are some differences between these definitions, metacognition is generally defined as "the individual's planning, monitoring and regulating how he knows by thinking about his own learning" (Melanlıoğlu, 2011).

In order to make metacognition more understandable, it is important to explain its relationship with cognition (Sarıkaya, 2021). While cognitive learning does not include any critical perspective, metacognitive learning is different in terms of pre-planning, monitoring and evaluating the process (Katrancı, 2012). However, it is also important to know metacognitive knowledge and metacognitive strategies in understanding metacognition. According to Akın (2006), metacognitive knowledge is the individual's knowledge and awareness about his own cognition. The individual knows what he can or cannot do about himself, and can compare his own cognition with other individuals. According to Schraw (1998), metacognitive knowledge is the individual's understanding and comprehension of his or her own thought processes.

Metacognitive strategies refer to the tools that individuals use to keep their learning processes under control. Thinking about learning, making a learning plan, detecting meaningful situations in the learning process, and checking whether a product is produced are using metacognitive strategies (Chamot *et al*., 1987). The task of these strategies is to control learning processes, and regulate cognition and thinking (Schraw & Moshman, 1995). Metacognitive strategies refer to the processes that enable cognition and regulate the cognition process (Brown & Palincsar, 1982). Metacognitive strategies are the ability of individuals to control themselves consciously and for certain periods of time in order to determine whether they have achieved their goals or not and to decide whether to make a change in their strategies (O'Neil & Abedi, 1998). Or making plans to learn and produce knowledge, developing awareness about the steps and strategies for solving problems and evaluating themselves are metacognitive strategies (Costa, 2008). According to Wenden (1998), metacognitive strategies are skills that consist of planning, monitoring evaluating and managing, directing and regulating the individual's learning. Hauck (2005) states that the number of strategies used and the intervals in which these strategies are used are distinguishing features for the success of individuals.

The main strategies are preparation, planning, control, problem-solving and monitoring. The general view is that individuals use metacognitive strategies to plan, monitor and evaluate his own learning (Brown & Palincsar, 1982; Brown *et al.*, 1982; Cohen, 1994; Deseote & Roeyers, 2002; Kim, 2013; O'Malley *et al*., 1987; Schraw & Moshman, 1995; Wenden, 1998; Wey, 1998). The development of the ability to use metacognitive strategies increases many skills such as more efficient use of previously known strategies, better understanding of the problem and finding different solutions (Schraw, 1998).

There are also various metacognitive strategy models, with a few differences. These are the Oxford model, O'Malley and Chamot model, Cohen model, Greenfell and Harris model (Liu, 2010), Rubin model, and Anderson model (Anderson, 2002; Chamot & Robins, 2005). The more accepted and referenced model in research is the O’Malley and Chamot model. The O'Malley and Chamot model is taken as a reference in this study as it includes forecasting, planning, monitoring and evaluation strategies.

1.2. Listening

Language consists of five basic skills. These basic skills are listening, speaking, reading, writing and visual literacy. Although listening comprehension was previously thought to be a passive skill that develops with speaking and reading, today this skill is recognized as an active skill that can be taught (Rost, 2013). This idea may also be an explanation for the limited number of studies on listening compared to other skills (Melanlıoğlu, 2011). Listening is a process that requires training. The training of this skill should begin at an early age (Melanlıoğlu, 2011). Before school age, this task falls to mothers and fathers, and at school age, it falls to teachers (Temur, 2001).

Listening is a critical component of effective communication and plays a vital role in our personal and professional lives (Arnold, 2014). Listening is a fundamental language skill that is often overlooked by language teachers despite its importance (Malureanu & Enachi-Vasluianu, 2016). Listening is not only a skill area in language performance but also an important way of acquiring a second language (Rost, 2001). Listening allows us to process language in real time, using the speed, coding units and pauses that characterize spoken language (Hattingh, 2014). In terms of all these functions, listening is an important language skill.

Metacognitive listening strategies have been used in many studies on listening education. Although it has been mostly used in experimental studies on foreign language teaching, there are also studies in which it is used in native language education (Berman, 1994; Chamot & Robbins, 2005; Cohen & Brooks-Carson, 2001; Manchon *et al*., 2009; Rubin, 2001; Wolfersberger, 2003). In studies investigating the effects of metacognitive strategies on listening, it has been concluded that the use of metacognitive strategies has a positive effect on listening skills. Birjandi and Rahimi (2012) stated that students who use metacognitive listening strategies more effectively are better listeners. Bozorgian (2012) stated that thanks to strategy teaching, especially less skilled listeners can become more efficient listeners. According to Coşkun (2010), strategy instruction should be included in curricula in order for students to become better listeners. Cross (2010) and Goh and Taib (2006) stated that while strategy-based instruction improves the listening skills of less skilled students, this improvement is very low in more skilled students. Ghapanchi and Taheryan (2012) stated that as individuals' metacognitive knowledge and their ability to use metacognitive listening strategies increase, their speaking and listening skills also increase. According to Imhof (2001), strategy use and self-assessment facilitate listening. According to Kurita (2012), metacognitive strategy use not only improves listening skills but also reduces anxiety. Strategy use in foreign language teaching enables individuals to become better listeners (Vandergrift, 2003; Vandergrift *et al.*, 2006).

1.3. Self-Efficacy

Self-efficacy is one of the concepts that Bandura (1977) attaches importance to in his Social Learning Theory; It expresses the individual's self-belief in doing a job and being successful in that job. Ermiş (2019) examined studies and determined that self-efficacy has been shown to affect individuals' motivation, cognitive skills, and behavior. Gülebağlan (2003) concluded that teachers with high levels of self-efficacy do not have difficulty in making certain decisions in teaching activities and show a more determined attitude in this regard. According to Klassen and Tze (2014), teachers' self-efficacy about teaching a subject or using a skill also affects their teaching efficiency.

In order to teach metacognitive strategies to students, teachers must first learn these strategies and be models for students by using these strategies. Teachers can use these strategies out loud if necessary, and make students feel what they are doing at each stage, which strategies they are using, or what questions they are asking themselves. By teaching metacognitive strategies, students can be enabled to use these strategies independently. In each of the stages of forecasting, planning, self-monitoring and evaluation, the teacher can contribute to the development of students' skills in using metacognitive strategies by giving explicit instructions. Thus, students will be able to learn which strategies to use when listening and which strategies improve their listening skills.

Studies should be conducted to organize activities that can improve the skills of both students and teachers in using metacognitive strategies (Melanlıoğlu, 2011). Determining teachers' self-efficacy levels in teaching metacognitive listening strategies will provide significant support to the studies. Determining teachers' self-efficacy levels in teaching metacognitive listening strategies and, if necessary, organizing training programs for teachers on the use and teaching of these strategies can contribute to more reliable studies that reveal the effects of metacognitive listening strategies on students' listening skills.

Self-efficacy determination tools enable individuals to determine their level of perception of their own skills in a certain field (Aypay, 2010). Thus, individuals will be able to identify their advantageous and disadvantageous aspects and take steps to eliminate them. After the literature review, scales related to metacognition were used to measure individuals' metacognitive beliefs in psychopathology (Tosun & Irak, 2008), and students' metacognitive awareness (Haghighi *et al.*, 2019; Kaplan & Duran, 2016; Nix, 2016; Vandergrift *et al.*, 2006; Zhang & Zhang, 2011), metacognitive self-efficacy (Thomas *et al.*, 2008), metacognition skills (Hameed & Cheruvalath, 2021) and teacher candidates' metacognitive skills (Melanlıoğlu, 2011; Okur & Azizoğlu, 2016; Topaç, 2019), but a scale to determine teachers' self-efficacy levels in teaching metacognitive listening strategies could not be reached. It was thought that determining teachers' self-efficacy levels in teaching these strategies would contribute to the evaluations regarding the teaching of the strategies, and in this study, an attempt was made to develop a measurement tool to determine teachers' self-efficacy in teaching metacognitive listening strategies.

2. METHOD

2.1. Study Group

The sample of the study consists of classroom teachers. Since Exploratory Factor Analysis and Confirmatory Factor Analysis will be conducted within the scope of the research, there are two sample groups in the research. In this study, the bisection method was used for the data obtained as a result of the same application. According to DeVellis (2016), even if there is no problem with the scale items, the mental states of two different groups of participants such as fatigue and boredom during answering may prevent the real situation from emerging. In addition, no matter how similar the two samples are, conducting the analyses by dividing the first sample gives valuable information about the stability of the scale. For this reason, the data were divided into two halves and reliability analysis was performed. Some information about the EFA and CFA study groups is presented in Table 1.

Data obtained from 205 participants were used for EFA. Of the 205 teachers, 51.7% are women (*n*=106) and 48.3% are men (*n*=99). 2.4% of the teachers have associate degrees (*n*=5), .5% have institute graduate degrees (*n*=1), 72.2% have undergraduate degrees (*n*=148), 22.9% have master's degrees (*n*=148). *n*=47) and 2% are PhD graduates (*n*=4). 15.2% of the participating teachers had 0-5 years of experience (*n*=31), 14.6% had 6-10 years of experience (*n*=30), 30.7% had 11-15 years of experience (*n*=63). 39.5% have 16 years or more (*n*=81) professional experience. 18.1% of the teachers work in the village (*n*=37), 11.7% in the town (*n*=24) and 70.2% in the city center (*n*=144).

Table 1. *Information on the study group.*

		First Study Group (EFA)		Second Study Group (CFA)	
		n	%	*n*	%
Gender	Female	106	51.7	142	57.3
	Male	99	48.3	106	42.7
	Total	205	100	248	100
Education Status	Associate Degree	5	2.4	2	.8
	Institute	1	.5	0	0
	Undergraduate	148	72.2	186	75
	Master’s Degree	47	22.9	57	23
	PhD	4	2	3	1.2
	Total	205	100	248	100
Professional Experience	0-5 Years	31	15.2	29	11.7
	6-10 Years	30	14.6	38	15.3
	11-15 Years	63	30.7	50	20.2
	16 Years and More	81	39.5	131	52.8
	Total	205	100	248	100
Region of Assignment	Village	37	18.1	40	16.1
	Town	24	11.7	40	16.1
	City Center	144	70.2	168	67.8
	Total	205	100	248	100

Data obtained from 248 participants were used for CFA. Of the 248 teachers, 57.3% are women (*n*=142) and 42.7% are men (*n*=106). .8% of the teachers had an associate degree (*n*=2), 75% had an undergraduate degree (*n*=186), 23% had a master's degree (*n*=57) and 1.2% had a doctorate degree (*n*=3). 11.7% of the participating teachers had 0-5 years of experience (*n*=29), 15.3% had 6-10 years of experience (*n*=38), 20.2% had 11-15 years of experience (*n*=50). 52.8% have professional experience of 16 years or more (*n*=131). 16.1% of the teachers work in the village (*n*=40), 16.1% in the town (*n*=40) and 67.8% in the city center (*n*=168).

2.2. Collection and Analysis of Data

Ethics Committee Permission was obtained for the scale on 16.12.2022 and an online form was created via Google Forms. For the validity and reliability studies of the scale, data were collected using these forms within 4 months. DeVellis (2016) suggested a 7-stage method for scale development studies. In this study, these 7-step scale development stages were used.

2.2.1. *Stage 1: Determination of the feature to be measured*

In this study, we tried to develop a valid and reliable measurement tool to determine teachers' self-efficacy in teaching metacognitive listening strategies. During the development stages of the Metacognitive Listening Strategies Teaching Self-Efficacy Scale for Teachers, the relevant literature was first examined, but a scale to determine teachers' self-efficacy levels in teaching metacognitive listening strategies could not be reached. It has been seen that there are scales mostly to measure the metacognitive skills of students and teacher candidates (Hameed & Cheruvalath, 2021; Haghighi, *et al.*, 2019; Kaplan & Duran, 2016; Karakelle & Saraç, 2007; Melanlıoğlu, 2011; Nix, 2016; Okur & Azizoğlu, 2016; Thomas *et al.*, 2008; Topaç, 2019; Vandergrift *et al.*, 2006; Zhang & Zhang, 2011). After this scanning, the features to be measured were determined.

2.2.2. *Stage 2: Creating the item pool*

At this stage, the item pool for the scale is created. An item pool of 78 items was created by using the scales developed in studies conducted for students and the information obtained from articles and theses covering metacognition teaching sections (Hameed & Cheruvalath, 2021; Haghighi *et al.*, 2019; Kaplan & Duran, 2016; Melanlıoğlu, 2011; Nix, 2016; Okur & Azizoğlu, 2016; Thomas *et al.*, 2008; Topaç, 2019; Tosun & Irak, 2008; Vandergrift *et al.*, 2006; Zhang & Zhang, 2011). The created item pool was examined by 1 measurement and evaluation expert and 2 Turkish education experts. As a result of the review, it was seen that there were items measuring the same skills and the number of items was reduced to 40.

2.2.3. *Stage 3: Determining the format of the scale*

In the third stage, the format of the scale is determined. It was decided that the scale to be developed to determine teachers' self-efficacy levels in teaching metacognitive listening strategies would be Likert type. The scale was created as a five-point Likert and the options "Never, Rarely, Sometimes, Often and Always" were selected.

2.2.4. *Stage 4: Submission of the article pool for expert opinion*

At this stage, the created items are presented to expert opinion. Content validity refers to the ability of a scale to measure the desired feature. In studies, when it is not possible to apply it during the scale development stages, content validity rates are used. Content validity rates are determined by statistically calculating expert opinions (Yurdugül, 2005).

To ensure the content validity of the scale, the Lawshe technique was used by utilizing expert opinions. The Lawshe technique consists of 6 stages.

a) Establishing a group of field experts

b) Preparation of candidate scale forms

c) Obtaining expert opinions

d) Obtaining content validity rates for the items

e) Obtaining content validity indexes for the scale

f) Creating the final form according to the content validity rates/index criteria.

The Lawshe technique requires the opinions of at least 5 and at most 40 experts. Experts' opinions about the items are collected and content validity rates are calculated. The content validity rate (CVR) is obtained by subtracting 1 from the ratio of the number of experts expressing a "Necessary" opinion on any article to half of the total number of experts expressing an opinion on the article. (Yurdugul, 2005).

These 40 items were presented to the opinions of 2 classroom education experts, 2 measurement and evaluation experts and 3 Turkish language teaching experts. In line with the recommendations of experts, phrases that may be difficult to understand were changed and 1 item with a KVR value below .99 was removed from the scale. Thus, the first draft of the 39-item scale was created.

2.2.5. *Stage 5: Finalizing the item pool*

At this stage, it is decided whether to add items to the scale. In this study, after expert opinions, it was concluded that there was no need to add anything to the scale.

2.2.6. *Stage 6: Implementation*

Researchers have different opinions about the required sample size in scale development studies. Field (2005) stated that there should be at least 300 participants for EFA. However, there are also researchers who suggest that the sample size should be determined according to a certain multiple of the number of items. Kline (1994) suggested that there should be 2 times the number of items, MacCallum *et al.* (2001) 4 times, Bryman and Cramer (2004) 5 times, and Nunnally (1978) 10 times the number of participants. In this study, it was aimed to reach 5 times the number of participants for EFA and the data obtained from 205 participants were used. However, there are different opinions about the sample size required for CFA. Anderson and Gerbing (1984) stated that it should be larger than 100, Boomsma (1985) stated that it should be 100-200 participants, Jackson (2001) stated that it should be larger than 200, Stevens (2002) stated that it should be 5-10 participants for each item, De Winter *et al*. (2009) stated that it should be 3, 6, 20 participants for each item. In this study, data obtained from 248 participants were used for CFA. Çokluk *et al.* (2010) stated that meeting at least two of the sample size criteria specified in the literature is appropriate for scale development studies. In this study, the number of participants was reached in a way to provide two of the opinions stated separately for EFA and CFA.

2.2.7. *Stage 7: Analyzing the scale and finalizing the scale*

In the seventh stage, validity and reliability analyzes of the scale to be developed are performed. At this stage, information about the analyzes performed and the procedures performed during the analyzes is given. Studies conducted to ensure content validity for the Metacognitive Listening Strategies Teaching Self-Efficacy Scale for Teachers were included in the previous stages. Statistics should be used to ensure the construct validity of the scales (Yurdubakan, 2010). Exploratory and Confirmatory Factor Analysis were conducted to ensure the construct validity of the scale tried to be developed in this study.

In order to determine the discriminatory power of the scale, it was checked whether the difference between the lower group and upper group scores was significant. 27% of 205 participants correspond to 55 participants. The averages of the scores received by the participants were listed from highest to smallest, and then the scores of the group with 55 participants in the lower group and the group with 55 participants in the upper group were calculated by independent sample t-test analysis. As a result of the analysis, the difference between the two groups was found to be significant (p = .00). According to this result, it was seen that the scale items enabled the measurement of the feature that was intended to be measured.

The reliability study of the scale was conducted with Exploratory Factor Analysis, and it was decided whether the items in the scale would be removed or not. The factor load values of the items obtained in the Exploratory Factor Analysis were .30, which was accepted as the limit value (Büyüköztürk, 2020). In this study, .40 was determined as the limit value for item loads, and EFA examined whether there were any items with item loads below .40. Since there was no item with an item load below .40, no item was removed due to this criterion.

Using one of the rotation techniques in factor analyzes makes it easier to interpret the analysis (Osborne, 2015). If the number of factors is thought to be more than 2, it is more useful to use one of these orthogonal rotation techniques. If one of the orthogonal rotation techniques is to be used in social science studies, the varimax technique is generally used (Çokluk, Şekercioğlu & Büyüköztürk, 2010). In this study, the Varimax technique, one of the orthogonal rotation techniques, was used, considering that factorization would give a more conceptually meaningful result. After this rotation process, the scale revealed a 4-factor structure. The difference between the loading values of items on more than one factor should be higher than 0.10. As a result of the analysis, UST4, UST6, UST11, UST12, UST14, UST15, UST18 and UST35 were removed from the scale because they were included in more than one factor and the difference between the load values was less than 0.10. As a result of these procedures, the scale showed a structure consisting of 4 factors and 31 items. In this form, EFA was applied and KMO Test and Bartlett Test were calculated. In the EFA results, the KMO value is expected to be greater than .60 and the Bartlett Test is expected to be significant (Büyüköztürk, 2020).

Confirmatory Factor Analysis (CFA) of the scale was conducted with the data obtained from 248 participants. CFA is an attempt to prove the accuracy of a theoretically based scale, thanks to the collected data (Gürbüz, 2021; Weston & Gore, 2006). In order to determine whether a scale model is appropriate or not, it must meet certain criteria as a result of CFA. As a model, CFA differs from Exploratory Factor Analysis in that it starts from a theoretical basis (Byrne, 2001; Schreiber *et al.*, 2006).

The criteria required to determine the suitability of the model in CFA are based χ^2/df, Root Mean Square Error of Approximation (RMSEA), Standardized Root Mean Square Residual (SRMR), Root Mean Square Residual (RMR), Normed Fit Index (NFI), Non-Normed Fit Index (NNFI), Comparative Fit Index (CFI), Goodness of Fit Index (GFI), Adjusted Goodness of Fit Index (AGFI), Incremental Fit Index (IFI) and Relative Fit Index (RFI) values. In the literature, researchers have expressed different opinions about the fit indices that should be looked at to determine fit (Baumgartner & Homburg, 1996; Bentler, 1980; Bentler & Bonett, 1980; Kline, 2015; Marsh *et al.*, 2006; Schermelleh, Moosbrugger, & Müller, 2003). These fit indices were taken as criteria for the fit of the model during CFA. Low factor loadings of the items may cause the reliability coefficient of the model to decrease. In such cases, removing the items would be a healthier method (Gürbüz, 2021).

Following the analyses, items with low item factor loadings (UST5, UST7, UST8, UST9, UST10, UST13, UST34 and UST39) were removed from the scale. Additionally, modifications must be made from time to time to ensure the compatibility of the model. The fewness of these modifications are important and affect reliability. Items that caused an increase in modifications and affected the fit of the model (UST16, UST 17, UST19, UST27, UST28, UST31, UST33) were also removed from the scale.

According to Gürbüz (2021), when an item or factor is removed as a result of CFA, EFA can be performed again, and the validity and reliability analyzes of the scale can be done again. Since item removal was in question in this study, Cronbach Alpha values and factor analyzes were re-done to calculate the internal consistency reliability of the scale to ensure structural reliability.

In the study, SPSS 25 program was used for Exploratory Factor Analysis and AMOS program was used for Confirmatory Factor Analysis. For EFA, Kaiser-Mayer-Olkin (KMO) test, Barlett test, total variance, item-total correlation, common factor variance, and factor loadings statistics were performed. While conducting CFA, researchers agree on reporting the χ^2/df value (İlhan & Çetin, 2014). In addition, McDonald and Ho (2002) suggested that CFI, GFI, NFI and NNFI (TLI) should be reported, Brown (2006) suggested that RMSEA, SRMR, CFI and NNFI (TLI) should be reported, and Iacobucci (2010) suggested that CFI and SRMR values should be reported. In the light of these opinions, RMSEA, SRMR, RMR, NFI, NNFI, CFI, GFI, AGFI, IFI, RFI, CR, AVE, MSV and ASV values and Cronbach Alpha were calculated in order to determine the convergent and divergent validity of the scale along with model fit.

3. RESULTS

KMO value and Barlett Sphericity Test, scale total variance and Cronbach's Alpha value were analyzed for the Self-Efficacy Scale for Teaching Metacognitive Listening Strategies for Teachers and presented in Table 2. The Scree Plot graph of the scale is shown in Figure 1.

Table 2. *KMO and Barlett Sphericity test results.*

KMO Sample Suitability Measure		.915
Barlett's Test of Sphericity	Chi-Square	2613.551
	fd	.120
	p	.000

Figure 1. *AFA scree plot graphic.*

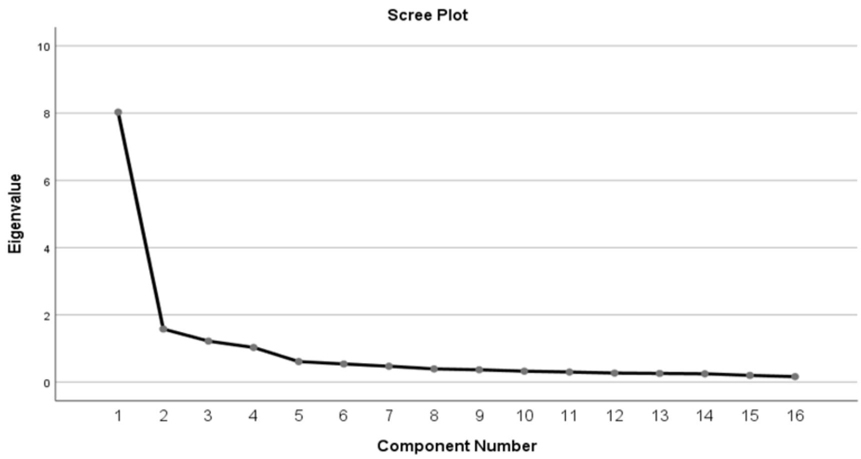


Table 3. *Rotated components table.*

Item	Factors			
	Factor 1	Factor 2	Factor 3	Factor 4
UST1			.821	
UST2			.870	
UST3			.738	
UST20		.793		
UST21		.749		
UST22		.733		
UST23		.687		
UST24		.589		
UST25	.722			
UST26	.834			
UST29	.838			
UST30	.772			
UST32	.799			
UST36				.574
UST37				.864
UST38				.809

After the analysis, the KMO value of the Metacognitive Listening Strategies Teaching Self-Efficacy Scale for Teachers was .915 and the Bartlett Sphericity Test result was significant (*p* = .00). According to Table 3 items of the scale, which has a 4-factor structure, had values between .574 and .870.

Table 4. *Reliability and total variance table.*

Factor	Cronbach Alpha	Explained Variance	General Cronbach Alpha	Total Explained Variance
Forecasting	.803	%15.07	.932	%74.10
Planning	.839	%13.41		
Monitoring	.917	%25.58		
Evaluation	.885	%20.02		

According to Table 3, as a result of the reliability and validity analysis, it was determined that the scale consists of 4 factors and 16 items. According to the expressions in the articles, the factors are named Forecasting, Planning, Monitoring and Evaluation. The Cronbach Alpha value of the scale was found to be .803 for the Forecasting factor, .839 for the Planning factor, .917 for the Monitoring factor and .885 for the Evaluation factor. The Cronbach Alpha value of the overall scale is .932. According to Table 4, it was determined that the scale explained 74.10% of the total variance. This value is 15.07% for the Forecasting factor, 13.41% for the Planning factor, 25.58% for the Monitoring factor and 20.02% for the Evaluation factor.

Sample items from some factors.

Forecasting- UST2- I think I can do the activities to be done during listening.

Plannig-UST37- I think that designing metacognitive activities requires a systematic approach.

Monitoring- UST25- I can create listening activities for teaching metacognitive listening strategies.

Evaluation- UST22- I can guide my students to think about what they would do differently the next time they listen.

After CFA analyses, the values of the scale according to various indices and its fit status are given in Table 5.

Table 5. *CFA Results of metacognitive listening strategies instruction self-efficacy scale for teachers.*

Indexes	Perfect Fit Criterion	Acceptable Fit Criterion	Scale Indexes	Compliance Status
χ^2/df	0-2.5	2.5-3	2.09	Perfect
RMSEA	≤05	≤08	.069	Acceptable
SRMR	≤05	≤08	.0513	Acceptable
RMR	≤05	≤08	.027	Perfect
NFI	≥95	≥90	.922	Acceptable
NNFI	≥95	≥90	.945	Acceptable
CFI	≥95	≥90	.956	Perfect
GFI	≥90	≥85	.907	Perfect
AGFI	≥90	≥85	.869	Acceptable
IFI	≥95	≥90	.956	Perfect
RFI	≥95	≥90	.902	Acceptable

It was concluded that the chi-square fit value (χ^2=209.361, *df*=96, *p*=.00) of the Metacognitive Listening Strategies Teaching Self-Efficacy Scale for Teachers was significant. The χ^2/df value for model fit is 2.09. It can be said that this value represents perfect fit (Kline, 2015). The RMSEA value of the scale is .069. This value represents acceptable fit. The SRMR value was calculated as .0513 and this value indicates acceptable fit. GFI and AGFI values close to 1 indicate perfect fit (Raykov & Marcaoulides, 2006). After the analysis, the GFI value of the scale is .907 and the AGFI value is .869. These values indicate perfect fit for GFI and acceptable fit for AGFI. NFI and CFI values being close to 1 indicate perfect fit (Kline, 2015; Raykov & Marcaoulides, 2006). The NFI value of the scale was calculated as .922 and the CFI value was .956. These values indicate acceptable fit for NFI and perfect fit for CFI. According to the results given in Table 5, as a result of the CFA performed on the specified sample, 6 of the findings obtained from the scale were determined to be acceptable and 5 of them to indicate perfect fit. Figure 2 shows the fit diagram of the scale.

Convergent validity expresses the relationships of the items with each other and the factors they form. Divergent validity refers to the low relationship of the items with other factors. CR, which expresses the combined reliability, and AVE, which expresses the average variance explained, are important to ensure the convergent validity of the scale (Hair *et al.*, 2014). According to Table 6, for each factor in the scale, the CR value is expected to be ≥ .70, the AVE value to be

≥.50, and the CR value to be greater than the AVE value (Fornell & Larcker, 1981). The fact that the CR value is greater than the AVE value for all factors in the scale indicates that the convergent validity of the scale is achieved.

Figure 2. *CFA Diagram of the scale.*

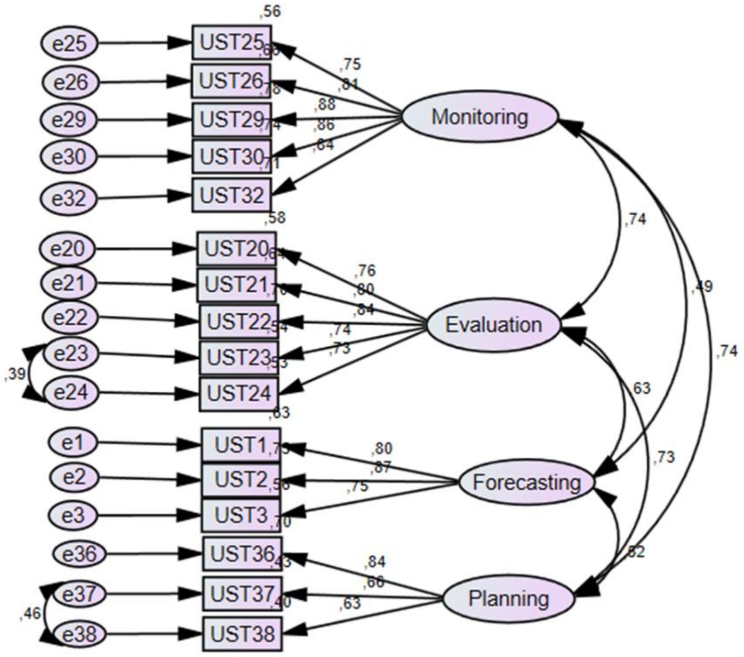


Table 6. *Convergent validity values of metacognitive listening strategies instruction self-efficacy scale for teachers.*

Factor	CR	AVE
Monitoring	.916	.688
Evaluation	.881	.549
Planning	.758	.515
Forecasting	.844	.645

CR: Composite Reliability/AVE: Average Variance Extracted

For divergent validity, MSV and ASV values need to be calculated. MSV, which expresses the Square of Maximum Shared Variance, is the square of the highest variance that a factor shares with one of the other factors. ASV, which expresses the Average of the Square of Shared Variance, is the sum of the squares of the variance shared by a factor with other factors, divided by the number of shared variances. To ensure divergent validity, MSV<AVE, ASV<MSV and the square root of AVE must be greater than the correlation between factors (Yaşlıoğlu, 2017). According to Table 7, it can be said that the scale provides divergent validity because it meets all these conditions.

Table 7. *Divergent validity values of metacognitive listening strategies instruction self-efficacy scale for teachers.*

Factors	Correlation Between Factors	MSV	ASV	Square Root AVE
Monitoring-Evaluation	.749	.561	.445	.8
Monitoring-Forecasting	.727	.528	.397	
Monitoring-Planning	.494	.244	.291	
Evaluation-Forecasting	.714	.509	.242	
Evaluation - Planning	.618	.381	.140	
Planning-Forecasting	.567	.321	.064	

MSV: Maximum Squared Variance/ ASV: Average Shared Square Variance

4. DISCUSSION and CONCLUSION

Listening skill is a skill that begins to develop in the womb and continues to develop throughout an individual's life. The limited number of studies on the development of listening skills over time has caused it to be perceived as a neglected skill. Metacognitive listening strategies are important for individuals in terms of monitoring the development of the learning process and guiding new learning. The use of metacognitive listening strategies can enable students to learn and develop their listening skills under their own control. Teaching these strategies by teachers at school will ensure that this development is rapid and planned. In this study, an attempt was made to develop a valid and reliable scale that can determine teachers' self-efficacy levels in teaching these strategies by developing the Metacognitive Listening Strategies Teaching Self-Efficacy Scale for Teachers. Following the literature review, measurement tools for measuring the metacognitive skills of students and teacher candidates were found (Melanlıoğlu, 2011; Okur & Azizoğlu, 2016; Topaç, 2019), but a measurement tool for determining the self-efficacy levels of teachers in teaching metacognitive listening strategies could not be found. Following these scales and literature review, an item pool consisting of 40 items was created. After the content validity study conducted with the Lawshe technique, one item was removed from the scale and the first draft of the scale consisting of 39 items was prepared.

EFA was performed on the scale with the data collected with the participation of 205 teachers, and after the analysis, 8 items that were included in more than one factor were removed from the scale. The item load limit for the items in the scale was determined as .40. Since it was seen that there was no item below this value, no item was removed from the scale due to the item load value. In the EFA results, the KMO value is expected to be greater than .60 and the Bartlett Test is expected to be significant (Büyüköztürk, 2020). In this form, the scale showed a structure consisting of 4 factors and 31 items.

Gürbüz (2021) stated that it would be appropriate to remove items or factors from the scale if necessary to ensure fit in the scale model. Therefore, 15 items that disrupted the fit in the CFA analyses were removed from the scale. After these procedures, the KMO and Barlett Sphericity Test results of the scale were examined again. The KMO value was .915 and Barlett's Test of Sphericity was significant (*p*=.00). The scale showed a structure consisting of 4 factors and 16 items. Scale items had item loadings between .574 and .870. The Cronbach Alpha value for the Forecasting factor of the scale was .803, .839 for the Planning factor, .917 for the Monitoring factor and .885 for the Evaluation factor. The Cronbach Alpha value of the overall scale is .932. After the analysis, it was determined that the scale explained 74.10% of the total variance. This value is 15.07% for the Forecasting factor, 13.41% for the Planning factor, 25.58% for the Monitoring factor and 20.02% for the Evaluation factor.

CFA was conducted on the scale with the data collected with the participation of 248 teachers. Within the framework of the opinions in the literature about the fit indices required to determine fit, χ^2/df, RMSEA, SRMR, RMR, NFI, NNFI, CFI, GFI, AGFI, IFI and RFI values were taken as basis to determine the suitability of the model (Baumgartner & Homburg, 1996; Bentler, 1980; Bentler & Bonett, 1980; Kline, 2015; Marsh *et al.*, 2006; Schermelleh *et al.*, 2003). The χ^2/df, RMSEA, SRMR, RMR, NFI, NNFI, CFI, GFI, AGFI, IFI and RFI values of the scale were calculated with CFA and it was determined that 6 of these values were acceptable and 5 were perfect fit. To determine the convergent validity of the scale, CR and AVE values for each factor were calculated. It was concluded that the CR value was greater than .70 for each factor, the AVE value was greater than .50 for each factor, and the CR value was greater than the AVE value for all factors. Accordingly, it can be said that the scale provides convergent validity. MSV and ASV values of the scale were calculated for divergent validity. To ensure divergent validity, MSV<AVE, ASV<MSV and the square root of AVE must be greater than the correlation between factors (Yaşlıoğlu, 2017). After the calculations, it can be said that the scale provides divergent validity.

Haghighi, Rashtchi, and Birjandi (2019) concluded that the scale they developed to determine students' metacognitive awareness had a 3-factor structure as Planning, Monitoring and Evaluation. The scale developed in this study showed a 4-factor structure. However, Planning, Monitoring and Evaluation factors are present on both scales. Kaplan and Duran (2016) stated that the scale named Mathematical Metacognition Awareness Inventory Towards Middle School Students consists of Mathematical Knowledge, Mathematical Monitoring and Mathematical Determination factors. Although the number of factors is different, the Mathematical Monitoring and Mathematical Determination factors are similar to the Monitoring and Evaluation factors. The scale developed by Nix (2016) to determine students' metacognitive awareness showed a 2-factor structure. Another scale developed to measure students' metacognitive awareness, MALQ, showed a five-factor structure (Vandergrift, Goh, & Mareschal, 2006). The scale prepared for university students learning a foreign language consists of Problem-solving, Planning and Evaluation, Translation, Person Knowledge, and Directed Attention factors. Although the number of factors is not the same, the Planning and Evaluation factor is also included in the scale developed in this study. Thomas, Anderson, and Nashon (2008) developed the SEMLI-S scale consisting of 30 items and 5 factors to determine students' metacognitive self-efficacy. Although it has more factors, it is similar to this scale in terms of the factor MEP (Monitoring, Evaluation, Planning) among the factors Cognitive Connectivity, MEP (Monitoring, Evaluation, Planning), Self-efficacy, Learning Risks Awareness and Control of Concentration. Hameed and Cheruvalath (2021) developed the MSI scale consisting of 12 items and one factor. The scale developed in this study is not compatible with MSI. Okur and Azizoğlu (2016) adapted the Metacognitive Listening Strategies Instrument (MLSI) into Turkish to determine the metacognitive skills of pre-service teachers and determined a structure consisting of 11 items and 3 factors. Among the 3 factors consisting of Attention, Planning and Evaluation and Problem-solving", the Planning and Evaluation factor is similar to our scale. The number of participants in these developed scales varies between 300 and 500. Our scale study is compatible with other scales in this respect.

In its final form, the scale showed a structure consisting of 4 factors and 16 items. The factors include metacognitive strategies of forecasting, planning, monitoring and evaluation. These sub-factors reveal teachers' self-efficacy levels in teaching these strategies. There are no reverse items in the scale. Therefore, the higher the average scores obtained from the scale, the higher the self-efficacy level. Factors can be examined in terms of the variable to be used in studies, and comments can be made about changes in self-efficacy levels according to these variables.

The concept of metacognition has been examined over time and studies on this subject are still continuing. The scale developed in this study was tried to be developed in the light of the studies carried out so far. The scale can be further developed with the contributions of future studies. However, studies can be conducted with different sample groups other than the sample group in this study. According to these results, it can be said that the Metacognitive Listening Strategies Teaching Self-Efficacy Scale for Teachers is a valid and reliable scale.

Acknowledgments

This article is derived from Murat Ermiş's PhD dissertation entitled "The Effect of Metacognitive Listening Strategy On 4th Grade Students' Listening Skills", prepared at Amasya University, Institute of Social Sciences under the supervision of Prof. Dr. Şafak Uluçınar-Sağır.

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**: Amasya University, E-30640013-108.01-108629.

Contribution of Authors

Murat Ermiş: Literature review, Investigation, Methodology, Item writing, Data collection, Receiving experts' opinions, Writing-original draft and Statistical analysis. **Şafak Uluçınar Sağır**: Data collection, Supervision and Critical review.

Orcid

Murat Ermiş https://orcid.org/0000-0002-8803-0612

Şafak Uluçınar Sağır https://orcid.org/0000-0003-3383-5330

REFERENCES

Akin, A. (2006). *Relationships between achievement goal orientations and metacognitive awareness, parental attitudes and academic achievement* [Unpublished master's thesis]. Sakarya University.

Anderson, J.C., & Gerbing, D.W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood con-firmatory factor analysis. *Psychometrika*, *49*(2), 155-173.

Anderson, N.J. (2002). The role of metacognition in second language teaching and learning. *ERIC Digest.*

Arnold, C.L. (2014). *Listening: The Forgotten Communication Skill*. OMICS Publishing Group, *04*(10). https://doi.org/10.4172/2165-7912.1000e155

Aypay, A. (2010). The adaptation study of General Self-Efficacy (GSE) Scale to Turkish. *Inonu University Journal of the Faculty of Education*, *11*(2), 113-132.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191. https://doi.org/10.1037/0033-295X.84.2.191

Baumgartner, H., & Homburg, C. (1996). Applications of structural equation modeling in marketing and consumer research: A review. *International Journal of Research in Marketing*, *13*(2), 139-161. https://doi.org/10.1016/0167-8116(95)00038-0

Bentler, P.M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology*, *31*(1), 419-456.

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

Berman, R. (1994). Learners' transfer of writing skills between languages. *TESL Canada Journal*, *12*(1), 29-46.

Birjandi, P., & Rahimi, A.H. (2012). The effect of metacognitive strategy instruction on the listening performance of EFL students. *International Journal of Linguistics*, *4*(2), 495-517.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika*, *50*(2), 229-242.

Bozorgian, H. (2012). Metacognitive instruction does improve listening comprehension. *ISRN Education*, 1-6.

Brown, A.L. (1978). Knowing when, where, and how to remember: A problem of metacognition. In R. Glaser (Ed.), *Advances in instructional psychology,* (Vol. 7, pp. 55–111). Academic Press.

Brown, A.L., & Palincsar, A.S. (1982). *Inducing strategic learning from texts by means of informed, self-control training*. Technical report No. 262.

Brown, A.L., Bransford, J.D., Ferrara, R.A., & Campione, J.C. (1982). *Learning, Remembering and Understanding.* Technical Report No. 244.

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

Bryman, A., & Cramer, D. (2004). *Quantitative data analysis with SPSS 12 and 13: A guide for social scientists*. Routledge. https://doi.org/10.4324/9780203498187

Büyüköztürk, Ş. (2020). Sosyal bilimler için veri analizi kitabi (28. basım) [*Handbook of data analysis for social sciences (28th Edition)*]. Pegem Akademi Publishing.

Byrne, B.M. (2001). Structural equation modeling with AMOS, EQS, and LISREL: Comparative approaches to testing for the factorial validity of a measuring instrument. *International Journal of Testing*, *1*(1), 55-86. https://doi.org/10.1207/S15327574IJT0101_4

Chamot, A. U., & Robbins, J. (2005). The CALLA Model: Strategies for ELL student success. *Workshop for Region 10*. New York City Board of Education.

Chamot, A.U., O'Malley, M.J., Kupper, L., & Impink-Hernandez, M.V. (1987). *A study of learning strategies in foreign language instruction: First Year Report*. ERIC Number: ED352824.

Cohen, A.D. (1996). Second language learning and use strategies: Clarifying the issues: Center for Advanced Research in Language Acquisition.

Cohen, A., & Brooks-Carson, A. (2001). Research on direct versus translated writing: students' strategies and their results. *The Modern Language Journal*, *85*, 169-188. https://doi.org/10.1111/0026-7902.00103

Cohen, A.D. (1996). Second language learning and use strategies: Clarifying the issues.

Costa, A.L. (2008). Describing the habits of mind. In A.L. Costa, B. Kallick (Eds.), *Learning and leading with habits of mind* (15-41). Association for Supervision and Curriculum Development.

Coşkun, A. (2010). The effect of metacognitive strategy training on the listening performance of beginner students. *Novitas-ROYAL (Research on Youth and language)*, *4*(1), 35-50.

Cross, J. (2010). Metacognitive instruction for helping less-skilled listeners. *ELT Journal*, *65*(4), 408-416. https://doi.org/10.1093/elt/ccq073

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2010). *Sosyal bilimler için çok değişkenli 497 istatistik. SPSS ve LISREL uygulamaları* [*Multivariate statistics for social sciences. SPSS and LISREL applications*]. Pegem Academy.

De Winter, J.C.F, Dodou, D., & P.A. Wieringa (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research*, *44*, 147-181.

Desoete, A., & Roeyers, H. (2002). Off-line metacognition-A domain-specific retardation in young children with learning disabilities? *Learning Disability Quarterly*, *25*(2), 123-139. https://doi.org/10.2307/1511279

DeVellis, R.F., & Thorpe, C.T. (2021). *Scale development: Theory and applications*. Sage publications.

Doğanay, A., & Kara, Z. (1995). Dimensions of thinking. *Çukurova University Journal of Faculty of Education*, *1*(11), 25–38.

Ermiş, M. (2019). *The relationship between teachers' self-efficacy levels and institutional commitment* [Unpublished master's thesis]. Amasya University.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage.

Flavell, J.H. (1976). Metacognitive aspects of problem solving. *The Nature of Intelligence*. Lawrence Erbaum.

Flavell, J.H. (1979). Metacognitive and cognitive monitoring: A new area of cognitive developmental inquiry. *American Psychologyst*, *34(10), 906-911*. https://doi.org/10.1037/0003-066X.34.10.906

Fornell, C., & Larcker, D.F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39-50. https://doi.org/10.1177/002224378101800104

Ghapanchi, Z., & Taheryan, A. (2012). Roles of linguistic knowledge, metacognitive knowledge and metacognitive strategy use in speaking and listening proficiency of ıranian efl learners. *World Journal of Education*, *2*(4), 64-75.

Goh, C., & Taib, Y. (2006). Metacognitive instruction in listening for young learners. *ELT Journal*, *60*(3), 222-232. https://doi.org/10.1093/elt/ccl002

Gülebağlan, C. (2003). *A research on the comparison of teachers' tendency to postpone the work to the last moment in terms of their perceptions of professional competence, professional experience and branches* [Unpublished master's thesis]. Ankara University.

Gürbüz, S. (2021). *Amos ile yapısal eşitlik modellemesi*. Seçkin Yayıncılık.

Hacker, D.J., & Dunlosky, J. (2003). Not all metacognition is created equal. *New Directions for Teaching and Learning*, *95*, 73-79.

Haghighi, M., Rashtchi, M., & Birjandi, P. (2019). Developing and validating a questionnaire to assess strategic competence in EFL listening performance: A structural equation modeling approach. *Research in English Language Pedagogy*, *7*(2), 336-362. https://doi.org/10.30486/relp.2019.665886

Hair Jr, J.F., Hult, G.T.M., Ringle, C.M., & Sarstedt, M. (2014). *A primer on partial least squares structural equation modeling (PLS-SEM).* Sage publications.

Hameed, H.A., & Cheruvalath, R. (2021). Metacognitive skills inventory (MSI): Development and validation. *International Journal of Testing*, *21*(3-4), 154-181. https://doi.org/10.1080/15305058.2021.1986051

Hattingh, S. (2014). The importance of teaching listening. *OIU Journal of International Studies*, *27*(3), 97-110.

Hauck, M. (2005). Metacognitive knowledge, metacognitive strategies, and CALL. *CALL Research Perspectives*, 65-86.

Iacobucci, D. (2010). Structural equations modeling: Fit indices, sample size, and advanced topics. *Journal of Consumer Psychology*, *20*(1), 90-98.

Imhof, M. (2000). How to Listen More Efficiently: Self-monitoring Strategies in Listening. *International Journal of Listening*, *15*(1), 2-19. https://doi.org/10.1080/10904018.2001.10499042

İlhan, M., & Çetin, B. (2014). Comparing the Analysis Results of the Structural Equation Models (SEM) Conducted Using LISREL and AMOS. *Journal of Measurement and Evaluation in Education and Psychology*, *5*(2), 26-42.

Jackson, D.L. (2001). Sample size and number of parameter estimates in maximum likeli-hood confirmatory factor analysis: A Monte Carlo investigation. *Structural Equa-tion Modeling*, *8*, 205-223.

Kaplan, A., & Duran, M. (2016). Mathematical metacognition awareness inventory towards middle school students: Validity and reliability. *Journal of Kazım Karabekir Education Faculty*, *32*, 1-17.

Katrancı, M., & Yangın, B. (2012). Effects of teachıng metacognıtıon strategıes to lıstenıng omprehensıon skılls and attıtude toward lıstenıng. *Adiyaman University Journal of Social Sciences*, (*11)*, 733-771.

Kim, S.H. (2013). *Metacognitive knowledge in second language writing*. [Unpublished master's thesis] Michigan State University.

Klassen, R.M., & Tze, V.M. (2014). Teachers' self-efficacy, personality, and teaching effectiveness: A meta-analysis. *Educational Research Review*, *12*, 59-76. https://doi.org/10.1016/j.edurev.2014.06.001

Kline, P. (1994). *An easy guide to factor analysis*. Routledge. https://doi.org/10.4324/9781315788135

Kline, R.B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Kurita, T. (2012). Issues in second language listening comprehension and the pedagogical implications. *Accents Asia*, *5*(1), 30-44.

Küçükahmet, L. (Ed.) (2001). *Dinleme Becerisi* [Listening skills]. Nobel Publication Distribution.

Liu, J. (2010). Language learning strategies and its training model. *International Education Studies*, *3*(3), 100-104.

MacCallum, R.C., Widaman, K.F., Preacher, K.J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate behavioral research*, *36*(4), 611-637. https://doi.org/10.1207/S15327906MBR3604_06

Malureanu, F., & Enachi-Vasluianu, L. (2016). The ımportance of elements of actıve lıstenıng ın dıdactıc communıcatıon: a student's perspectıve. *In CBU International Conference Proceedings.*, *4*, 332-335. https://doi.org/10.12955/cbup.v4.776

Manchon, R.M., Murphy, L., & de Larios, J.R. (2007). Lexical retrieval processes and strategies in second language writing: A synthesis of empirical research. International *Journal of English Studies*, *7*(2), 149-174.

Marsh, H.W., Hau, K.T., Artelt, C., Baumert, J., & Peschar, J.L. (2006). OECD's brief self-report measure of educational psychology's most useful affective constructs: Cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing*, *6*(4), 311-360. https://doi.org/10.1207/s15327574ijt0604_1

McDonald, R.P., & Ho, M.H.R. (2002). Principles and practice in reporting structural equation analyses. *Psychological methods*, *7*(1), 64.

Melanlıoğlu, D. (2011). *Impact of the metacognitive strategy instruction on secondary school students listening skill* [Unpublished Phd dissertation]. Gazi University.

Nix, J.M.L. (2016). Measuring latent listening strategies: Development and validation of the EFL listening strategy inventory. *System*, *57*, 79-97. https://doi.org/10.1016/j.system.2016.02.001

Nunnally, J.C. (1978). *An overview of psychological measurement*. Clinical diagnosis of mental disorders: A handbook. 97-146.

O'Malley, M.J., Chamot, A.U., Walker, C., Russo, R.P., & Kupper, L. (1987*). The role of learning strategies in second language acquisition: A selected literature review.* ARI. U.S. Army Research Institute for the Behavioral and Social Sciences, Technical Report 744.

Okur, A., & Azizoğlu, N.İ. (2016). Metacognitive Listening Strategies Instrument: Validity and Reliability Study. *Mehmet Akif Ersoy University Journal of Education Faculty*, *40*, 113-124.

O'Neil Jr, H.F., & Abedi, J. (1996). Reliability and validity of a state metacognitive inventory: Potential for alternative assessment. *The journal of educational research*, *89*(4), 234-245. https://doi.org/10.1080/00220671.1996.9941208

Osborne, J.W. (2015). What is rotating in exploratory factor analysis? *Practical Assessment, Research, and Evaluation*, *20*(1), 1-7. https://doi.org/10.7275/hb2g-m060

Raykov, T., & Marcoulides, G.A. (2006). On multilevel model reliability estimation from the perspective of structural equation modeling. *Structural Equation Modeling*, 13(1), 130-141. https://doi.org/10.1207/s15328007sem1301_7

Rost, M. (2001). *Listening*. Cambridge University Press, 7-13. https://doi.org/10.1017/cbo9780511667206.002

Rost, M. (2013). *Teaching and researching: Listening*. Routledge.

Rubin, J. (2001). Language learner self-management. *Journal of Asian Pacific Communication*, *11*(1), 25–37. https://doi.org/10.1075/japc.11.1.05rub

Sarıkaya, B. (2021). *Investigation of 7th grade elementary school students' metacognitive awareness levels and vocabulary learning strategies they use in English lessons*. [Unpublished Phd dissertation]. İnönü University.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*(2), 23-74.

Schraw, G. (1998). Promoting General Metacognitive Awareness. *Instructional Science*. *26*(1-2), 113–125. https://doi.org/10.1023/A:1003044231033

Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational psychology review*, *7*, 351-371. http://dx.doi.org/10.1007/s10648-017-9413-7

Schreiber, J.B., Nora, A., Stage, F.K., Barlow, E.A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research*, *99*(6), 323-338. https://doi.org/10.3200/JOER.99.6.323-338

Stevens, J. (2002). *Applied multivariate statistics for the social sciences* (Vol. 4). Lawrence Erlbaum Associates.

Taylor, S. (1999). Better learning through better thinking: Developing students' metacognitive abilities. *Journal of College Reading and Learning*, *30*(1), 34-45. https://doi.org/10.1080/10790195.1999.10850084

Thomas, G., Anderson, D., & Nashon, S. (2008). Development of an instrument designed to investigate elements of science students' metacognition, self-efficacy and learning processes: The SEMLI-S. *International Journal of Science Education*, *30*(13), 1701-1724.

Topaç, E. (2019). *The effect of metacognitive listening strategy instruction on EFL learnerslistening comprehension and awareness levels and the role of ted talks as a listening resource on studentsperceptions* [Unpublished master's thesis]. Yeditepe University.

Tosun, A., & Irak, M. (2008). Adaptation, validity, and reliability of the metacognition questionnaire-30 for the turkish population, and its relationship to anxiety and obsessive-compulsive symptoms. *Turkish Journal of Psychiatry*, *19*(1), 67-80.

Vandergrift, L. (2003). Orchestrating strategy use: Toward a model of the skilled second language listener. *Language Learning*, *53*(4), 463-496. https://doi.org/10.1111/1467-9922.00232

Vandergrift, L., Goh, C.C.M., Mareschal, C.J., & Tafaghodtari, M.H. (2006). The metacognitive awareness listening questionnaire: Development and validation. *Language Learning*, *56(*3), 431–462. https://doi.org/10.1111/j.1467-9922.2006.00373.x

Wenden, A.L. (1998). Metacognitive knowledge and language learning. *Applied Linguistics*, *19*(4), 515-537. https://doi.org/10.1093/applin/19.4.515

Weston, R., & Gore Jr, P.A. (2006). A brief guide to structural equation modeling. *The Counseling Psychologist*, *34*(5), 719-751. https://doi.org/10.1177/0011000006286345

Wey, S. (1998). *The effects of goal orientations, metacognition, self-efficacy and effort on writing achievement*. University of Southern California.

Wolfersberger, M. (2003). L1 to L2 writing process and strategy transfer: a look at lower proficiency writers. *TESL-EJ*, *7*(2), 1-12.

Yaşlıoğlu, M.M. (2017). Factor analysis and validity in social sciences: Application of exploratory and confirmatory factor analyses. *Istanbul Business Research*, *46*, 74-85.

Yıldırım A., Şimşek, H. (2008). *Sosyal bilimlerde nitel araştırma yöntemleri* [*Qualitative research methods in social sciences*]. Seçkin Yayıncılık.

Yurdubakan, İ. (2010) *Eğitimde kullanılan ölçme araçların nitelikleri* [*Qualifications of measurement tools used in education*]. In M. Gömleksiz & S. Erkan (Eds.). Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education]. (pp. 33-66). Nobel.

Yurdugül, H. (2005). *Ölçek geliştirme çalışmalarında kapsam geçerliği için kapsam geçerlik indekslerinin kullanılması* [*Using content validity indices for content validity in scale development studies*]. XIV. National Congress of Educational Sciences. Denizli.

Zhang, Z., & Zhang, L.J. (2011). Developing a listening comprehension problem scale for university students' metacognitive awareness. *The Journal of Asia TEFL*, *8*(3), 161-189.

International Journal of Assessment Tools in Education
2025, Vol. 12, No. 1, 165–179

https://doi.org/10.21449/ijate.1439457

journal homepage: https://dergipark.org.tr/en/pub/ijate **_Research Article_**

Adaptation of the attitudes towards mobile-assisted language learning scale to Turkish culture and language

Emirhan Bingöl[1*], **Adnan Taşgın**[2], **Savaş Yeşilyurt**[1]

[1]Atatürk University, Faculty of Education, Department of Foreign Language Teaching, Erzurum, Türkiye
[2]Atatürk University, Faculty of Education, Department of Education and Training Programs, Erzurum, Türkiye

ARTICLE HISTORY

Received: Feb. 20, 2024

Accepted: Oct. 21, 2024

Keywords:

Mobile-assisted language learning,
MALL,
Technology,
Smart devices,
CALL.

Abstract: The use of technological devices, especially mobile devices, in language learning has increased the number of studies in this field. In this regard, it is essential to identify students' attitudes towards mobile-assisted language learning (MALL). Therefore, the present study aimed to translate, adapt, and validate Gönülal's (2019) attitudes towards the MALL (A-MALL) scale to the Turkish language and culture. The study included 250 EFL learners from different cities in Türkiye who completed the adapted version of the 15-item A-MALL scale. To align the assumed factor loadings as closely as possible with the target matrix, confirmatory factor analysis was performed using the original study results as a calibration example. The results revealed that the adapted A-MALL scale has acceptable fit indexes; therefore, the Turkish version of the A-MALL scale is reliable and valid.

1. INTRODUCTION

During the last two decades, technology has become increasingly integrated into the teaching and learning of languages, and as a part of this process, computer-assisted language learning (CALL) has emerged (Kukulska-Hulme & Traxler, 2005). Then technology-assisted language learning has added new dimensions to the trend (Thorne & Smith, 2011). As a result of the ever-evolving and dynamic nature of technology, a new concept emerged in language learning: MALL (mobile-assisted language learning). Although MALL can be questionably considered another form of CALL (Gönülal, 2019), studies on MALL reveal that this concept has characteristics such as portability, interactivity, individuality, and wireless technologies (Chang *et al*., 2018; Jarvis & Achilleos, 2013; Kukulska-Hulme, 2009; Stockwell, 2013). Furthermore, the critical catchphrase in MALL studies is "anywhere, anytime" (Agca & Özdemir, 2013; Burston, 2014; Kolb, 2008; Stockwell, 2013). Thus, the concept of MALL is unique as it is easy to use, easy to access, flexible, helpful in facilitating collaboration in language learning, and independent of location.

*CONTACT: Emirhan BİNGÖL ✉ emirhan.bingol@hotmail.com Atatürk University, Faculty of Education, Erzurum, Türkiye

e-ISSN: 2148-7456

A different perspective on the importance of the studies and applications developed in the field is considering them as foundational knowledge for the unexpectedly emerging epidemic of Covid-19. In most countries, home-based learning has been adopted at all levels of education, as well as in informal institutions (Okmawati & Tanjak, 2020). Consequently, teachers and students faced the unfavorable prospect of switching from an offline, face-to-face teaching environment to a digital/virtual world (Amin & Sundari, 2020). Therefore, in such a situation, it has become even more essential to determine students' attitudes toward the digital education tools they use. Scales were developed to measure students' attitudes in this area (Croop, 2008; Çelik, 2013; Demir & Akpınar, 2016; Gönülal, 2019; Liu, 2017; Martin & Ertzberger, 2013; Yang, 2012). However, to the best of our knowledge, there are not enough scale adaptation studies that address the different dimensions of the feature to be measured in the context of Türkiye. Accordingly, the present study focused on adapting and validating an attitudinal scale to examine language learners' attitudes toward MALL. In doing so, this study adopted Gönülal's (2019) A-MALL scale measuring attitudes toward MALL.

1.1. Mobile-Assisted Language Learning and Attitudes

MALL is still a new area of investigation. Despite the growing interest in MALL, practitioners need to know more about what it can offer language learning different from traditional techniques. The MALL concept generally refers to a mobile-based approach to language learning that involves the use of portable handheld devices such as tablets, iPads, wireless laptop computers, portable MP3 players, mobile phones, and personal digital assistants (PDAs) to support language acquisition (Chang *et al*., 2018; Gönülal, 2019; Stockwell, 2010).

Palasas (2016) stated that "MALL learns from CALL but cannot be considered as merely a subset of CALL" (p. 45). Similarly, mobile learning is a natural extension of CALL since it incorporates all the benefits of CALL but with fewer time and space restrictions (Jarvis & Achilleos, 2013). In addition, mobile learning has various attributes, including spontaneity, personalization, informality, context, portability, ubiquitousness, and pervasiveness (Kukulska-Hulme & Traxler, 2005). Considering all these features of MALL, learning language items such as words and phrases in a different language with digital devices is essential. Nevertheless, technologies do not directly carry out learning (Jonassen, 1992), learners need to engage in some level of thinking, participation, and attraction to learn.

Understanding students' attitudes toward MALL is crucial for capturing their attention and engaging them in language-learning situations. As stated by Dörnyei (2003), attitude has a significant effect on the learning of a language, as it can either positively or negatively affect the learning process. Thus, several studies have been conducted to investigate the attitudes of teachers and students toward MALL (Alkhudair, 2020; Almudibry, 2018; Anwar *et al*., 2022; Aromaih, 2021; Pham, 2022). To illustrate, using a 21-item scale, Anwar *et al*. (2022) investigated the attitude of 310 female midwifery students toward MALL under six factors (i.e., self-efficacy, anxiety, self-regulation, usefulness, social interaction, behavioral acceptance). While the use of MALL has been shown to have a positive effect on language learning, its effect on anxiety was found to be small. Therefore, Anwar *et al*. (2022) suggested that anxiety must be taken seriously in every aspect of the learning process, whether the device is a MALL or not. Similarly, Pham (2022) investigated 116 university students' attitudes toward the MALL app Quizizz. The results revealed that participants had positive attitudes toward the application, and their satisfaction levels correlated strongly with attitude.

Studies on MALL have also attracted attention in Türkiye and have been the subject of several studies. For instance, Okumuş Dağdeler *et al*. (2020) examined the impact of a mobile application on improving English vocabulary knowledge and found positive short-term effects, but no significant differences in long-term retention or productive vocabulary knowledge. Similarly, the study by Şendağ *et al*. (2019) revealed that mobile extensive listening was less effective compared to teacher-centered intensive listening in enhancing listening skills.

Similarly, Özer and Kılıç (2020) reported positive effects on academic achievement and acceptance of mobile learning tools, though they underlined the need to investigate negative aspects as well. Özsarı and Saykılı (2020) stated that while mobile learning can be actively used for language learning, skills other than vocabulary learning, such as writing and listening, are largely neglected. These studies indicate that the impact of MALL in Türkiye is generally low or ineffective. In contrast, several studies from existing literature have demonstrated the effectiveness of MALL. For example, Solodka *et al*. (2022) showed that MALL supports interaction, communication, and resource access. Pratiwi *et al*. (2023) found significant impacts on learning outcomes in TOEFL preparation classes, albeit with limited effectiveness. Moreover, Phetsut and Waemusa (2022) emphasized the effect of MALL in improving the students' English-speaking skills in Thailand. Therefore, the overall low or ineffective results of MALL studies in Türkiye highlight the need for further research. The current scale may serve as an important tool to investigate why MALL yields negative or ineffective results in Türkiye.

As can be understood from the aforementioned research, attitude is a complex concept that needs to be determined, especially in newly developed learning applications. In Türkiye, a few researchers studied developing or adapting MALL scales. Çam *et al*. (2019) adapted the Mobile Learning Attitude Scale developed by Knezek and Khaddage (2013) to learn about general attitudes towards mobile learning in Turkish culture. In this scale, researchers focus primarily on how mobile technologies are used in educational settings as a whole. Nevertheless, the scale did not specifically address the unique features of MALL, such as its application in language learning situations. The scale does not take into account attitudinal factors like anxiety and motivation although it measures perceived usefulness, effectiveness, perceived control, and behavior. Similarly, Önal and Tanık Önal (2019) translated and validated an English mobile learning attitude scale for adult learners. A major focus of the scale is mobile learning experiences rather than specific attitudinal dimensions like anxiety, self-regulation, or social interaction. Demir and Akpınar (2016) also developed a mobile learning attitude scale that covers issues such as cognitive load and usability. However, this scale does not adequately cover affective factors that are critical for language learning environments, such as motivation and engagement. In their study, they emphasize the general use of mobile technologies in education, but they do not aim to explore the attitudinal factors that influence language acquisition.

This study, in contrast, adapts and validates Gönülal's (2019) A-MALL scale that focuses specifically on attitudes towards MALL as well as its cognitive and affective aspects. Unlike the abovementioned scales, the A-MALL scale addresses the portability, interactivity, and "anytime, anywhere" aspects of MALL, which are essential to language learning. In order to provide a more nuanced understanding of students' attitudes toward MALL, this tool includes detailed subscales that measure factors such as anxiety, self-efficacy, and social interaction. This adaptation study not only improves measurement precision but also contributes significantly to the literature by filling a gap in the cognitive and affective dimensions of MALL, which makes language education research and practice more effective and context-specific in Türkiye. Therefore, the adaptation of this scale to the Turkish language and culture will contribute to future studies in this field.

1.2. Adaptation Research

The adaptation process consists of translation, adaptation, and validation steps. In terms of terminology, adaptation is distinct from translation, and it is usually the former that is used since it refers to all aspects of cultural fit beyond mere translation (Hambleton, 2004). To avoid such confusion and to ensure that the process is carried out appropriately, the International Test Commission has developed guidelines on how psychological instruments should be translated and adapted cross-culturally (ITC, 2017). Further, adapting a scale is a long, demanding process that takes place with the involvement of more than one researcher. According to Hambleton

(2004), the process is so delicate that some researchers have argued that poorly adapted scales ruined their research.

Adapting an existing instrument can be more advantageous than developing a new one tailored to the target population (Borsa *et al*., 2012). The advantages such as time, cost, and effort are important for a researcher. Furthermore, in addition to being able to generalize more readily, the use of adapted instruments also permits analysis of the differences among a more diverse population (Hambleton, 2004). However, as well as its advantages, this process has several disadvantages or risks. For instance, Güngör (2016) stated that although it may seem more economical to adapt a scale whose validity and reliability have been proven in another language, problems such as the lack of measurement equivalence due to translation or cultural differences may arise. To minimize the abovementioned problems, as the International Test Commission suggested, the present study followed Hambleton and Patsula's (1999) guidelines in the adaptation process.

1.3. The Present Study

This study attempted to translate, adapt and validate an attitude toward the MALL scale (see Appendix 2) using an adaptation method. The rationale for adapting the A-MALL scale is in response to the growth of research in the field of MALL in Türkiye and the lack of a scale that measures a feature that has different components such as affective and cognitive aspects.

As Jarvis and Achilleos (2013) suggested, moving from CALL toward a well-supplied MALL, Gönülal (2019) replicated Vandewaetere and Desmet’s (2009) 20-item scale toward CALL and developed a valid and reliable A-MALL scale. During the replication process of Vandewaetere and Desmet’s (2009) study, Gönülal (2019) first performed the Exploratory Factor Analysis (EFA) and then the Confirmatory Factor Analysis (CFA). According to the EFA results, items 7, 16, and 17 were determined as complex variables and removed. Further, CFA results revealed that items 2 and 9 had low factor loadings; therefore, both were removed. Eventually, the final version of the developed A-MALL scale consists of 15 items and five factors (i.e., the effectiveness of MALL, teacher influence, degree of the exhibition to MALL, surplus value of MALL, orientation toward MALL). As in the original questionnaire, Gönülal (2019) used a seven-point Likert scale (1 = totally disagree, 7 = totally agree). All in all, the author's reporting practices and appropriate transparency were deemed to make this study suitable for adaptation in general.

2. METHOD

2.1. Participants

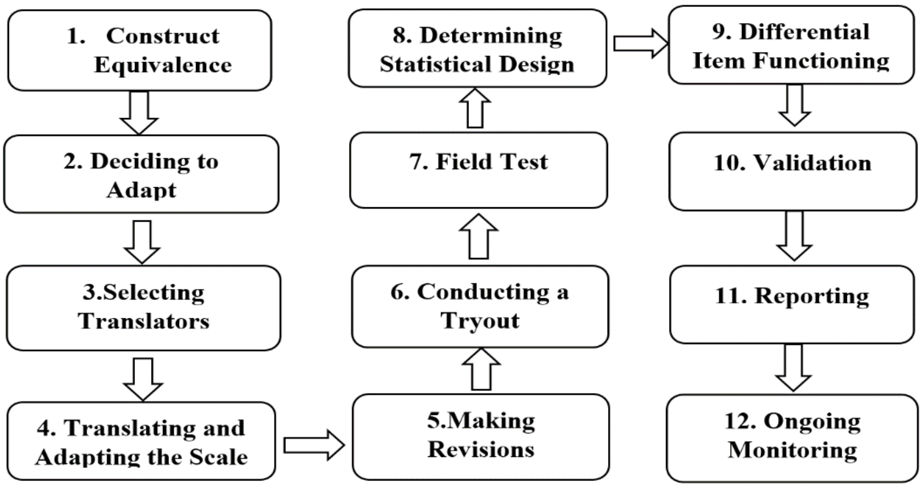
In total, 250 EFL learners in different cities of Türkiye participated in the study. Using Google Forms, the questionnaire was sent to students through the instructors, who reported using MALL applications in their classes. The majority of the participants were female (69.6%), and two participants (0.8%), did not want to indicate gender. The participants’ ages ranged from 17 to 50, and the average age was 22.01 (SD = 5.53). They participated in the study in 49 different cities from Türkiye; Erzurum (23.2%), Trabzon (20.8%), Hakkari (11.6%), Van (5.2%), Diyarbakır (4.4%), Samsun (2%), Batman, Bursa, and Şırnak (1.6%), Adıyaman, Iğdır, Kars, and Siirt (1.2%), to name a few. All participants had a mobile phone; some also had a tablet (21.2%) and a portable music player (9.6%). Participant education levels were as follows: high school (7.2%), associate degree (4%), undergraduate (77.2%), master's degree (6%), and Ph.D. (5.6%).

The participants were informed before completing the questionnaire that their participation was entirely voluntary, their names would not be taken, and the data would only be used for research purposes. The Ethics Committee of Ataturk University approved this research.

2.2. Translation and Adaptation of the Scale

In this study, the questionnaire was cross-culturally adapted in multiple steps following Hambleton and Patsula's (1999) guidelines. An illustration of these steps is provided in Figure 1.

Figure 1. *Illustration of the adaptation process.*



The adaptation process began with ensuring the construct equivalence; that is, the definition of the MALL and its extensions were checked to determine whether they were equally perceived in both languages and cultures. According to the literature review, the terminology used in MALL is universally similar, and adaptability is not a problem. As a next step, we did a review of the literature to find out whether there are any scales aiming to measure attitudes toward MALL in the Turkish literature. To the best of our knowledge, no adapted or developed scales to measure the MALL concept with the desired factor structure were found. However, it is worth noting that only the M-learning Attitude Scale developed by Çelik (2013) has similarities with the characteristics to be measured. Eventually, the A-MALL scale, consisting of 15 items, by Gönülal (2019) was decided to be adapted to Turkish culture and language.

In line with the recommendations of Hambleton and Patsula (1999), well-qualified translators were recruited to translate the questionnaire. First, 15 items were independently translated by two researchers with high proficiency levels in English and whose native language was Turkish. The two translations were compared, and only minor differences were identified in the level of synonymy. Thus, specialists reached a consensus. Afterwards, the translated copy was sent to the Turkish language expert to check for grammatical and semantical errors. According to the Turkish language expert's feedback, there were no semantic or structural problems, and the scale was sent to a scale development specialist to check its face validity. Following confirmation of the scale's positive face validity, the back translation process was initiated. A researcher from the field of English Language Education back-translated the last version of the scale. While comparing the translation copies, it was found that the item content was nearly identical to that of the original scale, with only minor differences identified.

2.3. Procedure

The procedure involved two EFL teachers simultaneously reading aloud the scale to a high school and a university class, and the students in each class were asked to indicate which concepts they did not understand. In response to the participants stating that they understood all of the points, a minor tryout was conducted with the same group of 61 students. While some minor issues were identified, the results suggest that the scale is generally comprehensible and applicable for its intended use. Subsequently, to collect the data and choose the best sampling, EFL teachers and lecturers working in different provinces were interviewed to determine whether they used MALL tools in their classes. Teachers from four cities where the participants were studying indicated that they utilized these tools in their lessons. In response, the teachers were asked to share the scale link, which also contained demographic information about the participants, including age, gender, education level, and mobile devices used by each participant, with their students. The data collection procedure was conducted via the Internet to make a reliable comparison between the collected data and the original data obtained from the calibration sample. Following the completion of the sampling, 250 participants filled in the questionnaire (it took nearly ten minutes), and there were no missing values. Finally, the data were prepared for analysis.

2.4. Data Analysis

2.4.1. *Confirmatory factor analysis*

The goal of a Confirmatory Factor Analysis is to fit the default factor loads as closely as possible to the target matrix (Kline, 2011). Thus, the researchers used the CFA results of the original study as calibration samples for testing the modified model in this study. To determine which probabilistic distribution and parameters best describe the observed data, the Maximum Likelihood Estimation Method was used. The proposed CFA model was evaluated for fit by estimating a number of fit indices such as Chi-square (χ^2), Chi-square divided by the degrees of freedom (χ^2/df), Root Mean Square Error of Approximation (RMSEA), Adjusted Goodness-of-fit Index (AGFI), Comparative Fit Index (CFI), Tucker-Lewis Coefficient (TLI), and Goodness-of-fit Index (GFI) (Kline, 2011; Tabachnick & Fidell, 2013).

3. RESULTS

The AMOS v23 statistical package was used for the CFA. Two hundred fifty samples from EFL students were included in the analysis. According to Kline (2011), a sample size of 200 people is usually sufficient to extract reliable factors. Another common practice is to study with a sample of 3-10 times the number of items (Cattell, 1978; Everitt, 1975). Therefore, the study’s sample size met these conditions with 250 participants. Afterwards, factor loadings were determined, and fit indexes were checked. As for the results of testing the assumptions of the CFA, AMOS v23 was employed to perform the CFA. Prior to conducting the CFA, all assumptions such as the presence of univariate and multivariate outliers, distribution normality, and the absence of multicollinearity were examined and met (Tabachnick and Fidell, 2013). Specifically, boxplots revealed no univariate outliers, and Cook’s distance values, ranging from .00 to .73, fell within the acceptable range of −1 to +1, indicating no significant multivariate outliers. The skewness values, which ranged from −.42 to .15, and the kurtosis values, ranging from −.73 to .30, were both within the acceptable range of −1 to +1, demonstrating that the dataset was normally distributed. Lastly, the Variable Inflation Factor (VIF) values, ranging from 1.82 to 3.77, were below the threshold of 4, suggesting no issues with multicollinearity.

The path diagram in Figure 2 also illustrates the intercorrelations, fit indexes, and factor loadings. As indicated in the path diagram, all the factor loadings are more than .30 and generally, a factor loading greater than .30 indicates that the item and the factor are moderately correlated (Tavakol & Wetzel, 2020). According to the analysis, the following fit indexes were obtained: χ^2/df=2,606, RMSEA=.080, SRMR=.0622, CFI=.954, GFI=.897, AGFI=.846, NFI=.929, TLI=.940.

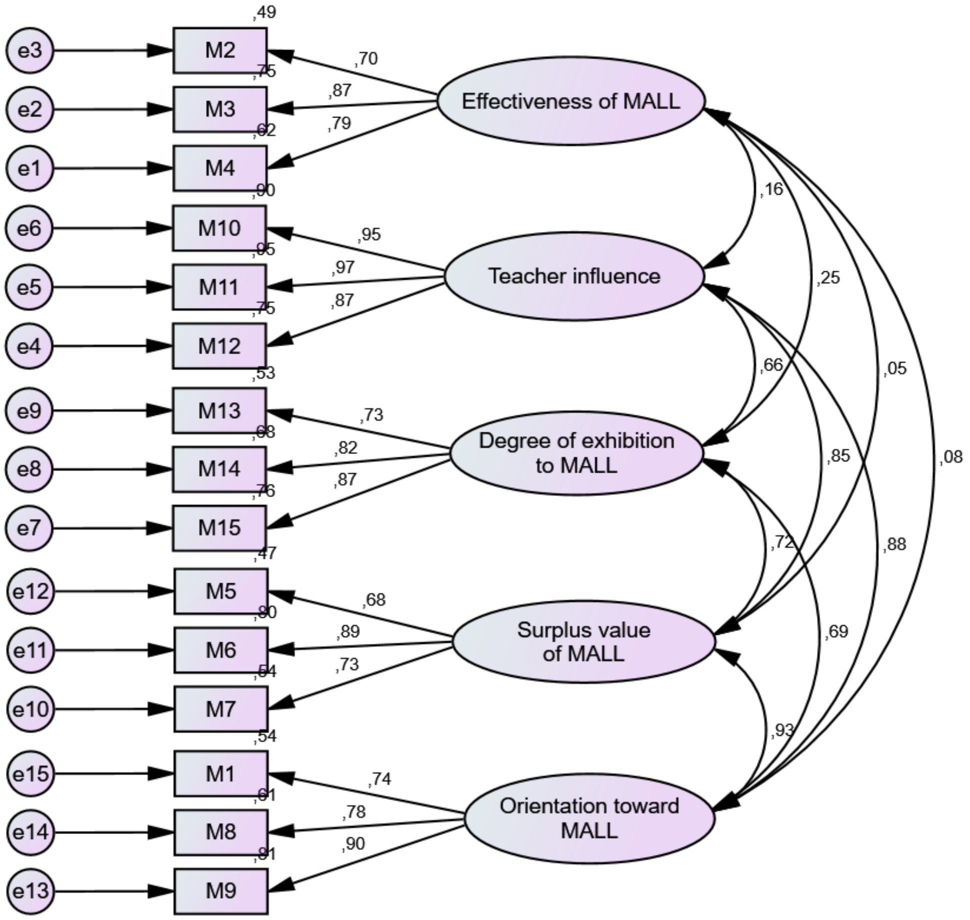
As shown in Table 1, except for AGFI (poor fit), the abovementioned values have a good and acceptable fit to the reference ranges. Consequently, modification indices were not required between variables.

Table 1. *Fit statistics for both calibration (original scale) and validation (adapted scale) samples.*

Index	Current levels		Perfect fit	Good fit	Evaluation
	Calibration	Validation			
χ^2/df	1.49	2.60	$\chi^2/df \leq 2$	$\chi^2/df \leq 3$	Good fit
RMSEA	.064	.080	RMSEA ≤ .05	RMSEA ≤ .08	Good fit
GFI	.88	.897	GFI ≥ .95	GFI ≥ .90	Acceptable fit
AGFI	.82	.846	AGFI ≥ .95	AGFI ≥ .90	Poor fit
CFI	.95	.954	CFI ≥ .95	CFI ≥ .90	Perfect fit
TLI	.93	.940	NNFI ≥ .95	NNFI ≥ .90	Good fit

The fit indices (Hair *et al*., 2010; Hooper *et al*., 2008; Hu & Bentler, 1999; Kline, 2011; Tabachnick & Fidell, 2013; as cited in Gönülal, 2019)

Figure 2. *Path diagram illustrating factor model of adapted A-MALL scale.*



CMIN=208,501; DF=80; P=,000; RMSEA=,080; GFI=,897; TLI=,940; AGFI=,846; CFI=,954; NFI=,929

The default model needs to be checked for validity and reliability in the next step of the CFA. Thus, the original A-MALL scale scores were used as a calibration and compared with the current findings to examine the two concepts better. Further, the Cronbach Alpha coefficient and the Composite Reliability (CR) were calculated to assess reliability. Compared with the original study, the current study produced higher reliability rates (i.e., Cronbach Alpha coefficient ranged from .80 to .94., CR .817-.951). Additionally, the overall Cronbach Alpha coefficient of the adapted A-MALL scale is .917. A Cronbach Alpha coefficient between .80 and 1 is considered highly reliable (Erkuş *et al*., 2017). Similarly, internal consistency reliability greater than .70 indicates good internal consistency (Hair *et al*., 2010). As a result, the adapted A-MALL scale is internally consistent, and comparative values are shown in Table 2.

As a measure of convergent validity, Average Variance Extracted (AVE) helps assess the relationship between factors (Gönülal, 2019). According to Fornell and Larcker (1981), AVE values of more than .5 indicate that the factor is well explained by its items/variables. In the case of the adapted scale, the AVE values fall between .601 - .866, which is higher than the calibration values (i.e., .532 - .757). This suggests that the items within each factor are highly correlated. Furthermore, an Excel tool designed by Gaskin (2011) was used to find discriminant validity measures.

Table 2. *Reliability and validity values of calibration and validation sample (in parentheses).*

Factor	Item	Factor loading	Reliability		Convergent validity
			A	CR	AVE
Factor 1	Item 2	.70 (.70)	.78 (.82)	.793 (.830)	.564 (.621)
	Item 3	.89 (.87)			
	Item 4	.69 (.79)			
Factor 2	Item 10	.84 (.95)	.90 (.94)	.903 (.951)	.757 (.866)
	Item 11	.89 (.97)			
	Item 12	.83 (.87)			
Factor 3	Item 13	.76 (.73)	.79 (.84)	.792 (.851)	.559 (.656)
	Item 14	.74 (.82)			
	Item 15	.73 (.87)			
Factor 4	Item 5	.76 (.68)	.78 (.80)	.804 (.817)	.586 (.601)
	Item 6	.91 (.89)			
	Item 7	.67 (.73)			
Factor 5	Item 1	.61 (.74)	.71 (.84)	.760 (.850)	.532 (.656)
	Item 8	.55 (.78)			
	Item 9	.86 (.90)			

CR composite reliability, AVE average variance extracted (Gönülal, 2019)

The discriminant validity of a construct can be defined as the extent to which those constructs are empirically distinct from one another (Ab Hamid *et al*., 2017). According to Table 3, the adapted A-MALL, as in the original scale, displays good discriminant validity since the square of AVE is greater than the inter-factor correlation. Finally, thanks to the transparency and reproducibility of the study, the order, types, and reporting format of the analysis were chosen to be similar to the original scale for comparison purposes.

Table 3. *Discriminant validity measures for the calibration sample and the validation sample (in parentheses).*

Factor	Factor 4	Factor 1	Factor 2	Factor 3	Factor 5
Factor 4	**.766 (.775)**				
Factor 1	.101 (.054)	**.751 (.788)**			
Factor 2	.086 (.850)	.213 (.160)	**.869 (.930)**		
Factor 3	.329 (.723)	.015 (.246)	.462 (.662)	**.748 (.810)**	
Factor 5	.419 (.928)	.051 (.080)	.479 (.881)	.600 (.692)	**.730 (.810)**

The square root of AVE is given in bold at diagonal

4. DISCUSSION and CONCLUSION

As technology has advanced, people's lifestyles, habits, and needs have evolved, leading to the emergence of new research areas and approaches aimed at meeting these changing needs and demands. One of these areas is language learning, which has seen the shift from Computer-Assisted Language Learning (CALL) to Mobile-Assisted Language Learning (MALL), as mobile devices offer ease of use, spontaneity, flexibility, and privacy. Consequently, it has become crucial to determine the attitudes of students towards MALL. However, there is no existing scale to measure students' attitudes towards MALL in Türkiye that takes into account, in particular, the cognitive and affective aspects of MALL. Although Çelik has developed (2013) a scale named the M-learning Attitude Scale, it was not designed to measure the abovementioned concepts. Therefore, the current study aimed to fill this gap by translating, adapting, and validating Gönülal's (2019) A-MALL questionnaire. Overall, this study contributes to the literature on language learning and technology by providing a comprehensive and context-specific instrument to measure students' attitudes towards MALL in Türkiye.

The adapted A-MALL scale consists of 15 items and five factors as in the original scale. After providing the necessary assumptions, the data collected from 250 English foreign language students were tested by CFA with the scale prepared according to the 7-point Likert type. The original scale data was used as a calibration sample to compare CFA results. Nearly all factor loadings were higher than the calibration sample values. Additionally, Cronbach Alpha coefficients and CR values met the reference ranges. Similarly, discriminant validity tests (i.e., AVE and the square root of AVE) again met the acceptable values. All in all, we adapted a valid and reliable Attitudes towards MALL scale (see Appendix 1).

In order to improve the effectiveness of the language acquisition process and to influence the results of second and foreign language proficiency, empirical research on the possible changes in individuals' learning strategies when using mobile devices in their language learning is required (Viberg & Grönlund, 2013). Therefore, the present study may help increase the empirical research in the Türkiye context and understand the effectiveness of mobile devices in language learning. Moreover, policymakers would benefit from these studies to prepare new language learning programs, develop new web tools, and implement new technological items into the curriculum.

The translated, adapted and validated A-MALL scale, as presented in Appendix 1 can be used to determine the attitudes of foreign language learners towards MALL, especially in terms of its cognitive and affective aspects in the context of Türkiye, and it can help both to increase research in this field and to use these tools in language education. Furthermore, considering the increasing proliferation of mobile technology, language learning may increasingly be integrated into everyday life. In light of this fact, it may be beneficial for all stakeholders within language education to determine students' perspectives on MALL by assessing five different dimensions and three different components of attitude prior to or during the learning process.

Acknowledgments

We would like to express our sincere appreciation to Dr. Talip Gönülal for granting us permission to use the scale in our research.

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**: Atatürk University, Social Sciences and Humanities Ethics Committee, 02.|2.2022-13.

Contribution of Authors

Each author has made an equal contribution to the research.

Orcid

Emirhan Bingöl https://orcid.org/0000-0001-9161-2317
Adnan Taşgın https://orcid.org/0000-0002-3704-861X
Savaş Yeşilyurt https://orcid.org/0000-0001-6871-8842

REFERENCES

Ab Hamid, M.R., Sami, W., & Sidek, M.M. (2017). Discriminant validity assessment: Use of Fornell & Larcker criterion versus HTMT criterion. In *Journal of Physics: Conference Series* (Vol. 890, No. 1, p. 012163). IOP Publishing.

Agca, R.K., & Özdemir, S. (2013). Foreign language vocabulary learning with mobile technologies. *Procedia - Social and Behavioral Sciences, 83*(0), 781-785. https://doi.org/10.1016/j.sbspro.2013.06.147

Alkhudair, R.Y. (2020). Mobile assisted language learning in Saudi EFL classrooms: Effectiveness, perception, and attitude. *Theory and Practice in Language Studies*, *10*(12), 1620–1627. https://doi.org/10.17507/tpls.1012.16

Almudibry, K. (2018). Exploring university EFL learners' experiences and attitudes towards using smart phones for English learning. *The Asian EFL Journal Quarterly*, *20*(6), 347-362.

Amin, F.M., & Sundari, H. (2020). EFL students' preferences on digital platforms during emergency remote teaching: Video Conference, LMS, or Messenger Application?. *Studies in English Language and Education*, *7*(2), 362-378. https://doi.org/10.24815/siele.v7i2.16929

Anwar, K., Wardhono, A., & Budianto, L. (2022). Attitude and social context in MALL classes: A view from midwifery learners. *Cypriot Journal of Educational Sciences*, *19*(9), 3048-3066. https://doi.org/10.18844/cjes.v17i9.7332

Aromaih, A. (2021). University EFL learners' attitudes towards using smart phones for developing language learning skills during the COVID-19 pandemic. *Asian EFL Journal*, *28*(11), 144–160.

Borsa, J.C., Damásio, B.F., & Bandeira, D.R. (2012). Cross-cultural adaptation and validation of psychological instruments: Some considerations. *Paidéia (Ribeirão Preto)*, *22*, 423-432.

Burston, J. (2014). MALL: The pedagogical challenges. *Computer Assisted Language Learning*, *27*(4), 344–357. https://doi.org/10.1080/09588221.2014.914539

Cattell, R.B. (1978). Fixing the number of factors: The most practicable psychometric procedures. In *The scientific use of factor analysis in behavioral and life sciences* (pp. 72-91). Springer, Boston, MA.

Chang, C.C., Warden, C.A., Liang, C., & Chou, P.N. (2018). Performance, cognitive load, and behaviour of technology-assisted English listening learning: From CALL to MALL. *Journal of Computer Assisted Learning*, *34*(2), 105-114. https://doi.org/10.1111/jcal.12218

Croop, F.J. (2008). *Student perceptions related to mobile learning in higher education* [Unpublished doctoral dissertation]. Northcentral University.

Çam, E., Uysal, M., Kıyıcı, M., & İşbulan, O. (2019). Mobil öğrenme tutum ölçeğinin Türk kültürüne uyarlanması [Adaptation of mobile learning attitude scale to Turkish culture]. *International Journal of Turkish Education Sciences*, *7*(13), 114-125. https://doi.org/10.46778/goputeb.408408

Çelik, A. (2013). M-Öğrenme tutum ölçeği: Geçerlik ve güvenirlik analizleri [M-Learning attitude scale: validity and reliability analyses]. *Journal of Research in Education and Teaching*, *2*(4), 172-185.

Demir, K., & Akpınar, E. (2016). Mobil öğrenmeye yönelik tutum ölçeği geliştirme çalışması [Development of attitude scale towards mobile learning]. *Educational Technology–Theory and Practice*, *6*(1), 59-79. https://doi.org/10.17943/etku.83341

Dörnyei, Z. (2003). Attitudes, orientations, and motivations in language learning: Advances in theory, research, and applications. *Language Learning*, *53*(S1), 3–32.

Erkuş, A., Sünbül, Ö., Sünbül, Ö.S., Yormaz, S., & Aşiret, S. (2017). *Psikolojide ölçme ve ölçek geliştirme 2 ölçme araçlarının psikometrik nitelikleri ve ölçme kuramları* [Psychological measurement and scale development 2- Psychometric properties of measurement instruments and measurement theories]. Pegem Akademi. https://doi.org/10.14527/9786053188186

Everitt B.S. (1975). Multivariate analysis: the need for data, and other problems. *The British Journal of Psychiatry: The Journal of Mental Science*, *126*, 237-240. https://doi.org/10.1192/bjp.126.3.237

Fornell, C., & Larcker, D.F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39-50.

Gaskin, J. (2011). "Validity during CFA made easy," *Gaskination's Statistics*. http://youtube.com/Gaskination

Gönülal, T. (2019). The development and validation of an attitude towards MALL instrument. *Educational Technology Research and Development*, *67*(3), 733-748. https://doi.org/10.1007/s11423-019-09663-6

Güngör, D. (2016). Psikolojide ölçme araçlarının geliştirilmesi ve uyarlanması kılavuzu [A guide to scale development and adaptation in psychology]. *Turkish Psychological Articles*, *19*(38), 104-112.

Hair, J., Anderson, R.E., Tatham, R.L., & Black, W.C. (2010). *Multivariate data analysis*. Prentice-Hall.

Hambleton, R.K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Applied Testing Technology, 1(1)*, 1–16.

Hambleton, R.K., Merenda, P.F., & Spielberger, C.D. (2004). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3-38). Psychology Press.

Jarvis, H., & Achilleos, M. (2013). From computer assisted language learning (CALL) to mobile assisted language use (MALU). *The Electronic Journal for English as a Second Language*, *16*(4), 1–18. https://doi.org/10.7575/aiac.alls.v.7n.2p.76

Jonassen, D.H. (2013). Evaluating constructivistic learning. In T. M. Duffy & D. H. Jonassen (Eds.), *Constructivism and the technology of instruction: A conversation* (pp. 137-148). Routledge. https://doi.org/10.4324/9780203461976

International Test Commission. (2017). *The ITC Guidelines for Translating and Adapting Tests* (Second edition). Retrieved January 7, 2023, www.InTestCom.org

Kline, R.B. (2011). Principles and practice of structural equation modeling (3rd Ed). *Guilford*, *14*, 1497–1513.

Knezek, G., & Khaddage, F. (2012). Bridging formal and informal learning: A mobile learning attitude scale for higher education (Version 1). *Deakin University*. https://hdl.handle.net/10536/DRO/DU:30055285

Kolb, L. (2008). *Toys to tools: Connecting student cell phone to education in and out of the classroom*. USA: International Society for Technology in Education.

Kukulska-Hulme, A. (2005). Mobile usability and user experience. In A. Kukulska-Hulme & J. Traxler (Eds.), *Mobile learning: A handbook for educators and trainers* (61-72). Routledge.

Kukulska-Hulme, A. (2009). Will mobile learning change language learning? *ReCALL*, *21*(2), 157–165. https://doi.org/10.1017/S0958344009000202

Liu, T.Y. (2017). Developing an English mobile learning attitude scale for adult learners. *Journal of Educational Technology Systems*, *45*(3), 424-435. https://doi.org/10.1177/0047239516658448

Martin, F., & Ertzberger, J. (2013). Here and now mobile learning: An experimental study on the use of mobile technology. *Computers & Education*, *68*, 76-85. https://doi.org/10.1016/j.compedu.2013.04.021

Okmawati, M. (2020). The use of Google Classroom during pandemic. *Journal of English Language Teaching*, *9*(2), 438–443. https://doi.org/10.24036/jelt.v9i2.109293

Okumuş Dağdeler, K., Konca, M.Y., & Demiröz, H. (2020). The effect of mobile-assisted language learning (MALL) on EFL learners' collocation learning. *Journal of Language and Linguistic Studies*, *16*(1), 489-509. https://doi.org/10.17263/jlls.712891

Önal, N., & Önal, N.T. (2019). İngilizce mobil öğrenme tutum ölçeğinin Türkçeye uyarlanması: Geçerlik ve güvenirlik çalışması [Adaptation of English mobile learning attitude scale into Turkish: Validity and reliability study]. *Journal of Social Sciences of Muş Alparslan University*, *7*(1), 283-289.

Özer, O., & Kılıç, F. (2018). The effect of mobile-assisted language learning environment on EFL students' academic achievement, cognitive load and acceptance of mobile learning tools. *EURASIA Journal of Mathematics, Science and Technology Education*, *14*(7), 2915-2928. https://doi.org/10.29333/ejmste/90992

Özsarı, G., & Saykılı, A. (2020). Mobile learning in Turkey: Trends, potentials and challenges. *Journal of Educational Technology and Online Learning*, *3*(1), 108-132. https://doi.org/10.31681/jetol.670066

Palasas, A. (2013). The ecological perspective on the "anytime anyplace" of mobile assisted language learning. In E. Gajek (Ed.), *Tecnologie mobilne w ksztalceniu jezykowym*, (pp.29-48). Texter.

Pham, A.T. (2022). University students' attitudes towards the application of Quizizz in learning English as a foreign language. *International Journal of Emerging Technologies in Learning*, *17*(19). https://doi.org/10.3991/ijet.v17i19.32235

Phetsut, P., & Waemusa, Z. (2022). Effectiveness of mobile assisted language learning (MALL)-based intervention on developing Thai EFL learners' oral accuracy. *International Journal of Technology in Education (IJTE)*, *5*(4), 571-585. https://doi.org/10.46328/ijte.271

Pratiwi, D.I., Amumpuni, R.S., Fikria, A., & Budiastuti, R.E. (2023). Enhancing students' learning outcomes through MALL in TOEFL preparation class for railway mechanical technology. *International Journal of Language Education*, *7*(2), 185-198. https://doi.org/10.26858/ijole.v7i2.22839

Solodka, A., Ruskulis, L., Demianenko, D., & Zaskaleta, S. (2022). MALL instructional course design: Constructing out-of-class experience. *Arab World English Journal (AWEJ) Special Issue on CALL (8)* 40-55. https://dx.doi.org/10.24093/awej/call8.3

Stockwell, G. (2010). Using mobile phones for vocabulary activities: Examining the effect of platform. *Language Learning & Technology*, *14*(2), 95-110. http://dx.doi.org/10125/44216

Stockwell, G. (2013). Tracking learner usage of mobile phones for language learning outside of the classroom. *Calico Journal*, *30*, 118–136.

Şendağ, S., Gedik, N., Caner, M., & Toker, S. (2019). Mobil destekli dil öğrenmede podcast kullanımı: Öğretici merkezli yoğun dinleme ve mobil kapsamlı dinleme [Use of podcasts in mobile assisted language learning: Instructor-led intensive listening and mobile extensive listening]. *Mersin Üniversitesi Eğitim Fakültesi Dergisi*, *15*(1), 1-27. https://doi.org/10.17860/mersinefd.455649

Tabachnick, B., & Fidell, L. (2013). *Using multivariate statistics*. Pearson Education Inc.

Tavakol, M., & Wetzel, A. (2020). Factor Analysis: a means for theory and instrument development in support of construct validity. *International Journal of Medical Education*, *11*, 245–247. https://doi.org/10.5116/ijme.5f96.0f4a

Thorne, S., & Smith, B. (2011). Second language development theories and technology mediated language learning. *CALICO Journal*, *28*(2), 268–277.

Vandewaetere, M., & Desmet, P. (2009). Introducing psychometrical validation of questionnaires in CALL research: The case of measuring attitude towards CALL. *Computer Assisted Language Learning*, *22*(4), 349-380. https://doi.org/10.1080/09588220903186547

Viberg, O., & Grönlund, Å. (2013). Cross-cultural analysis of users' attitudes towards the use of mobile devices in second and foreign language learning in higher education: A case from Sweden and China. *Computers & Education*, *69*, 169–180.

Yang, S.H. (2012). Exploring college students' attitudes and self-efficacy of mobile learning. *Turkish Online Journal of Educational Technology-TOJET*, *11*(4), 148–154.

APPENDIX

Appendix 1. Mobil Destekli Dil Öğrenimine (A-MALL) Yönelik Tutum Ölçeği.

	Kesinlikle Katılmıyorum						**Kesinlikle Katılıyorum**
1. Dil öğrenimim bir mobil cihaz tarafından desteklendiğinde daha fazla ilerleyecektir.	1	2	3	4	5	6	7
2. Mobil teknoloji tabanlı yapılan dil testleri, asla kâğıt kalemle yapılan testler kadar iyi değildir.	1	2	3	4	5	6	7
3. Mobil destekli dil öğrenimi, geleneksel dil öğreniminden daha elverişsizdir.	1	2	3	4	5	6	7
4. Mobil destekli öğrenim yoluyla bir dil öğrenen kişiler, geleneksel dil öğrenicilerine göre daha yeteneksizdirler.	1	2	3	4	5	6	7
5. Mobil destekli dil öğrenimi, klasik öğrenme yöntemlerinin değerli bir uzantısıdır.	1	2	3	4	5	6	7
6. Mobil destekli dil öğrenimi, dil öğrenimine daha çok kolaylık sağlar.	1	2	3	4	5	6	7
7. Mobil cihaz ile yabancı dil öğrenmek daha rahat ve stressiz bir ortam oluşturur.	1	2	3	4	5	6	7
8. Mobil cihazlarla yabancı dil öğrenmek zekânızı geliştirir.	1	2	3	4	5	6	7
9. Mobil cihazlarla yeni bir dil öğrenmeyi severim.	1	2	3	4	5	6	7
10. Öğretmenin MALL'a karşı tutumu, dil öğreniminde mobil cihazların kullanımına yönelik tutumumu büyük ölçüde etkiler.	1	2	3	4	5	6	7
11. Öğretmenin MALL'a karşı hevesi, dil öğreniminde mobil cihazları kullanma motivasyonumu büyük ölçüde etkiler.	1	2	3	4	5	6	7
12. Öğretmenin dil öğreniminde mobil cihazları kullanma yeterliliği, dil öğreniminde mobil cihaz kullanımına karşı tutumumu büyük ölçüde etkiler.	1	2	3	4	5	6	7
13. Yüz yüze öğrenmeye kıyasla mobil cihazlar aracılığıyla yabancı dilde iletişim kurarken daha az cesaretimin kırıldığını hissediyorum	1	2	3	4	5	6	7
14. Yüz yüze öğrenme durumunda, yabancı dilde konuşmakta sık sık endişe duyarım.	1	2	3	4	5	6	7
15. Benim için yüz yüze bir sohbet başlatmaya karar vermek, mobil destekli sanal bir ortamda sohbet başlatmaya karar vermekten daha zordur.	1	2	3	4	5	6	7

Appendix 2. Attitudes towards mobile assisted language learning (A-MALL) questionnaire.

	Totally disagree						Totally agree
1. My language learning will proceed more when this is assisted by a mobile device.	1	2	3	4	5	6	7
2. Mobile-technology-based language tests can never be as good as paper-and-pencil tests.	1	2	3	4	5	6	7
3. Mobile-assisted language learning is less adequate than traditional language learning.	1	2	3	4	5	6	7
4. People who learn a language by mobile-assisted learning are less proficient than traditional learners.	1	2	3	4	5	6	7
5. Mobile-assisted language learning is a valuable extension of the classical learning methods.	1	2	3	4	5	6	7
6. Mobile-assisted language learning gives more flexibility to language learning.	1	2	3	4	5	6	7
7. Learning a foreign language with a mobile device constitutes a more relaxed and stress-free atmosphere.	1	2	3	4	5	6	7
8. Learning a foreign language by mobile devices enhances your intelligence.	1	2	3	4	5	6	7
9. I (would) like to learn a new language on mobile devices.	1	2	3	4	5	6	7
10. Teacher's attitude towards MALL largely defines my attitude towards the use of mobile devices in language learning.	1	2	3	4	5	6	7
11. The teacher's enthusiasm towards MALL largely defines my motivation for using mobile devices in language learning.	1	2	3	4	5	6	7
12. The teacher's proficiency in using mobile devices in language learning largely defines my attitude towards mobile device use in language learning.	1	2	3	4	5	6	7
13. I feel less inhibited when communicating in a foreign language via mobile devices than in face-to-face learning.	1	2	3	4	5	6	7
14. In a face-to-face learning situation (classroom) I often experience anxiety when speaking in a foreign language.	1	2	3	4	5	6	7
15. For me, the threshold to start a face-to-face conversation is bigger than starting a virtual (mobile-assisted) conversation.	1	2	3	4	5	6	7