e-ISSN: 2148-7456

a peer-reviewed
online journal

hosted by DergiPark

International Journal of Assessment Tools in Education

Volume: 10 Issue: 3 September 2023

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

e-ISSN: 2148-7456

Volume 10 Issue 3 2023

Editor Dr. Ömer KUTLU

Address Ankara University, Faculty of Educational Sciences, Cebeci Yerleşkesi, Çankaya, Ankara, Türkiye

E-mail ijate.editor@gmail.com
omerkutlu@ankara.edu.tr

Publisher Dr. İzzet 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.editor@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

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 Merve SENTURK

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

CONTENTS

Research Articles

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 3, 395–412

https://doi.org/10.21449/ijate.1215747

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

Deconstructing Learner Engagement: An Expanded Construct Model for Higher Education Learners

Meral Şeker [1,*]

[1]Alanya Alaaddin Keykubat University, Faculty of Education, Department of Foreign languages Teaching, Antalya, Türkiye

ARTICLE HISTORY

Received: Dec. 07, 2022
Revised: June 10, 2023
Accepted: June 30, 2023

Keywords:
Higher education,
Learner Engagement,
Metacognitive Engagement,
Engagement Indicators,
Construct Model.

Abstract: Despite the unanimous agreement regarding the positive outcomes of learner engagement, theorists and researchers draw attention to the disparate conceptualizations and structural models of "engagement" construct. The present study, in this respect, attempts to contribute to the development of a theoretical framework by suggesting a multidimensional overarching model for assessing higher education learner engagement. Following the descriptive research design, the study reports the initial model construction and validation results. The findings show significant differences from the earlier conceptualizations indicating a five-dimension model: academic-functional, cognitive, meta-cognitive, collaborative-social, and collaborative- academic engagement. While metacognitive engagement indicators form a distinct but integral dimension in the construct, the social dimension displays an idiosyncratic structure, implying that the multidimensional nature of the engagement construct has a situated nature. Pedagogical implications are discussed based on the engagement model validated.

1. INTRODUCTION

An increasing amount of research has reported the significant role of learner engagement in attaining `success`, which is usually accepted as the ultimate goal of both learners and educational institutions. High levels of learner engagement are found to correlate with numerous positive learning outcomes. For example, engagement is reported to enhance cognitive and metacognitive abilities such as critical thinking; developing practical competence; spending more time and energy on educationally meaningful tasks; learning actively and in collaboration with others and exploring and sharing ideas in and out of class; establishing relationships with the newly learned materials and professional lives (Mazer, 2013). With the help of enhanced cognitive involvement in academic tasks, engaged learners are more likely to exhibit increased performance and productivity (Kuh, 2009; Lam et al., 2012). In addition, engagement is also shown to be a significant predictor in other academic outcomes such as higher graduation and lower drop-out rates (Appleton et al., 2008; Padilla Rodriguez et al., 2020).

Studies have also revealed that engagement is an important mediator between contextual influences and satisfactory psychological and psycho-social states for learners (Fredricks et al.,

*CONTACT: Meral ŞEKER ✉ meral.seker@alanya.edu.tr Alanya Alaaddin Keykubat University, Faculty of Education, Department of Foreign languages Teaching, Antalya, Türkiye

2004; Appleton et al., 2006). Engaged learners are reported to have higher self-esteem and satisfaction rates, to develop a positive identity as a member of the school community and feel connected while they become more confident to establish social relationships and more motivated to participate in extracurricular activities (Lam et al., 2012).

Although the substantial agreement among educators and researchers on the positive outcomes of engagement has proliferated the studies on the concept, theorists and researchers draw attention to the disparate conceptualizations and structural models of "engagement" construct (Aubrey et al., 2020; Dao et al., 2021; Tian & Zhou, 2020). The operationalizations of the research tools developed as a result of incomplete and/or weak conceptualizations and models present inconsistent and questionable findings, which leads to clouding the educational implications to improve learning conditions for higher levels of learner engagement (Kahu, 2013; Krause, 2012; Tian & Zhou, 2020; Zepke, 2014). Furthermore, recent studies report that learners' engagement levels have decreased substantially during compulsory online education because of the Covid-19 pandemic (Chiu, 2022; Yang et al., 2020) and thus, there is a need for instructional interventions to enhance learners' engagement, particularly in online education (Deng et al., 2020; Sun et al., 2020). The present study, in this respect, attempts to contribute to the development of a theoretical framework for engagement construct by proposing a social-constructivist perspective where specific context-related indicators are added to the model structure. It specifically aims to investigate the properties of learner engagement at higher education levels from learners' perspectives while exploring the psychometric qualities of the proposed learner engagement instrument. Following descriptive research design, the study presents the initial construction and validation results of the model. Pedagogical implications are presented based on the engagement model validated.

1.1. Deconstructing Learner "Engagement"

Depending on the base perspective, the concept of learner "engagement" is characterized quite differently. For researchers opting for a psychological perspective, engagement is mainly regarded as an emotional state. Schaufeli et al. (2002, 74), for instance, define engagement as a "... positive, fulfilling, work-related state of mind that is characterized by vigor, dedication, and absorption". The model and the tool developed, Engagement Scale, is based on this conceptualization with a three-dimensional structure involving vigor, dedication, and absorption. Engagement is also regarded as positive or negative feelings towards school such as a state of interest and willingness to participate in learning or negative feelings such as boredom or developing a sense of belonging to the school (Askham, 2008). The psychological perspective, however, fails to account for indicators beyond feeling or emotions and overlooks at cognitive, behavioral or social involvement of learners in learning, and thus, the conceptualizations and models based on the psychological perspective are criticized for their limited account of the construct (Llyod, 2014).

Adapting behavioral perspective, other researchers regard engagement as an effort, time and energy spent or reactions displayed to actively participate in learning activities. Theoretically based on a behavioristic perspective, the National Survey of Student Engagement (NSSE), as one of the most popular learner engagement tools in higher education, was developed in a project in the United States by Kuh (2009) and has been widely used since then. Viewing engagement as a dynamic construct conveying not only learner behaviors but also institutional and teaching practices, it was developed as a measurement tool to identify engagement rates and tendencies of college students to improve education quality (Zhoc et al., 2019). The survey has five scales: academic challenge, active learning, interactions with students and staff, enriching educational experiences, and supportive learning environment (NSSE, 2010). As another popularly used tool, The Australasian Survey of Student Engagement (AUSSE) was developed based on NSSE. AUSSE has added one scale to NSSE, work-integrated learning, to identify engagement in regard with students' career planning (Coates, 2010).

Despite the popularity of these two scales, they are not without criticisms. The main line of the criticisms is related to the way learner engagement is conceptualized. It is claimed that defining engagement within behavioral perspective as "[the] time and the effort students devote to educationally purposeful activities" (AUSSE, 2010, 1) is limited as it does not represent the psychological or the affective dimensions of engagement (Axelson & Flick 2011; Hagel et al., 2012; Kahu, 2013). It is also debated that the scales' domain definition is too broad, which leads to confusion and to questioning the theoretical bases of the items (Zhoc et al., 2019). The NSSE is also found to have intermingled learner engagement as a dependent variable with independent variables such as features related to the learning environment (Lam et al., 2012; Zhoc et al., 2019). Another criticism is directed towards the predictive validity of the survey claiming that the scale's benchmarks show weak correlations with academic success (Hagel et al., 2012) as it fails to acknowledge all the interacting dimensions of the engagement construct. As Kahu (2013) also notes, focusing on a single facet of the construct and overlooking at the other interlinked dimensions results in a limited understanding of this complex construct.

The behavioristic perspective on learner engagement also fails to reflect contextual influences and thus misses the situational and individual factors as well. As Appleton et al. (2006) also suggest, the validation procedures carried out are contingent on the sample from which the data was obtained. That is, engagement is thought to be in a cyclical interaction with contextual variables. Thus, the validity of the operationalizations of these tools in different contexts and the implications drawn are criticized as they do not account for variables such as cultural or linguistic features of the learners and the institutions involved (Glanville & Wildhagen, 2007; Krause, 2012) while leaving the differences in the qualities of different disciplines out (Nelson Laird et al., 2008).

Some researchers view engagement as a combination of behavioral and psychological involvement in academic work (e.g., Appleton et al., 2006; Glanville & Wildhagen, 2007). The models proposed are two-, three-, or four-dimensional. In the two-dimensional models, behaviors and emotions constitute the construct. The three-dimensional models, on the other hand, include behavioral, cognitive, and emotional or affective dimensions (e.g., Fredricks et al., 2004). Others propose four-dimensional models including either an academic component (Appleton et al., 2006) or a social component (Finn & Zimmer, 2012; Zhoc et al., 2019) in addition to behavioral, cognitive, and psychological components.

However, the tools developed based on these multidimensional models have also been questioned, particularly in terms of validity. In fact, researchers have pointed out that such tools need to have a clear distinction between the indicators and the facilitators based on clearly determined criteria to distinguish among the indicators and/or among the facilitators (Fredricks et al., 2004). For example, the model proposed by Appleton et al. (2006), which has a taxonomy for engagement including four subtypes: academic, behavioral, cognitive, and psychological, sets a clear distinction between the indicators and the outcomes of engagement. However, while accounting for the multiple dimensions of engagement, the taxonomy fails to have definite criteria to distinguish among the indicators. For example, while 'credits hours towards graduation' is considered as an academic indicator, 'extra credit options' is regarded as a behavioral indicator. Furthermore, applying self-regulated strategies is categorized under cognitive engagement. However, self-regulation covers not only cognitive involvement but metacognitive, social and affective activation of strategies as well (e.g., Oxford, 2011).

2. METHOD

The present study has been conducted following descriptive research design and reports the initial model construction and validation results of a multidimensional construct model with the aim to contribute to the development of a theoretical framework for assessing higher education learner engagement. The first step to generate the indicators involved an extensive review of literature pertaining to learner engagement in order to examine the conceptualizations and the models presented in the field of educational research as well as to build a theoretical framework in order to guide in the development of the item pool for the scale.

Following the review of relevant literature, semi-structured interview questions were prepared to identify (a) the learners' perceptions of the concept of engagement, (b) their levels of engagement, (c) and how they actualize engagement. The interviews were recorded and the participants were asked to write a composition of 250-350 words on how they define an engaged learner and 21 of them volunteered to participate. They wrote their essays the next day after the interviews and submitted them anonymously. Following verbatim transcription of the recorded data and the first analysis of the learner compositions, systematic content analysis was conducted by the researcher and a fellow researcher separately to identify the emerging themes. Here, in order to identify the degree of agreement between the themes elicited by the two researchers, the inter-coder reliability was measured using Cohen's kappa. The agreement value indicated high reliability (.83) (Cohen, 1969).

After the themes gathered from the literature review and the learner interviews were compared, the next step involved categorizing the common themes under the relevant groups. The themes that were confusing or that seemed too abstract or irrelevant were excluded from the scale. After the items and the dimensions were identified, the accuracy and clarity of the items were revised first by the researcher. Upon the modifications made, two other researchers working at the same university revised the scale: one was an expert in statistics and the other was an expert in educational assessment. After the revisions and the alterations suggested were completed, the questionnaire at this stage had two parts. The first part included five questions related to learners' demographic information: age, gender, their universities and departments. The second part included 45 items in a five-point Likert Scale format, anchored by 'always' (1), 'often' (2), 'sometimes' (3), 'rarely' (4), and 'never' (5). The scale was developed and presented to the participants in Turkish in order to obtain accurate and precise responses and also to avoid any possible language obstacle. The scale was then transformed into Google Forms and was sent to be completed online by the second sample group, which consisted of 496 higher education learners.

2.1. Sampling

The study was conducted with the participation of 554 higher education learners in total formed via convenience sampling method. The inclusion criteria consisted of accessibility, availability at the time of data collection, and consent to participate. For the collection of the data, two different samples were formed. The first sample group consisted of 58 learners at two different state universities while the second sample group involved 496 students at 40 different universities studying at 51 different departments. Both sample groups were previously informed about the research and were invited to participate. The learners who signed the consent form were included in the samples while the necessary ethical permissions were obtained from the Research Ethics Committee of the university.

The first group of participants (n=58) was interviewed previous to the development of the items (indicators of engagement) in order to acquire situational insights into learner engagement. While the extensive review of related research conducted previous to and during the interviews provided theoretical and conceptual perspectives on engagement from various contexts around the world, the data gathered from the interviews with this sample of Turkish higher education learners enabled us to gain a contextual perspective from learners' own perspectives. The second sample consisted of 496 higher education learners who were asked to complete the questionnaire following the development of the indicators. The majority were females (n=292) while male participants constituted the smaller share (n=204). Their ages varied between 18 and 23.

2.2. Data Analysis

Within the aim to explore the psychometric qualities of the scale, analyses were conducted to find out construct validity, reliability in terms of internal consistency, and item distinctiveness. In order to determine the internal consistency, the Cronbach Alpha method was used as the scale has a five-point Likert design. Item-total test score correlation was calculated to identify the item distinctiveness of the scale. As for construct validity, Explanatory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) were conducted. Determining the number of factors is an important decision to make in scale development; thus, in order to determine the number of the factors in this study Horn's Parallel Analysis and MAP (Minimum Average Partial) Analysis were used to guide in identifying the number of factors.

Before conducting the analyses on the data sets, they were analyzed in terms of missing data. As the sets did not have any missing data, no tests were run for missing item issues. The next step involved the identification of outliers. As two participants' responses were found to be outliers, they were excluded from further analyses. Following normality testing of the data sets, the data were analyzed using Lisrel 8.51 Program for CFA and the "psych" package in R program for the other analyses.

It is suggested that the data set obtained from EFA be validated by a different data set, i.e., running EFA and CFA on data gathered from two different groups of samples (Macfarlane et al., 2014). To do this, a large data set could be split randomly into two sets, one of which is used for EFA and the other for CFA. In line with this suggestion, the data set obtained from 496 participants were divided randomly into two sets by including the odd-numbered participants in DataSet1 and the even numbers in DataSet 2. DataSet1 (n=248) was used for EFA while DataSet 2 (n=248) was used for CFA.

3. RESULTS

Previous to EFA analysis, Kaiser–Meyer–Olkin (KMO) Test and Bartlett's Test were run to find out whether the data set was suitable for factor analysis (Table 1).

Table 1. *KMO and Bartlett's test results***.**

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.911
Bartlett's Test of Sphericity	Approx. Chi-Square	1177
	Degree of Freedom	247
	Sig.	0.000

The results displayed in Table 1 reveal that the data set is suitable for EFA (KMO = 0.911; Bartlett's *df* = 247; *p* = 0.00 < 0.05) as KMO value is above .50 (Pallant, 2001). Therefore, EFA was conducted and a six-dimensional construct was obtained (Table 2).

Table 2. *The results of factors loadings and the total variance explained.*

Factor	Eigenvalue	Variance Explained (%)	Total Variance Explained
Factor 1	9.533	30.752	58.685
Factor 2	2.988	9.640	
Factor 3	2.583	8.331	
Factor 4	1.700	5.485	
Factor 5	1.388	4.477	
Factor 6	0.992	3.201	

Table 2 shows the factors obtained as a result of EFA, the variance explained, and the total variance explained by five factors that loaded greater than 1 eigenvalue and were accepted as valid based on K1 method criteria. For the total variance explained, values between 40 % and 60 % are accepted as ideal. EFA analysis results show that the first five factors explained 58 % of the total variance. When the eigenvalues are analyzed, it can be seen that factor 6 loaded just under 1. Therefore, in order to validate and finalize the number of factors, Horn's Parallel Analysis and Velicer's Minimum Average Partial (MAP) Test were used. These methods are used to identify the number of dimensions of a construct when trying to define it, especially for the first time. In Horn's Parallel Analysis, an artificial data set was generated parallel to the original data set to be analyzed using EFA. After EFA was conducted on both the original and the artificial data sets, the eigenvalues obtained for each factor in the two data sets were compared in order to confirm the number of the factors. Accordingly, the factors in the original data with eigenvalues greater than the corresponding eigenvalues of the parallel data were retained and the number of the factors was confirmed.

Table 3. *The results of Parallel analysis*

	Eigenvalue					
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Original Data	9.533	2.988	2.583	1.700	1.388	0.992
Parallel Data	1.782	1.621	1.348	1.334	1.217	1.101

According to the results in Table 3, the eigenvalues of the first five factors in the original data set are greater than the ones in the parallel data set. In Factor 6, the eigenvalue of the parallel data is higher than the one in the original data. As a result, the parallel analysis method suggests that the number of factors is five. Table 4 displays the results of MAP Test.

Table 4. *The results of MAP test.*

MAP Criteria					
Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
0.0542	0.3009	0.0286	0.0277	0.0214	0.0298

As shown in Table 4, in line with the results obtained from EFA and, the parallel analysis, MAP test confirm that the scale has five factors with 31 items in total. Out of 45 items in the original scale, 14 were eliminated for statistical reasons such as having less than .30 item correlation, loading under more than one factor at high levels, or having a low item distinctiveness value (Pallant, 2001). Table 5 displays the factors (n=5) and the items (n=31) with their item distinctiveness and loading values in the final version of the scale.

When analyzing data sets using EFA, the Varimax method is used especially when some of the items have high factor loading values to rotate the data and to form the items in groups to constitute different factors. Thus, rotation using the Varimax method was conducted to anchor the loadings of the factors. Item-total test score correlation is conducted to identify the item distinctiveness of the scale (Pallant, 2001). When the value is over .30, an item is considered to have a good distinctiveness value. According to the results displayed in Table 6, item-total test scores of the items in the scale are between 0,303 and 0,711, which indicates that the items have suitable distinctiveness values.

Table 5. *Item distinctiveness and item factor distribution.*

Items	Item Distinc-tiveness	Factor Loadings				
		Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
I27 I attend my lessons regularly.	.653	.840				
I26 I search for online resources to complete my assignments.	.676	.824				
I31 I complete my assignments on time.	.616	.751				
I29 I follow my teacher's instructions in class.	.678	.670				
I17 I study online to support my lessons.	.627	.606				
I41 I prepare the required materials (e.g., textbooks, tools, etc.) for my lessons before classes start.	.624	.605				
I10 I organize my study environment to concentrate better before starting to study.	.608		.753			
I9 I prepare the necessary lesson materials before starting to study.	.521		.676			
I11 I organize my lesson notes while studying.	.653		.653			
I5 I prepare a study plan before starting to study.	.449		.630			
I12 I try to find links between new learning materials and the previous ones.	.596		.620			
I8 What I learn in my lessons is important for me.	.604		.614			
I4 What I learn in my lessons will be useful for my future career.	.558		.609			
I14 If I have difficulty in understanding the study materials, I try to find alternative ways that can make it easier for me.	.564		.570			
I1 If I miss a class, I study individually to compensate for what I have missed.	.481		.523			
I35 I revise my notes after my classes.	.370			.815		
I30 I study regularly out of class.	.322			.809		
I38 I study regularly for my exams/tests.	.303			.744		
I18 I study regularly not to fall behind my lessons.	.567			.636		
I39 I reread my studying materials whenever I can.	.547			.551		
I42 I attend my classes having completed my assignments and readings.	.545			.815		
I34 I leave the campus as soon as my classes finish.	.711				.801	
I24 I like spending time in the campus.	.381				.776	
I37 I go to the campus only if I have classes.	.380				.738	
I25 I participate in the extracurricular activities (e.g., sports, student clubs, music festivals, etc.) in the campus.	.314				.630	
I23 I feel I belong to my university.	.457				.598	
I15 I come together with my peers to study.	.371					.837
I16 If I have difficulty in my lessons, I ask for help from my friends or teachers.	.539					.835
I22 I like studying with my friends.	.377					.537
I43 I like discussing our lessons performances with my friends and teachers.	.309					.455
I40 We discuss what we learn in our lessons with my friends.	.438					.439

Table 6. *The results for internal consistency of the model and the dimensions.*

Factors		Cronbach Alpha
Factor1	(Academic-Exertive Engagement)	0.887
Factor2	(Metacognitive Engagement)	0.869
Factor3	(Cognitive Engagement)	0.875
Factor4	(Collaborative- Social Engagement)	0.782
Factor5	(Collaborative- Academic Engagement)	0.724
Total Scale		0.870

Table 6 shows that the internal consistency of the whole scale is 0.870 while the values for the factors vary between 0.724 and 0.887, which indicates that the whole scale and the subscales are reliable.

As the final step of the analyses, the five-factor 31-item scale obtained from EFA was tested using CFA. The results were evaluated based on various goodness of fit criteria. Table 7 presents the fit indices of the construct.

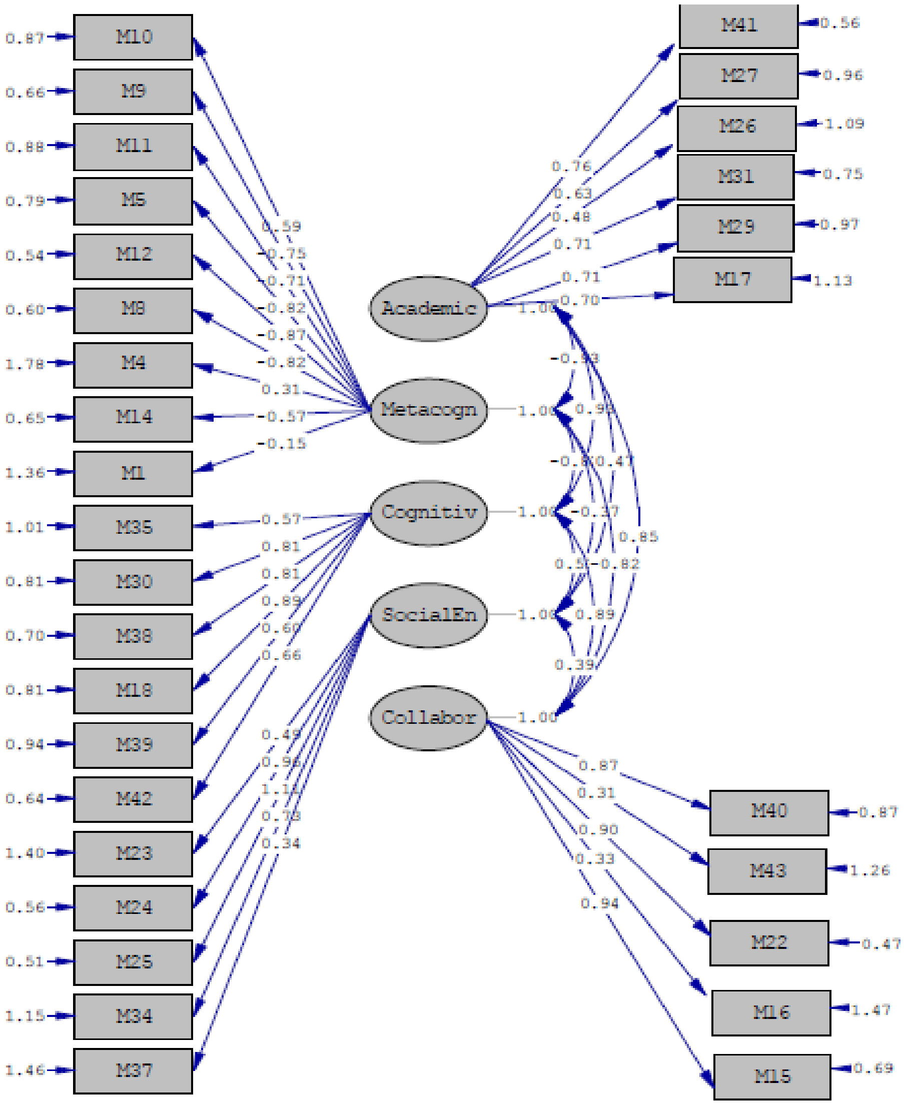
Table 7. *Evaluation of fit indices obtained from CFA.*

Fit Indices		
χ^2 (df)	465 (3)	Acceptable fit
χ^2/df	1.88	Good fit
RMSEA	0.073	Acceptable fit
NFI	0.95	Good fit
NNFI	0.96	Acceptable fit
CFI	0.95	Acceptable fit

The values obtained are within the interval of acceptable fit and good fit (Stevens, 2002). RMSEA value is .073 and the ratio of χ^2 to df is 1.88. The results that the root mean squared error of approximation value is lower than .08 and the ratio of χ^2 to df is lower than 2 indicate a good model fit (Tabachnick & Fidell, 2012). As the majority of the fit indices of the scale show good or acceptable values, the model proposed for higher education learner engagement with five factors is confirmed. The path diagram of the scale obtained from CFA is displayed in Figure 1.

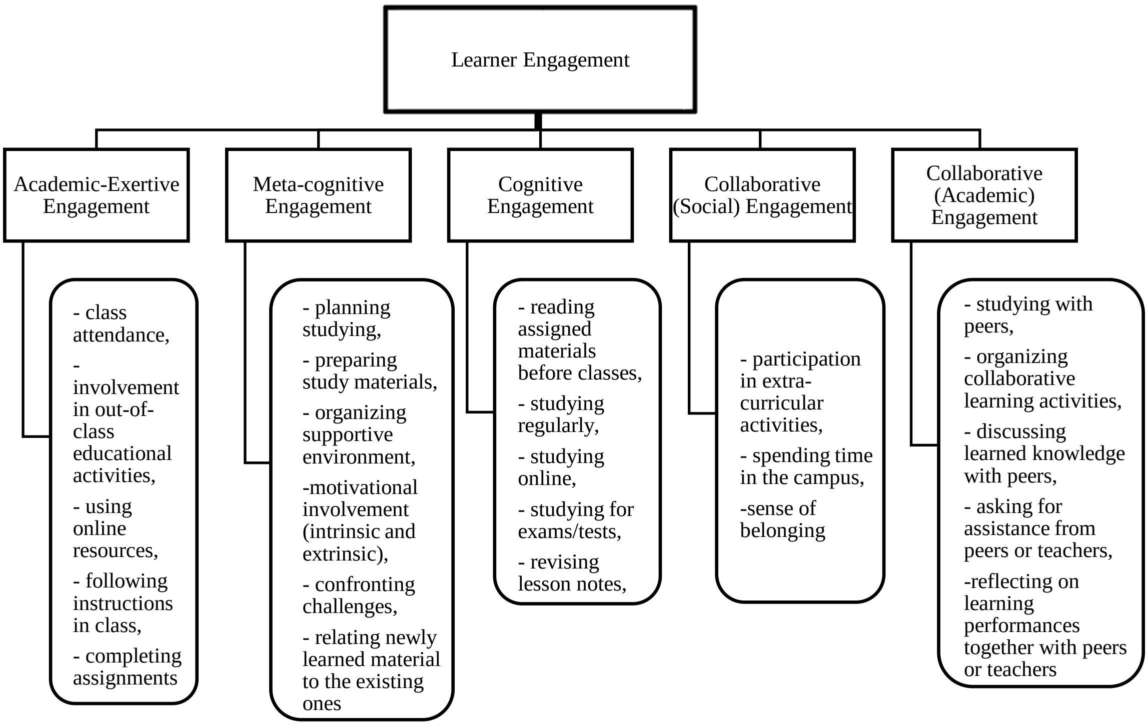
According to the analyses, the psychometric properties of the higher education learner engagement construct revealed a five-dimensional conceptual framework. The model is schematized in Figure 2.

Figure 1. *The path diagram of the model.*



The first dimension “Academic-Exertive Engagement” comprises six items related to class attendance, preparations for lessons, instructions in class, persistence, and completion of assignments. “Metacognitive Engagement”, the second dimension, has nine items related to meta-cognitive efforts such as preparing a study plan, having intrinsic and extrinsic motivation, or compensating for missed classes. Another six items formed “Cognitive Engagement” dimension and addressed different cognitive efforts such as studying regularly or revising lesson notes. While the “Collaborative (Academic) Engagement” dimension included items related to academic social gatherings such as studying with peers or organizing collaborative learning activities, the fifth dimension “Collaborative (Social) Engagement” are towards behavioral or emotional social involvement in campus life.

Figure 2. *The representation of the five-dimensional learner engagement conceptual model*


Learner Engagement
Academic-Exertive Engagement
Meta-cognitive Engagement
Cognitive Engagement
Collaborative (Social) Engagement
Collaborative (Academic) Engagement
- class attendance,
- involvement in out-of-class educational activities,
- using online resources,
- following instructions in class,
- completing assignments
- planning studying,
- preparing study materials,
- organizing supportive environment,
-motivational involvement (intrinsic and extrinsic),
- confronting challenges,
- relating newly learned material to the existing ones
- reading assigned materials before classes,
- studying regularly,
- studying online,
- studying for exams/tests,
- revising lesson notes,
- participation in extra-curricular activities,
- spending time in the campus,
-sense of belonging
- studying with peers,
- organizing collaborative learning activities,
- discussing learned knowledge with peers,
- asking for assistance from peers or teachers,
-reflecting on learning performances together with peers or teachers

4. DISCUSSION and CONCLUSION

The present study attempted to develop a situated model that can guide measuring higher education learners' engagement levels. To this end, the results of this study regarding the initial model construction and validation have yielded a five-dimensions including academic-exertive, cognitive, meta-cognitive, collaborative-social, and collaborative-academic dimensions of learner engagement at higher education level.

The first dimension comprises six items related to "Academic-Exertive Engagement", where learners' behavioral efforts towards educational activities and tasks are grouped. A significant number of models proposed for learner engagement construct incorporate behavioral component (e.g. Appleton et al., 2006; Finn & Zimmer, 2012; Fredricks et al., 2004). Yet, the conceptualization of behavioral component in the models proposed and its indicators included in the measurement tools vary considerably. From a broader perspective, the behavioral dimension is considered to encompass all efforts exhibited in school and towards school work such as attending classes, spending time on tasks, taking an active part in lessons, persistence, participation in academic and out-of-class educational activities as well as in extra-curricular activities (e.g. Appleton et al., 2006; Fredricks et al., 2004). Although some models distinguish between the efforts spent on actual educational tasks like completing assignments or showing persistence in study and the ones devoted to non-academic tasks such as participation in social activities or taking part in student-clubs (Finn & Zimmer, 2012), these indicators are still considered to be within behavioral engagement dimension. The model proposed by Reschly and Christenson (2006), on the other hand, includes efforts spent for academic tasks such as homework completion under academic engagement and categorizes other efforts such as attendance and extracurricular participation under behavioral engagement. In the model proposed in this research, however, academic-functional dimension covers the efforts energized and/or exhibited directly towards academic work, i.e., completing assignments, attending classes, active class participation, using online resources, preparing for lessons, and persistence. Other efforts such as participation in extra-curricular or social activities are grouped under Collaborative- Social Engagement dimension.

The second emerging dimension is “Cognitive Engagement” with six indicators related to cognitive involvements. These are revising notes, studying regularly out-of-class, studying online, trying to keep up with lessons, reading assigned materials, and studying for exams/tests. The indicators in this dimension are related to self-regulation cognitive strategies that are used to learn, process, understand, and remember learning materials.

Within the theoretical frameworks for self-regulation, cognitive regulation has been reported to significantly correlate with engagement (e.g., Cleary & Zimmerman, 2012; Cobos & Ruti-Garcia, 2021). However, cognitive engagement is frequently used to refer “… to the extent and consumption of an intellectual effort that students spent in learning projects (e.g. students’ efforts to incorporate the new knowledge into previously well-known patterns and guide their understanding from a study through the use of cognitive and metacognitive strategies)” (Pellas, 2014, 159). According to this conceptualization, cognitive involvement entails not only the execution of intellectual efforts via utilizing cognitive strategies but the use of meta-cognitive strategies as well. For some models, cognitive engagement indicators also include psychological states or involvement such as motivation or expectation or connecting learning with personal experiences (Kahu, 2013; Pellas, 2014). Furthermore, extra cognitive efforts displayed are considered to be cognitive engagement indicators as well (Finn & Zimmer, 2012). In fact, Zhoc et al. (2019, 225) emphasizes a distinction between academic engagement, which involves “… behaviors exhibited to achieve a minimal ‘threshold’ level of learning …” and cognitive engagement, which “… refers to an internal investment of cognitive energy to attain more than a minimal understanding of the course content”. In other words, efforts displayed to achieve regular or ‘minimal’ learning requirements are indicators of academic engagement whereas efforts to go beyond minimal requirements and to extend learning by facing challenges and enduring learning commitments are accepted to be indicators of cognitive engagement. However, the model fails to clearly explain what is ‘minimal’ and what is ‘beyond minimal’ when it comes cognitive involvement. For example, reading an assigned article could be an easy task for some higher education learners who read such articles regularly and are extremely interested in the topic whereas for other learners who may do better when learning audio-visually or who are not interested in the topic, this task could be quite challenging. Basing cognitive regulation and cognitive strategy use categorization on “easy or complex” tasks would depend on multiple factors such as the specific learner or the duration for assignment completion, etc. For example, the study conducted by Aubrey et al (2020) shows that learners reported higher engagement levels towards easier and more familiar speaking tasks while having lower levels of engagement for unfamiliar tasks or topics. Other studies also report that learners may exhibit different levels of task engagement depending on the task characteristics (Butler, 2017). In the present model, all engagement indicators referring to the exertion of cognitive effort, however simple or complex the cognitive involvement may be, have loaded significantly under cognitive engagement as a distinct factor from meta-cognitive engagement. Also, the findings indicated that indicators of motivational engagement are related to meta-cognitive engagement, rather than belonging to cognitive engagement or forming a separate factor.

“Metacognitive Engagement” formed the third dimension within the construct with nine items related to meta-cognitive efforts activated for learning. Three sub-categories can be found under this dimension: (a) preparative meta-cognitive involvement such as preparing a study plan, or a supportive learning environment, (b) motivational involvement including intrinsic and instrumental motivation, and (c) confronting challenges - persistence such as trying to find alternative ways to learn difficult materials or compensating for missed classes. Although previous studies have indicated that metacognitive involvement is correlated with engagement and has a significant role in predicting learner achievement (Caroll et al., 2021; Hiver et al., 2020; Pellas, 2014), meta-cognitive indicators did not form a distinct dimension in the previous models. Similar indicators have been included in the models and the scales proposed for learner engagement, however, rather than being grouped under a single dimension, these indicators were included in affective (e.g., Kahu, 2013), or in behavioral dimension (e.g., Appleton et al., 2006). For example, in the model developed by Martin (2008), persistence, planning, and study management were three different factors out of 11 dimensions while in the model developed by Appleton et al. (2006), extrinsic motivation was shown to be a separate dimension in learner engagement. The findings in this study, on the other hand, indicate that meta-cognitive engagement is a distinct dimension in the construct including indicators for motivational engagement, preparatory meta-cognitive engagement, and persistence.

The items under the fourth dimension "Collaborative - Academic Engagement" are mostly related to socially shared learning efforts with peers or teachers such as studying with peers or organizing collaborative learning activities, discussing learned knowledge with peers, asking for assistance from peers or teachers, and discussing learning performances with peers or teachers. As the last dimension, "Collaborative-Social Engagement" includes behavioral and emotional involvement in social life on campus such as participation in extra-curricular activities, spending time on campus, and feeling a sense of belonging to the university. Although the last two factors could form a single strand under the "social" dimension, the analyses indicated that they are distinctive properties and that the Turkish higher education context may require a distinction among collaborative engagement indicators as being academically driven or socially (non-academically) driven. The reason for attaining different categorizations in the properties of engagement construct, as Appleton et al. (2006) state, should be the differences in the sample that provided the data and the context. This finding highlights the significant role of contextual factors on engagement, as frequently emphasized in recent research on learner engagement (Aubrey et al., 2020; Sato & Storch, 2020; Sun et al., 2020; Zhang, 2022). Social interaction is highly valued in Turkish culture (Şişman & Turan, 2004) and higher education Turkish learners are found to be keen on working collaboratively for academic tasks (Taşdemir & Yıldırım, 2017) and therefore, being engaged in academic activities could be cohesive to social involvement.

In conclusion, the previous conceptualizations of engagement include indicators either mainly related to informal social involvement such as participation in social activities or interactions with socialization agents on campus or directed solely towards academic interactions such as discussing grades with teachers (e.g., Finn & Zimmer, 2012). The model proposed by Zhoc et al. (2019), on the other hand, includes two sub-categories under the social dimension with indicators for both formal/academic involvements and informal involvements. These two sub-categories are distinguished based on the involved parties in the interactions: Social Engagement with Peers (i.e., informal interactions with peers both in academic and social spheres) and Social Engagement with Teachers (i.e., interactions with academic staff within academic spheres). However, the model obtained in this study indicates a distinction based on the nature of the interaction: social involvement directed towards academic-collaborative activities like studying with peers and social involvement in non-academic activities such as participation in extra-curricular activities in the campus.

Research on learner engagement is particularly vital for higher education institutions in order to optimize learning conditions and opportunities as well as being able to retain the students they already have (Deng et al., 2020; Padilla Rodriguez et al., 2020; Zepke, 2014). However, assessing learner engagement requires defining the construct accurately by considering contextual factors. Assuming that all learners possess similar qualities and exhibit similar behaviors and dispositions across different contexts has led to the misassumption that a single engagement scale could be used for any given context, which also overlooks the interrelated dynamic dimensions of the engagement construct (Aubrey et al., 2020; Kahu, 2013; Zhang, 2022). A considerable majority of the research findings so far point out that engagement is a “meta-construct” embodying multiple constructs (e.g. Zhoc et al., 2019). Determining what these constructs are and understanding how they interrelate with each other will likely to expand our understanding of learner engagement while contributing to the development of engagement pedagogy.

The overall findings of the study indicate two important suggestions to consider when measuring learner engagement. First, since understanding learners and the diverse properties they possess require an extensive consideration of contextual factors and the situated nature of learner behaviors, both contextual and individual factors interacting and shaping learner engagement should be acknowledged. As Zepke (2014, 704) states, “… engagement is more than a ‘one size fits all’ set of ‘how to’ suggestions”. The findings of the present study suggest that metacognitive engagement, which conveys a significant number of engagement indicators that are closely related to contextual factors, forms an integral dimension in engagement construct.

Secondly, the multidimensional nature should be recognized in structuring the concept while accounting for the dynamic interrelation among diverse dimensions (Glanville & Wildhagen, 2007). Rather than trying to draw sharp lines between these dimensions and trying to categorize specific indicators under them, categorizations with broader scopes for each dimension that allow modifications depending on the context at hand could be proposed. This could be achieved by developing an overarching model for engagement in order to have a better understanding of the construct and its role in learning.

The present study is not without limitations. Firstly, the results presented rely on learners’ self-reports and thus, the implications cannot be generalized to the whole population. The model and the tool proposed are subject to further validation through mixed methodological approach. This might entail including in-depth learner perspectives obtained through observations and/or interviews as well as the perspectives of other parties involved in the learning process such as teachers, peers, or administrators. Such a broader scope is expected to yield in more valid results advancing the efforts to understand engagement construct. Also, the model presented includes limited number of indicators for online learner engagement. As online learning has become an integral part of higher education, particularly since the onset of Covid-19 pandemic, the use of digital technologies in education need to be considered as a central part of learner engagement (Deng et al., 2020; Zhoc et al., 2019). Thus, more online engagement indicators should be included in further research conducted to develop models and measurement tools for learner engagement.

Acknowledgments

I am grateful to the team of Le Laboratoire Dynamique Du Langage (DDL) de Lyon for providing a supporting and stimulating working atmosphere as a visiting researcher during the writing process of this paper.

Declaration of Conflicting Interests and Ethics

The author declares no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author. **Ethics Committee Number**: Kocaeli University, Social and Human Sciences Ethics Committee, E-10017888-204.01.07-319718.

Orcid

Meral Şeker https://orcid.org/0000-0001-7150-4239

REFERENCES

Appleton, J.J., Christenson, S.L., & Furlong, M.J. (2008). Student engagement with school: Critical conceptual and methodological issues of the construct. *Psychology in the Schools, 45,* 369–386.

Appleton, J.J., Christenson, S.L., Kim, D., & Reschly, A.L. (2006). Measuring cognitive and psychological engagement: Validation of the student engagement instrument. *Journal of School Psychology, 44,* 427-445.

Askham, P. (2008). Context and identity: Exploring adult learners' experiences of higher education. *Journal of Further and Higher Education, 32,* 85–97.

Australian Council for Educational Research, (2010). Doing more for learning: Enhancing engagement and outcomes. *Australasian Student Engagement Report*. ACER.

Aubrey, S., King, J. & Almukhaild, H. (2020). Language learner engagement during speaking tasks: A longitudinal study. *RELC Journal*. https://doi.org/10.1177/0033688220945418

Axelson, R.D., & Flick, A. (2011). Defining student engagement. *Change*, *43*(1), 38–43.

Carroll, M., Lindsey, S., Chaparro, M. & Winslow, B. (2021). An applied model of learner engagement and strategies for increasing learner engagement in the modern educational environment. *Interactive Learning Environments*, *29*(5), 757-771.

Chiu, T.K. (2022). Applying the self-determination theory (SDT) to explain student engagement in online learning during the COVID-19 pandemic. *Journal of Research on Technology in Education*, *54*(1), 14-30.

Cleary, T.J., & Zimmerman, B.J. (2012). A cyclical self-regulatory account of student engagement: Theoretical foundations and applications. In S.L. Christenson, A. L Reschly & C. Wylie (Eds), *Handbook of research on student engagement* (pp. 237-257). Springer.

Coates, H. (2010). Development of the Australasian survey of student engagement (AUSSE). *Higher Education, 60,* 1–17.

Cobos, R., & Ruiz-Garcia, J.C. (2021). Improving learner engagement in MOOCs using a learning intervention system: A research study in engineering education. *Computer Applications in Engineering Education*, *29*(4), 733-749.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. Academic Press.

Dao, P., Nguyen, M.X.N.C., & Iwashita, N. (2021). Teachers' perceptions of learner engagement in L2 classroom task-based interaction. *The Language Learning Journal*, *49*(6), 711-724.

Deng, R., Benckendorff, P., & Gannaway, D. (2020). Linking learner factors, teaching context, and engagement patterns with MOOC learning outcomes. *Journal of Computer Assisted Learning*, *36*(5), 688-708.

Finn, J.D. & Zimmer, K.S. (2012). Student engagement: What is it? Why does it matter? In S.L. Christenson, A.L. Reschly & C. Wylie (Eds), *Handbook of research on student engagement* (pp. 97-131). Springer.

Fredricks, J.A., Blumenfeld, P.C., & Paris, A.H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*(1), 59–109.

Glanville, J.L., & Wildhagen, T. (2007). The measurement of school engagement: Assessing dimensionality and measurement invariance across race and ethnicity. *Educational and Psychological Measurement*, *67*(6), 1019-1041.

Hagel, P., Carr, R., & Devlin, M. (2012). Conceptualising and measuring student engagement through the Australasian Survey of Student Engagement (AUSSE): A critique. *Assessment and Evaluation in Higher Education, 37*(4), 475–486.

Hiver, P., Zhou, S.A., Tahmouresi, S., Sang, Y., & Papi, M. (2020). Why stories matter: Exploring learner engagement and metacognition through narratives of the L2 learning experience. *System*, *91*, 102260.

Kahu, E.R. (2013). Framing student engagement in higher education. *Studies in Higher Education, 38*(5), 758–773.

Krause, K.L. (2012). Student engagement: A messy policy challenge in higher education. In I. Solomonides, A. Reid, & P. Petocz (Eds), *Engaging with learning in higher education* (pp. 457–474). Libri Publishers.

Kuh, G.D. (2009). The national survey of student engagement: Conceptual and empirical foundations. *New Directions for Institutional Research, 141*, 5–21.

Lam, S., Wong, B., Yang, H. & Liu, M. (2012). Understanding student engagement with a conceptual model. In S. Christenson, A. Reschly, and C. Wylie (Eds), *Handbook of research on student engagement* (pp.403–420). Springer.

Macfarlane, I., Meach, P.M., & Leroy, B. S. (2014). *Genetic counselling research: A practical guide*. Oxford University Press.

Martin, A.J. (2008). Enhancing student motivation and engagement: The effects of a multidimensional intervention. *Contemporary Educational Psychology, 33*(2), 239–269.

Mazer, J.P. (2013). Validity of the student interest and engagement scales: Associations with student learning outcomes. *Communication Studies, 64*(2), 125-140.

National Survey of Student Engagement. (2010). *Major differences: Examining student engagement by field of study: Annual results 2010*. Indiana University Center for Postsecondary Research.

Nelson Laird, T.F., Chen, P.D. & Kuh, G.D. (2008). Classroom practices at institutions with higher than expected persistence rates: What student engagement data tell us". *New Directions for Teaching and Learning, 115*, 85-99.

Oxford, R.L. (2011). *Teaching and researching language learning strategies.* Pearson Longman.

Padilla Rodriguez, B.C., Armellini, A., & Rodriguez Nieto, M.C. (2020). Learner engagement, retention and success: why size matters in massive open online courses (MOOCs). *Open Learning: The Journal of Open, Distance and e-Learning, 35*(1), 46-62.

Pallant, J. (2001). *SPSS survival manual: A step-by-step guide to data analysis using SPSS for windows*. Open University Press.

Pellas, N. (2014). The influence of computer self-efficacy, metacognitive self-regulation and self-esteem on student engagement in online learning programs: Evidence from the virtual world of Second Life. *Computers in Human Behavior, 35,* 157-170.

Reschly, A., & Christenson, S.L. (2006). School completion. In G. Bear, & K. Minke (Eds), *Children's needs III: Development, prevention, and intervention*. National Association of School Psychologists.

Sato, M., & Storch, N. (2020). Context matters: Learner beliefs and interactional behaviors in an EFL vs. ESL context. *Language Teaching Research.* Advance online publication. https://doi.org/10.177/1362168820923582

Schaufeli, W.B., Martinez, I.M., Pinto, A.M., Salanova, M., & Bakker, A.B. (2002). Burnout and engagement in university students: A cross-national study. *Journal of Cross-Cultural Psychology, 33*(5), 464-481.

Şişman, M., & Turan, S. (2004). A study of correlation between job satisfaction and social-emotional loneliness of educational administrators in Turkish public schools. *Eskisehir Osmangazi University Journal of Social Sciences, 5*(1), 117-128.

Stevens, J.P. (2002). *Applied multivariate statistics for the social sciences*. Lawrence Erlbaum Associates.

Sun, Y., Guo, Y., & Zhao, Y. (2020). Understanding the determinants of learner engagement in MOOCs: An adaptive structuration perspective. *Computers & Education, 157,* 103963.

Tabachnick, B.G., & Fidell, L.S. (2012). *Using multivariate statistics*. Allyn & Bacon.

Taşdemir, H., & Yıldırım, T. (2017). Collaborative teaching from English language instructors' perspectives. *Journal of Language and Linguistic Studies*, *13*(2), 632-642.

Tian, L. & Zhou, Y. (2020). Learner engagement with automated feedback, peer feedback and teacher feedback in an online EFL writing context. *System, 91*, 102247.

Yang, L. (2020). Practice and exploration of online teaching during epidemic period. *In 2020 6th International Conference on Social Science and Higher Education,* 420-423.

Zepke, N. (2014). Student engagement research in higher education: Questioning an academic orthodoxy. *Teaching in Higher Education, 19*(6), 697-708.

Zhang, Z. (2022). Learner engagement and language learning: A narrative inquiry of a successful language learner". *The Language Learning Journal*, *50*(3), 378-392.

Zhoc, K.C., Webster, B.J., King, R.B., Li, J.C., & Chung, T.S. (2019). Higher education student engagement scale (HESES): Development and psychometric evidence. *Research in Higher Education*, *60*(2), 219-244.

APPENDIX

Learner Engagement Scale (Turkish)

AKADEMİK KATILIM ÖLÇEĞİ

Sevgili Öğrenciler,

Bu anket akademik çalışmalarınızdaki akademik katılım seviyenizi ölçmek için hazırlanmıştır. Çalışmaya katılım gönüllülük esasına dayalıdır. Katkılarınız için teşekkür ederiz.

A. Kişisel Bilgiler

1. Cinsiyetiniz
   - o Kadın
   - o Erkek
   - o Söylememeyi tercih ediyorum
   - o Diğer:

2. Yaşınız
   - o 17 - 19
   - o 20 - 25
   - o 26 - 30
   - o 31 +

3. Bölümünüz ………………………………………..

B. Akademik Katılım Maddeleri

Lütfen aşağıdaki ifadeleri dikkatli okuyun ve size en uygun seçeneği işaretleyin.

1. Her zaman **2.** Sık sık **3.** Bazen **4.** Nadiren **5.** Hiçbir zaman

No	**Ölçek Maddeleri**	**1**	**2**	**3**	**4**	**5**
1	Eğer bir dersime katılamazsam, o derste kaçırdığım konuları öğrenmek için kendim çalışırım.					
2	Derslerde öğrendiklerim gelecekteki kariyerim için önemlidir.					
3	Ders çalışmaya başlamadan önce kendime bir çalışma planı hazırlarım.					
4	Derslerimi desteklemek için online (çevrimiçi) çalışırım.					
5	Derslerde öğrendiklerim benim için önemlidir.					
6	Ders çalışmaya başlamadan önce gerekli çalışma materyallerini hazırlarım.					
7	Daha iyi yoğunlaşabilmek için, ders çalışmaya başlamadan önce çalışma ortamımı düzenlerim.					
8	Ders çalışırken ders notlarımı düzenlerim.					
9	Yeni öğrendiklerim ile daha önce öğrendiklerim arasında ilişki kurmaya çalışırım.					
10	Çalışma konularını anlamakta güçlük çekersem, anlamamı kolaylaştıracak alternatif yollar ararım.					
11	Arkadaşlarımla ders çalışmak için bir araya gelirim.					
12	Derslerimde güçlük çekersem, arkadaşlarımdan veya öğretmenlerimden yardım isterim.					
13	Derslerimde geri kalmamak için düzenli olarak ders çalışırım.					
14	Arkadaşlarımla birlikte ders çalışmayı severim.					
15	Kendimi üniversiteme ait hissederim.					
16	Kampüste vakit geçirmek hoşuma gider.					

17	Kampüsteki müfredat dişi etkinliklere (örneğin spor, öğrenci kulüpleri, müzik festivalleri, vb.) katılırım.					
18	Ödevlerimi tamamlamak için çevrimiçi kaynaklar ararım.					
19	Derslerime düzenli olarak katılırım.					
20	Derslerde öğretmenlerimin açıklamalarını takip ederim.					
21	Okul dışında düzenli olarak ders çalışırım.					
22	Ödevlerimi zamanında tamamlarım.					
23	Derslerim biter bitmez kampüsten ayrılırım.					
24	Derslerden sonra ders notlarımı gözden geçiririm.					
25	Kampüse sadece derslerim için giderim.					
26	Sınavlarıma düzenli olarak çalışırım.					
27	Her müsait olduğum zaman ders notlarımı tekrar gözden geçiririm.					
28	Arkadaşlarımla derslerde öğrendiklerimiz üzerine konuşuruz.					
29	Dersler başlamadan önce gerekli ders materyallerini (örneğin ders kitapları, ders araçları, vb.) hazırlarım.					
30	Derslere ödevlerimi ve okumalarımı tamamlamış olarak katılırım.					
31	Arkadaşlarım ve öğretmenlerimle derslerdeki performanslarımız üzerine konuşmayı severim.					

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 3, 413–433

https://doi.org/10.21449/ijate.1218659

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

Examining the Factors Affecting Students' Science Success with Bayesian Networks

Hasan Aykut Karaboğa [1,*], **Ibrahim Demir** [2]

[1]Amasya University, Faculty of Education, Department of Educational Sciences, Amasya, Türkiye
[2]Turkish Statistical Institute, Ankara, Türkiye

ARTICLE HISTORY

Received: Dec. 14, 2022
Revised: May 11, 2023
Accepted: Aug. 11, 2023

Keywords:
Bayesian Networks,
Educational Data Mining,
PISA,
Science Success.

Abstract: Bayesian Networks (BNs) are probabilistic graphical statistical models that have been widely used in many fields over the last decade. This method, which can also be used for educational data mining (EDM) purposes, is a fairly new method in education literature. This study models students' science success using the BN approach. Science is one of the core areas in the PISA exam. To this end, we used the data set including the most successful 25% and the least successful 25% students from Turkey based on their scores from Program for International Student Assessment (PISA) survey. We also made the feature selection to determine the most effective variables on success. The accuracy value of the BN model created with the variables determined by the feature selection is 86.2%. We classified effective variables on success into three categories; individual, family-related and school-related. Based on the analysis, we found that family-related variables are very effective in science success, and gender is not a discriminant variable in this success. In addition, this is the first study in the literature on the evaluation of complex data made with the BN model. In this respect, it serves as a guide in the evaluation of international exams and in the use of the data obtained.

1. INTRODUCTION

The world is constantly changing with technological developments. In today's technology-oriented society, especially success in science is directly related to understanding and applying basic scientific knowledge and ensuring the scientific progress of the country by utilizing science and technology in daily life (OECD, 2019a, 2019b). People in 21st century, have to solve a continuous series of daily problems for living in the today's world (Gilbert et al., 2000). International exams such as Program for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS) measure and compare the science successes of the countries.

PISA has an important impact on educational systems and policies (Deng & Gopinathan, 2016). A functional and well-structured educational system is the way to increase achieving future goals set by a country (Sağlam & Aydoğmuş, 2016). The development levels of societies are closely related to the education their students receive. Quality education increases career opportunities, affects economic and cultural development and helps people to increase their

*CONTACT: Hasan Aykut Karaboğa ✉ h.aykut.karaboga@amasya.edu.tr 🖃 Amasya University, Faculty of Education, Department of Educational Sciences, Amasya, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

social status. Determining the factors that enhance the quality of education and improving those factors influences the international success of a country.

International exams measure many factors that affect student success. These exams enable educational authorities in countries to not only assess students' achievements against basic educational standards but also compare the results with those of other countries (Gamazo & Martínez-Abad, 2020; Schleicher, 2019). Assessments in international exams help us explore the relationships between student achievement and students themselves, as well as between student achievement and both schools and education systems. By identifying the factors that influence student success, stakeholders can take necessary steps to increase the levels of low achievement in education (Aşkın & Öz, 2020). It is important to determine the functioning or problematic parts of the education system according to scientific data. Thus, shaping future education policies according to the available data will increase the quality of education (Üstün et al., 2020). Researchers have conducted numerous modeling studies using data derived from international exams. Furthermore, the literature has explored various studies examining the factors that influence academic success (Altun & Kalkan, 2019; Carnoy et al., 2015; Chen et al., 2019; Gamazo & Martínez-Abad, 2020; Karakoç Alatlı, 2020, 2020; Kilic Depren, 2018; Kiray et al., 2015; Kjærnsli & Lie, 2004; Lee & Shute, 2010; Rastrollo-Guerrero et al., 2020; Sebastian et al., 2017; Sheldrake et al., 2017; Sirin, 2005; Tang & Zhang, 2020; Topçu et al., 2015; Torrecilla Sánchez et al., 2019; Yip et al., 2004; Yıldırım, 2012). However, researchers face difficulty in modeling the complex relationships (Kiray et al., 2015; Lee & Shute, 2010) between the variety of factors that impact success (Martínez Abad & Chaparro Caso López, 2017). Choosing the appropriate modeling strategy ensures that the findings are a source for educational systems (Aşkın & Öz, 2020).

Successful modeling results were obtained using EDM. The biggest advantages of EDM methods are that they can work with complex related data sets and do not have restrictive statistical assumptions such as variance homogeneity and linearity (Sinharay, 2016). EDM, the use of classical data mining techniques in education (Baker & Yacef, 2009; Romero & Ventura, 2010; Shin & Shim, 2021), provides practical information for educational policy makers and researchers in increasing success (Peña-Ayala, 2014; Romero & Ventura, 2010). In this study, Bayesian Networks (BN), an EDM approach used in few studies in the educational literature, was preferred to model the relationships of variables affecting high and low science success scores. BN create graphical models of the dependency relations of all variables (Nielsen & Jensen, 2009). Unlike other machine learning models, BN enables queries which explicitly reveal variables’ cause-effect relationships (Pearl, 2014). So, BN provides an advantage over other machine learning methods in revealing complex relationships (Karaboga et al., 2021).

1.1. Research Problem

According to the PISA 2018 results, Turkey increased its performance in all fields compared with the 2015 results, but remained below the average of the OECD countries. Based on the 2018 PISA results, Turkey has an average of 468 points in science and OECD average is 489 points. The country became 39th in science with this score. However, Turkey ranked 54th among 72 countries in science in PISA 2015. According to the OECD's report, Turkey was the only country to experience improvement in all three areas, in despite of the number of students in the 15-year-old group increased significantly between 2003 and 2018 (MEB, 2019).

Despite the significant structural changes made in the Turkish education system, the desired level of success could not be reached in international exams such as PISA. For this reason, in order to achieve a higher performance, factors influencing success should be determined and Turkey should focus on areas of improvement to obtain better results from international exams for years to come. We used the BN model to identify key factors for improving students' science success in international exams, as we believe it is an effective tool for this purpose.

BNs are used in different applications in education literature. Researchers have used BN for various purposes in EDM such as predicting student proficiency levels (Almond & Mislevy, 1999; Desmarais & Baker, 2012), predicting course performance (Xing et al., 2021; Zwick & Lenaburg, 2009), smart classroom applications (Saini & Goel, 2019), evaluation of intelligent tutoring systems (Ramírez-Noriega et al., 2021), student knowledge assessment system (Levy, 2016; Millán et al., 2013; Xing et al., 2021), cognitive diagnostic modeling (Almond et al., 2007) and educational assessment (Culbertson, 2016). But, only one BN study using the PISA data to investigate the relationship of influential factors with mathematics achievement conducted by Tingir and Almond (2017) was found. However, we could not find any study in the literature on science achievement.

This study seeks to contribute to the literature on modeling science success by using BN with pre-determined variables. Initially, a BN model was developed with these variables, but it was found to be difficult to interpret the model structure and the interrelationships between variables. In order to create a more comprehensible and practical model, we utilized feature selection to identify the most effective variables and reduce their total number. It is noteworthy that BN with feature selection has not been previously applied to PISA data to determine the factors influencing science success and their interrelationships. As such, we conducted a comparative analysis of two distinct BN models for science success using PISA 2018 Turkey data, with one model including all variables and the other model including only selected features. This study is the first of its kind to examine the combined use of feature selection and BN modeling in evaluating science success. The BN model we developed evaluates student science success in PISA 2018 by modeling complex relationships among science-effective variables. Compared to other statistical models, BN offers advantages by transforming complex relationships into interpretable knowledge. We argue that our research will have significant implications for evaluating studies using educational data sets.

1.2. Research Focus

In this study, the following research questions guided the study to investigate the factors affecting the science success of Turkish students in PISA 2018:

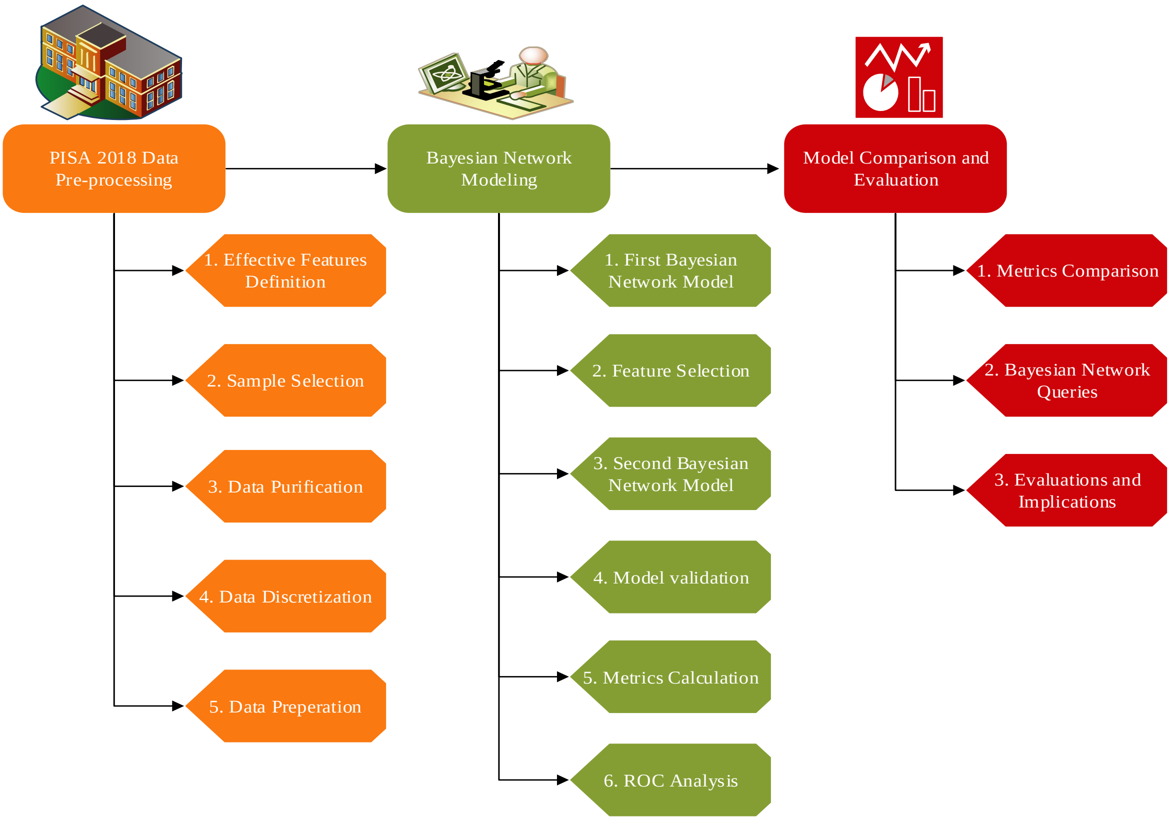
1. What are the factors affecting students' science success?
2. Do the factors affecting students' science success differ according to their high and low success levels? If so, what are the factors that increase success?
3. Is there any performance difference between the two models in terms of science success?
4. Are there any performance differences between male and female students in models?
5. What measures will raise the success level of Turkish students in science?

2. METHOD

In this section, we summarized the sample and explain the steps of the BN design with feature selection methods used to model science success with success-related feature interactions. We describe the implementation stages of the study in Figure 1.

In the first step, we preprocessed the data for BN modelling. The second step, we created the first BN model after the determination of the data related to science success. In the next step, we evaluated the relations in the BN model and obtained results. Considering the obtained results, the next step was feature selection. After this step, we created a second BN model with the selected features. In the last step, we compared the obtained results for these two models and evaluated their implications.

Figure 1. *Graphical Representation of Research Process.*


PISA 2018 Data Pre-processing
1. Effective Features Definition
2. Sample Selection
3. Data Purification
4. Data Discretization
5. Data Preperation
Bayesian Network Modeling
1. First Bayesian Network Model
2. Feature Selection
3. Second Bayesian Network Model
4. Model validation
5. Metrics Calculation
6. ROC Analysis
Model Comparison and Evaluation
1. Metrics Comparison
2. Bayesian Network Queries
3. Evaluations and Implications

2.1. Research Sample

The OECD has conducted PISA every 3 years since 2000. The PISA test consists of school, student, and teacher questionnaires. We employed student's school, family and individual evaluations in our study. In the questionnaire data set, besides demographic variables such as gender and age of the student, there are also indexes such as socio-economic status, family wealth, highest parental occupational status which were constructed through Item-Response Theory (OECD, 2019a).

In statistical analysis, missing values should be excluded from the dataset. Thus, 4276 students who remained after the missing values were eliminated. After that, 2138 students representing the most successful 25% and the least successful 25% were included in the analysis. In the analysis, personal variables such as the students gender and study time as well as school and family variables were used. In order to work with BN, the data was discretized (Nojavan et al., 2017; Yang & Webb, 2002). Modeled variables are shown in Table 1.

Table 1. *Student-Related Model Variables*

	Code	Label	Variable Type
Q1	ATTLNACT	Attitude toward school	*Student*
Q2	BEINGBULLIED	Student experience of being bullied	*Student*
Q3	BELONG	Sense of belonging to school	*Student*
Q4	CLSIZE	Class size	*School*
Q5	COMPETE	Competitiveness	*Student*
Q6	CREACTIV	The index of creative extracurricular activities at school	*School*
Q7	EDUSHORT	The scale of the shortage of educational material	*School*
Q8	EMOSUPS	Parents' emotional supports perceived by students	*Parental*
Q9	ESCS	Economic, social and cultural status index	*Parental*
Q10	EUDMO	Eudaemonia: meaning in life	*Student*
Q11	GFOFAIL	The general fear of failure	*Student*
Q12	HISCED	Highest parental education	*Parental*
Q13	HISEI	Highest parental occupational status index	*Parental*
Q14	IC150Q03HA	Digital devices using time during science lessons (In a typical week)	*School*
Q15	ICTHOME	ICT available at home	*Parental*
Q16	MASTGOAL	Mastery goal orientation	*Student*
Q17	PARED	Highest parental education in years of schooling index	*Parental*
Q18	PERCOMP	Perception of competitiveness at school	*Student*
Q19	PERCOOP	Perception of cooperation at school	*Student*
Q20	RESILIENCE	Resilience	*Student*
Q21	ST004D01T	Student gender	*Student*
Q22	STAFFSHORT	The scale of staff shortage	*Student*
Q23	STRATIO	The student-teacher ratio	*School*
Q24	STUBEHA	Student behavior hindering learning	*Student*
Q25	SWBP	Subjective well-being: Positive effect	*Student*
Q26	TEACHBEHA	Teacher behavior hindering learning	*School*
Q27	TEACHINT	Perceived teacher interest	*School*
Q28	TMINS	Total learning time (minutes per week)	*Student*
Q29	SUCCESS	Science Success (lowest 25% and highest 25%)	*Dependent*

2.2. Bayesian Network

Bayesian Networks (BN) are statistical models that graphically display the common probability distributions of variables in addition to their dependency relations of variables (Nielsen & Jensen, 2009). In a BN model, variables are represented as nodes and relationships between variables are represented as edges. Edges are oriented as one-way arrows and indicate the structure of the network. Structure of the BN is specified as DAG (Directed Acyclic Graph) (Neapolitan, 2009). The established DAG structure can be used to make inferences on the parameters of the model using mathematical equations. In other words, the most important feature of BN is the ability to update the probabilities of each node in the entire model with new information (Sener et al., 2019).

As a graphical model, the DAG structure is shown as ***G=(A, B)***, where ***A*** is the set of nodes and ***B*** is the set of edges that provides the nodes' connections. In a BN -containing the variable ***M***-each node ***X*** is associated with the conditional probability distribution of the corresponding variable considering its parents. The conditional probability of a node is given in Equation 1. This probability value is called conditional probability distribution when the $pa(X_i)$ values of the X node are given.

$$p\left(X|pa\left(X_i\right)\right) \qquad \left(i=1,\ldots,M;M\in A\right) \tag{1}$$

The joint probability distribution calculated for the $(X_1,...,X_M)$ nodes in the whole model is given in Equation 2.

$$p\left(X_1,\ldots,X_M\right)=\prod_{i\in A}p\left(X_i|pa\left(X_i\right)\right) \tag{2}$$

The contribution of variables to the model originates from the conditional probability values, which are calculated when $pa(X_i)$ is given.

2.3. Feature Selection

Feature selection methods play an important role in machine learning, particularly in situations where the number of features is high relative to the number of observations. The feature selection aims to identify the most relevant and informative subset of features, which can improve the model's accuracy, reduce overfitting, and enhance interpretability. In this study, we used correlation-based feature selection named the CFS subset algorithm.

Correlation is one of the most important indicators showing the relationship between two variables. One popular feature selection algorithm is the Correlation-based Feature Selection (CFS) algorithm. CFS subset algorithm was introduced by Hall (1999a). CFS is a filter method that evaluates the features based on their correlation with the class variable and with each other. CFS aims to identify features that are highly correlated with the class variable while minimizing redundancy among the features. The CFS algorithm works by calculating a merit score for each feature, which is based on the correlation between the feature and the class variable, as well as the correlation between the feature and the other features. The merit score is used to rank the features, and a subset of the top-ranked features is selected. CFS is effective in improving the performance of machine learning algorithms by reducing the number of irrelevant and redundant features. This algorithm selects features with low correlation between them and high correlation between class tags (Hall, 2000). The CFS selection coefficient - the equation is the standardized Pearson correlation of all variables- was calculated for each subset (Hall, 1999b).

2.4. Classification Criteria

In this study, accuracy, F-Measure, Mean Absolute Percentage Error (MAPE), Kappa (κ), Root Mean Square Error (RMSE), and ROC area were used in the evaluation of model performance. Accuracy is the overall correct classification rate in the positive and negative cluster which is one of the most common performance measure (Ferri et al., 2009). F-measure is the harmonic mean of correctly classified positive and negative values (Hossin & Sulaiman, 2015). The Kappa coefficient deals with the prediction performance of an algorithm. The closer the Kappa coefficient is to 1, the higher predictive performance of the model. The MAPE value could measure the difference between the expected and predicted results. The MAPE value of models with high predictive performance converges to zero. The RMSE is a quadratic metric that measures the magnitude of error by finding the distance between predicted and actual values. RMSE is a measure of how far these errors are propagated.

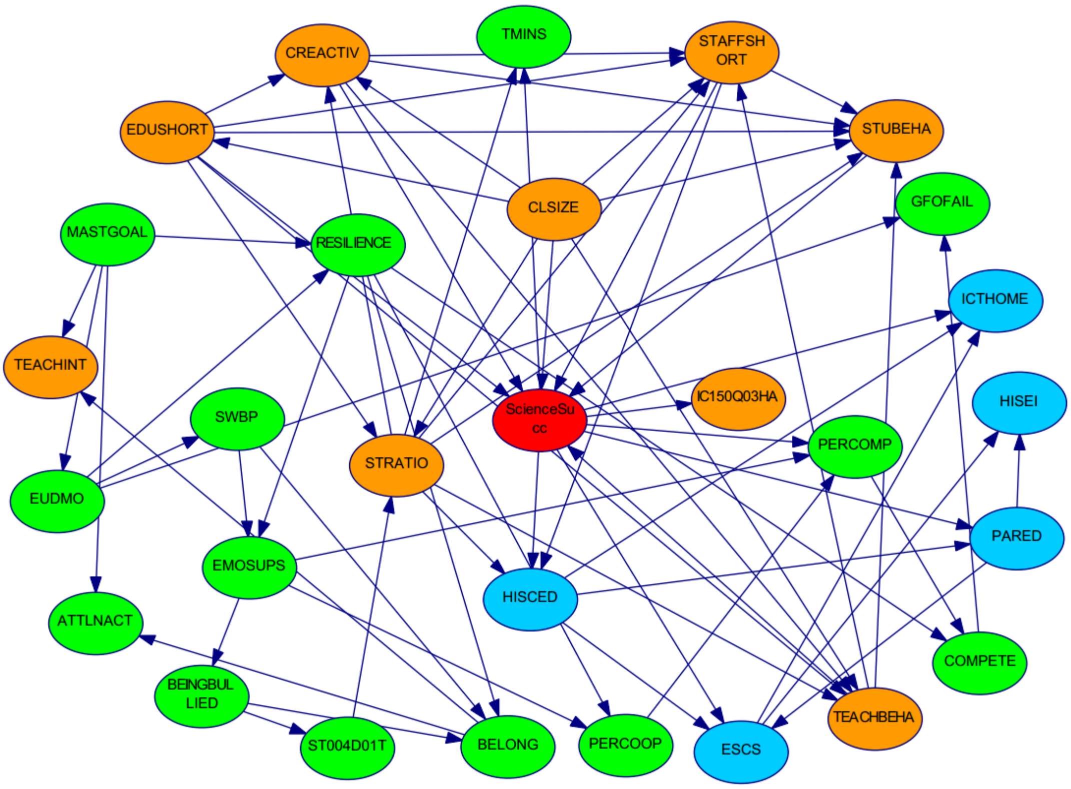
The ROC area namely ‘Area under the receiver operating curve (AUC) value’ measures the ability of the model to avoid errors during class estimation. The AUC is closely related to specificity and sensitivity values. This value is a measure used in conjunction with the ROC curve to show whether a perfect classification has been made (Marsland, 2015).

3. RESEARCH RESULTS

It is not only variables related to students themselves that affect their science success but family and school-related variables are closely related to their success (Kilic Depren, 2018). Variables are grouped under three sub-headings: variables about the student themselves, variables about his/her family, and variables about his/her school. In the first stage, these variables are discrete to be used in Bayesian networks. Therefore, we preferred the quantile discretization method commonly used in discretizing (Lima, 2014; Ropero et al., 2018).

We utilized academic version of the GeNIe program for BN modeling and we preferred k-fold cross-validation method for model evaluation (BayesFusion, 2017). The quartile values were used to discretize the variables. Thus, the variables were represented in 4 different ways from Q1 to Q4 (very low, low, high, very high). The greedy tick thinning algorithm was used in the analysis. In this technique, the data set is divided into k parts; k-1 parts of the data are used for training and the other part of the data are used for testing (Wong, 2015). Finally, we obtained the classification success performance by calculating the mean error of the k tests pieces (Karaboga et al., 2021). In this study, the k value was taken as 10.

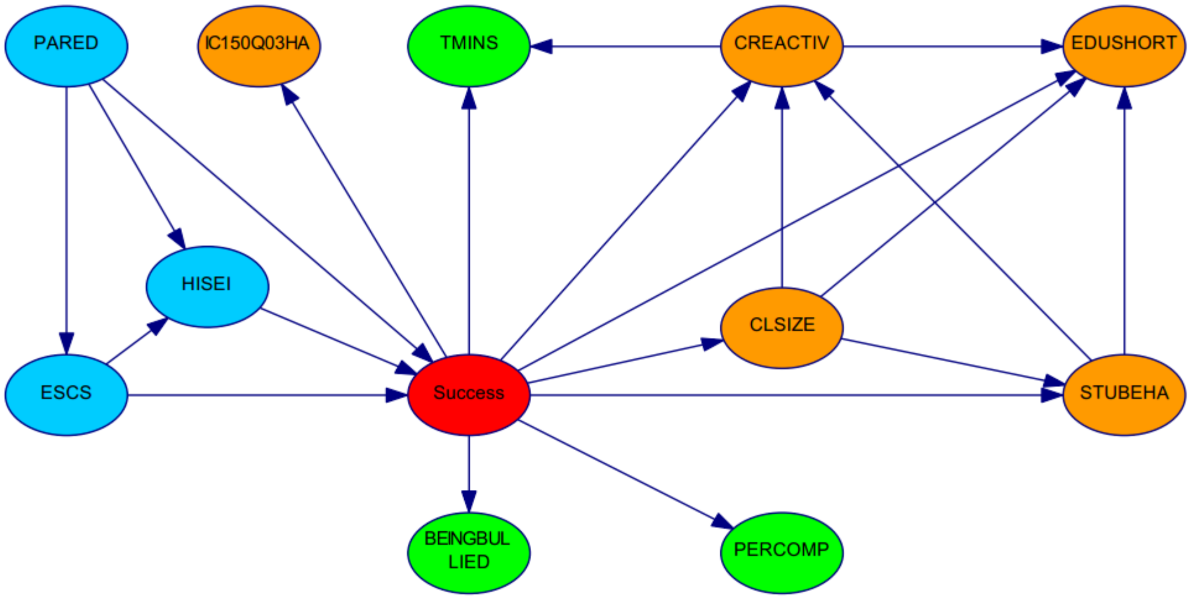
Figure 2. *BN Model with 28 Variables.*



In the first step, we constructed a BN model with 28 variables that affect science success. The variables were divided into 3 groups in the model: blue group variables are the student's family related variables, green group variables are the student's individual variables, and orange group variables are the student's school related variables. As a result, we obtained 89.4% accuracy from the model shown in Figure 2. However, the model and relationships of the variables were quite complex to understand and interpret.

As the Parsimony principle requires (Zhang, 1992), we reduced the number of variables by using expert knowledge and feature selection for simplifying the complex model structure suggested by the algorithm as well as making it more meaningful. In the second step, we selected 11 variables due to feature selection performed using the CFS subset algorithm. In the last step, we reconstruct the model with 11 effective variables. The final model is shown in Figure 3. The performance of this model is also close to the first model (Accuracy = 86.2%).

Figure 3. *BN Model with 11 Variables.*



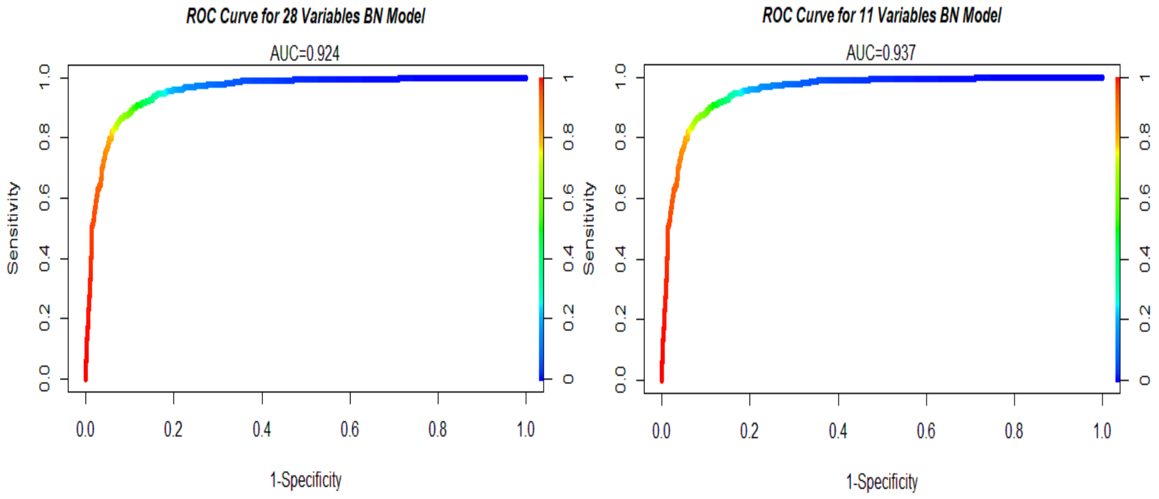
The reduced model produced a more meaningful with fewer variables. The comparison results of the models are shown in Table 2. When the models are compared, the success variable's prediction performance of the models is close to each other.

Table 2. *Model Comparison Results.*

	BN with 28 variables			BN with 11 variables		
	Female	Male	Overall	Female	Male	Overall
Accuracy	0.859	0.840	0.849	0.863	0.862	0.862
F-Measure	0.859	0.840	0.849	0.863	0.862	0.862
Kappa	0.717	0.679	0.699	0.726	0.723	0.725
RMSE	0.376	0.400	0.388	0.370	0.372	0.371
MAPE	10.664	11.319	10.992	10.617	9.916	10.267

We applied models separately for male and female students to investigate model performance differences, and no differences were found in terms of evaluation criteria. In the first and second models, we observed that male and female students differed by approximately 1% according to the MAPE value. In the literature, gender is effective on science success (Aşkın & Öz, 2020; Harker, 2000; Kilic Depren, 2020; Kjærnsli & Lie, 2004; Reilly et al., 2019; Torrecilla Sánchez et al., 2019; Yip et al., 2004). In this study, however, gender was not an effective variable on science success and prediction performance.

Figure 4. *ROC Curves of BN Models.*



ROC curves of the BN models obtained with ROCR package (Sing et al., 2005) are given in Figure 4. It was understood that the second model produced 0.937 AUC in the prediction of students' science success. AUC of the second model is better than that of the first model. In a BN model, if we know the value of the any factors, we can build scenarios to predict the student's high and low success probabilities with this new knowledge. The success prediction scenarios of the student-based variables are given in Table 3.

Table 3. *Success Prediction Scenarios with Student-Based Variables.*

	Evidence	SUCCESS	
		Very Low	Very High
BEING BULLIED	Very Low	0.731	0.269
	Low	0.498	0.502
	High	0.317	0.683
	Very High	0.613	0.387
PERCOMP	Very Low	0.612	0.388
	Low	0.449	0.551
	High	0.499	0.501
	Very High	0.402	0.598
TMINS	Very Low	0.653	0.347
	Low	0.578	0.422
	High	0.348	0.652
	Very High	0.665	0.335

In Table 3, we examined low and high science success according to the values of the variables of peer bullying (BEING BULLIED), perceived competition (PERCOMP), and total studying time in minutes (TMINS). When perceived bullying is very low, the probability of low success is 0.731 and the probability of high success is 0.269. Also, in the case of very high perceived bullying, the probability of low science success is 0.613 and the probability of high science success is 0.387. On the other hand, it is seen that successful students are more likely to be exposed to bullying. In other words, unsuccessful students fail not because of being bullied but because of other reasons. It is understood that successful students are exposed to more intense peer bullying. Considering perceived competitiveness, the probability of low success in the case of the perceived low competitiveness is 0.612, whereas it is calculated as 0.402 in the case of high competitiveness. It is seen that perceived competitiveness increases high success (0.598). Also, too much or too little studying of the student affects success negatively. We found that studying time above average positively affects science success (0.652).

Table 4. *Success Prediction Scenarios with Parental Variables.*

	Evidence	Success	
		Very Low	*Very High*
ESCS	Very Low	0.699	0.301
	Low	0.634	0.366
	High	0.522	0.478
	Very High	0.229	0.771
HISEI	Very Low	0.668	0.332
	Low	0.619	0.381
	High	0.534	0.466
	Very High	0.249	0.751
PARED	Low	0.617	0.383
	Moderate	0.680	0.320
	High	0.549	0.451
	Very High	0.257	0.743

The relationship between the student's family-related variables and science success is shown in Table 4. We observed that when the student’s ESCS value is low, their success is also low, and when the student’s ESCS value is high, their success is also high. Considering the index highest parental occupational status (HISEI) value, we found that if this value is too low, the success is also low (0.668) and that if high, the science success is very high. Finally, when we investigated the relationship between education level of family (PARED) and science success, we revealed that the student's science success was low (0.617) in the case of a low level of parental education, and high when the level of parental education was very high (0.743).

The relationship between the student's school-related variables and science success is given in Table 5. As seen in the table, the probability of science success is quite low in classes with fewer than 25 students (0.169). It is seen that the ideal class size is between 31 and 35. In schools where no creative activities (CREATIV) are carried out, the probability of students' science success is low (0.756). Choir and music events are generally held in Turkish schools. Therefore, no positive effect of these activities on success has been observed. However, artistic activities had a very positive effect on students' science success (0.706). In other words, artistic activities carried out at school should play an active role in increasing student success. When the dataset is examined in detail, science high schools and private high schools have more artistic activities and more successful students. In addition, it is observed that the families of the students in these schools are educated and the lack of educational materials is less.

Shortage of educational material (EDUSHORT) also has a negative impact on science success. In this sample, we observed 3 parts of shortage: low, high, and very high. The probability of science success increases (0.610) when the shortage is low. However, students show low success when the lack of teaching and learning materials is very high (0.715). Digital device use in lessons positively affects science success. Moreover, we have seen that using digital devices for at least 60 min weekly in science lessons increases the students’ science success (0.804). Students who declare that they do not work are more likely to fail (0.701). It could be stated that supporting the course with a digital device in science lessons increases the student's learning and thus the possibility of high success.

Student behavior hinder learning negatively influence success. As a result of the study, when students have fewer disruptive behaviors, their success probability increases (0.773), and when students display too many disruptive behaviors, science success is quite low (0.779).

Table 5. *Success Prediction Scenarios with School-Related Variables.*

Evidence		Success	
		Very Low	*Very High*
CLSIZE	Less than 25	0.831	0.169
	Between 26-30	0.432	0.568
	Between 31-35	0.368	0.632
	Between 36-50	0.700	0.300
	More than 50	0.468	0.532
CREACTIV	None	0.756	0.244
	Art club activities	0.294	0.706
	Band orchestra choir	0.612	0.388
	School play musical	0.664	0.336
EDUSHORT	Low	0.390	0.610
	High	0.486	0.514
	Very high	0.715	0.285
IC150Q03HA	I don’t study	0.701	0.299
	No time	0.622	0.378
	Between 1-30 min	0.664	0.336
	Between 31-60 min	0.449	0.551
	More than 60	0.196	0.804
STUBEHA	Very Low	0.227	0.773
	Low	0.330	0.670
	High	0.610	0.390
	Very High	0.779	0.221

4. DISCUSSION

Primarily, this is the first BN study that has been conducted with this dataset. Although different data mining methods were used in previous studies, BN was not used to model science success. Unlike rule-based machine learning such as support vector machines, logistic regression and artificial neural networks, it enables queries which explicitly reveal cause-effect relationships between variables (Pearl, 2014). Besides, the posterior probabilities are updated with each new information, allows more accurate estimations (Korb & Nicholson, 2010). Hence, modeling with BN provides an advantage over other machine learning methods in revealing complex relationships (Karaboga et al., 2021). BN, which is widely used in a variety of fields, has been used in a small number of studies in the field of education (Almond et al., 2015; Culbertson, 2016; Reichenberg, 2018). However, BN is more advantageous than other methods with its ability to model students in the field of education (Levy, 2016; Lytvynenko et al., 2019;

Sinharay, 2006) and to evaluate the model quickly (Kenekayoro, 2018; Kustitskaya et al., 2020; Millán et al., 2013; Nguyen & Do, 2009).

Essential improvements in the education system are vital to enhance students' success. Therefore, educators, researchers, and government agencies should prioritize research for identifying factors to improve success. Especially, enhancing science success is considered as a key to the scientific and economic progress of countries (Sjøberg, 2019). Students' individual, family-related and school-related factors are effective on science success (Beese & Liang, 2010; Kiray et al., 2015; Lee & Shute, 2010; Yıldırım, 2012). PISA aims to help explain the differences in student performance by collecting data on students' successes, as well as collecting each student, family and personal information (Beese & Liang, 2010). The effect of interaction among these variables, which are normally effective separately, on success has been investigated using the advantages of BN. In this study, we used the dataset of Turkey obtained from the PISA 2018 survey. In the first step of the study, we discretized the variables.Then, we constructed a dataset that included the most successful 25% and the least successful 25% students. As a consequence, we examined the effects of the factors which influence science success by creating a model with 28 variables. The most effective variables determined with the CFS subset algorithm were BEING BULLIED, PERCOMP, TMINS, ESCS, HISEI, PARED, CLSIZE, CREACTIV, EDUSHORT, IC150Q03HA, and STUBEHA. A more effective model was obtained with these 11 determined variables.

Bullying is a type of violence which disrupts school climate and harms students' physical or mental states (Fry et al., 2018; Wachs et al., 2019). The student's success is low in the case of high perceived bullying (Clarke & Kiselica, 1997; Jan, 2015; Sudrajad et al., 2020). Successful students are exposed to more intense bullying than unsuccessful students.. Also, perceived high competitiveness increases success (Karataş & Ergin, 2018; Muñoz-Merino et al., 2014; OECD, 2020). Less disruptive behaviors of the students increase their science success. On the contrary, in schools with too many disruptive behaviors, the science success decreases (Ertem, 2021; Özdemir et al., 2019). We observed that too much or too little study of the student has a negative impact on success.

The increase in parents' socio-economic and cultural status increases the students' science success. We found that if the student's parental occupational status is too low, their success is also low, and that a very high status of parental occupation correlates with students' high science success. Similarly, students with a low level of family education have low science success, and when their family's education level is very high, their science success is very high. As a result, the economic and socio-cultural status of the student's family, their educational background, and occupational status are effective upon students' science success. According to the literature, it is clear that students having families with high educational, and sociaconomic status are more successful (Gamazo & Martínez-Abad, 2020; Lee & Shute, 2010; Sirin, 2005; Topçu et al., 2015; Yıldırım, 2012). The high science success of these students is related to their awareness of science and education. Children of educated families are also conscious about science education (O'Connell, 2019).

The class sizes of the students who participated in the survey were generally more than 50. Classes are smaller in vocational schools located in small settlements. The success level of students studying in those schools is generally low (Suna et al., 2020). The probability of success is quite low in classes where there are fewer than 25 students. Classes in Anatolian high schools are also larger than those in other schools. Medium-sized classes are provided in science high schools and private colleges. It is stated in the literature that as the classes get smaller, the success increases, but this effect is low (Borland et al., 2005; Hanushek & Woessmann, 2017; Hattie, 2005). Also, reducing class size is quite costly (Ehrenberg et al., 2001; White, 2018). Because of that, educators should determine the ideal class size, considering the situation of the students and the school (Borland et al., 2005; Wößmann, 2005). This study reveals that the ideal class size is between 31 and 35.

Generally, choir and musical events are held in Turkish schools. In these schools where there are no other activities, it is impossible to encourage students to participate in different activities. Artistic activities other than music should have a positive impact on students' science success. Extracurricular artistic school activities play an active role in increasing student success (Tang & Zhang, 2020). That’s because, according to the hidden curriculum (Margolis, 2001), extracurricular activities ensure a rise in success by increasing concentration and motivation (Stearns & Glennie, 2010).

Another effective factor on science success is the lack of educational material shortage (Altun & Kalkan, 2019; Archibald, 2006). There may be a lack of educational materials at schools in various disadvantaged regions. Particularly, in socioeconomically disadvantaged regions, schools cannot fill those deficiencies by getting support from families (van der Berg, 2008). In schools where there exist few artistic activities, educational material shortage such as digital devices for lessons is higher. The use of digital devices in lessons has also been identified as a variable that positively affect success. Accordingly, we have seen that the use of digital devices in science classes increases the probability of student success. Supporting science lessons with digital devices boosts student success by facilitating their learning (Bingimlas, 2009; Chen et al., 2019; Odell et al., 2020).

Apart from the studies we mentioned, studies have been conducted on factors affecting science success such as teachers, school, school curriculum (Cansiz & Cansiz, 2019; Tatar et al., 2016). Numerical content of science subjects and the intensity of curriculum are important predictors of science success (Tatar et al., 2016). If we identify the factors affecting student success, we will guide the reforms that need to be made in the curriculum to increase students' low success level (Topçu et al., 2016).

5. CONCLUSIONS and IMPLICATIONS

Science literacy requires students to explain various phenomena scientifically, design and evaluate the scientific method, and interpret the findings scientifically (OECD, 2019a). The relationship of students' background of knowledge and skills with other variables obtained is one of the main indicators of PISA (MEB, 2019; Schleicher, 2019). To sum up, PISA evaluates how students could use their scientific content knowledge in their daily life by combining methodological and epistemic knowledge (OECD, 2019a, 2019c). In this respect, science literacy examines whether students could go beyond the school curriculum.

Based on the PISA 2018 results, Turkish students scored lower than the OECD average. Although some progress has been made compared to previous years, this progress is inadequate. The main purpose of the Turkish science curriculum is to raise science-literate students (MEB, 2018). However, science literate individuals do not grow as the curriculum aims. Hence, it is necessary to explore how students could improve their ability to use information and interpret it in real life.

Even though the most significant source of student success is internal motivation (Augustyniak et al., 2016), school and family variables are also important. In particular, opportunities provided to students by their families, and schools are a major key to success. Low-income families are a significant issue here. Nonetheless, no short-term solution to this problem exists. Instead, it is required to raise awareness in cooperation with the families of students and to organize activities that will encourage them to study. Providing an optimum studying environment and ideal teaching and learning materials will be encouraging for students.

Increasing opportunities in schools will also increase student success. Computer-assisted classes have demonstrated significant potential in enhancing students' problem-solving abilities, particularly in the domain of science education (Bayrak & Bayram, 2010; Chang, 2002). Schools must take measures to prevent peer bullying and disruptive student behaviors that hinder learning. To achieve this, collaborative efforts between schools and families are crucial in devising various policies aimed at safeguarding students. The study time of students should be maintainedat a sufficient level. Too little or too much work should harm student success. Competition and cooperation among students should be encouraged through various activities to increase science success. Motivating students to study more emerges as a key factor in attaining a long-term success rate. For this purpose, education politicians should prepare a rich curriculum based on experiments and observations to have more fun in science lessons.

PISA's science literacy qualifications are almost non-existent within the scope of Turkish science curricular outcomes (Cansiz & Cansiz, 2019). The curriculum is not sufficient to raise scientifically literate individuals. To raise individuals who research, question and use 21st century information and technologies, changes should be made and implemented in education systems. Thus, we can use the assessments obtained using the BN model to increase students' science success in future exams. We proved that the results of this study will provide effective clues for innovations in the educational system. We hope that it will be a useful model for the evaluation of international exams and contributions to educational systems not only for Turkey but also for all OECD countries.

This study is not without limitations. First, the study was conducted with PISA data only from Turkey. Nevertheless, students from different countries could also be analyzed to make the study comparative. Besides, models could be made more successful by combining BN algorithms with newly developed machine learning methods. In addition, different results can be obtained by repeating this study for different data sets.

Acknowledgments

Acknowledgments of people, grants, and funds should be placed in a separate section before the References. If the study has been previously presented at a conference or a scholarly meeting, it should be mentioned here.

Declaration of Conflicting Interests and Ethics

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

Authorship Contribution Statement

Hasan Aykut Karaboga: Investigation, Resources, Methodology Visualization, Software, Formal Analysis, and Writing-original draft. **Ibrahim Demir**: Supervision, and Validation.

Orcid

Hasan Aykut Karaboga https://orcid.org/0000-0001-8877-3267
Ibrahim Demir https://orcid.org/0000-0002-2734-4116

REFERENCES

Almond, R.G., DiBello, L.V., Moulder, B., & Zapata-Rivera, J.-D. (2007). Modeling Diagnostic Assessments with Bayesian Networks. *Journal of Educational Measurement*, *44*(4), 341–359. https://doi.org/10.1111/j.1745-3984.2007.00043.x

Almond, R.G., & Mislevy, R.J. (1999). Graphical Models and Computerized Adaptive Testing. *Applied Psychological Measurement*, *23*(3), 223-237. https://doi.org/10.1177/01466219922031347

Almond, R.G., Mislevy, R.J., Steinberg, L.S., Yan, D., & Williamson, D.M. (2015). *Bayesian Networks in Educational Assessment*. Springer.

Altun, A., & Kalkan, Ö.K. (2019). Cross-national study on students and school factors affecting science literacy. *Educational Studies*, 1-19. https://doi.org/10.1080/03055698.2019.1702511

Archibald, S. (2006). Narrowing in on Educational Resources That Do Affect Student Achievement. *Peabody Journal of Education*, *81*(4), 23-42. https://doi.org/10.1207/s15327930pje8104_2

Aşkın, Ö.E., & Öz, E. (2020). Cross-National Comparisons of Students' Science Success Based on Gender Variability: Evidence From TIMSS. *Journal of Baltic Science Education*, *19*(2), 186–200. https://doi.org/10.33225/jbse/20.19.186

Augustyniak, R.A., Ables, A.Z., Guilford, P., Lujan, H.L., Cortright, R.N., & DiCarlo, S.E. (2016). Intrinsic motivation: An overlooked component for student success. *Advances in Physiology Education*, *40*(4), 465–466. https://doi.org/10.1152/advan.00072.2016

Baker, R.S.J.d, & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, *1*(1), Article 1. https://doi.org/10.5281/zenodo.3554657

BayesFusion, L. (2017). GeNIe modeler user manual. *BayesFusion, LLC, Pittsburgh, PA*.

Bayrak, B.K., & Bayram, H. (2010). The effect of computer aided teaching method on the students' academic achievement in the science and technology course. *Procedia - Social and Behavioral Sciences*, *9*, 235–238. https://doi.org/10.1016/j.sbspro.2010.12.142

Beese, J., & Liang, X. (2010). Do resources matter? PISA science achievement comparisons between students in the United States, Canada and Finland. *Improving Schools*, *13*(3), 266–279. https://doi.org/10.1177/1365480210390554

Bingimlas, K.A. (2009). Barriers to the Successful Integration of ICT in Teaching and Learning Environments: A Review of the Literature. *Eurasia Journal of Mathematics, Science & Technology Education*, *5*(3), 235–245.

Borland, M.V., Howsen, R.M., & Trawick, M.W. (2005). An investigation of the effect of class size on student academic achievement. *Education Economics*, *13*(1), 73–83. https://doi.org/10.1080/0964529042000325216

Cansiz, N., & Cansiz, M. (2019). Evaluating Turkish science curriculum with PISA scientific literacy framework. *Turkish Journal of Education*, *8*(3), Article 3. https://doi.org/10.19128/turje.545798

Carnoy, M., Khavenson, T., & Ivanova, A. (2015). Using TIMSS and PISA results to inform educational policy: A study of Russia and its neighbours. *Compare: A Journal of Comparative and International Education*, *45*(2), 248-271. https://doi.org/10.1080/03057925.2013.855002

Chang, C.-Y. (2002). Does Computer-Assisted Instruction + Problem Solving = Improved Science Outcomes? A Pioneer Study. *The Journal of Educational Research*, *95*(3), 143–150. https://doi.org/10.1080/00220670209596584

Chen, J., Zhang, Y., Wei, Y., & Hu, J. (2019). Discrimination of the Contextual Features of Top Performers in Scientific Literacy Using a Machine Learning Approach. *Research in Science Education*. https://doi.org/10.1007/s11165-019-9835-y

Clarke, E.A., & Kiselica, M.S. (1997). A systemic counseling approach to the problem of bullying. *Elementary School Guidance & Counseling*, *31*(4), 310–325.

Culbertson, M.J. (2016). Bayesian Networks in Educational Assessment: The State of the Field. *Applied Psychological Measurement*, *40*(1), 3-21. https://doi.org/10.1177/0146621615590401

Deng, Z., & Gopinathan, S. (2016). PISA and high-performing education systems: Explaining Singapore's education success. *Comparative Education*, *52*(4), 449-472. https://doi.org/

10.1080/03050068.2016.1219535

Desmarais, M.C., & Baker, R.S.J.d. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, *22*(1), 9–38. https://doi.org/10.1007/s11257-011-9106-8

Ehrenberg, R.G., Brewer, D.J., Gamoran, A., & Willms, J.D. (2001). Class Size and Student Achievement. *Psychological Science in The Public Interest*, *2*(1), 30.

Ertem, H.Y. (2021). Examination of Turkey's PISA 2018 reading literacy scores within student-level and school-level variables. *Participatory Educational Research*, *8*(1), 248–264. https://doi.org/10.17275/per.21.14.8.1

Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, *30*(1), 27–38. https://doi.org/10.1016/j.patrec.2008.08.010

Fry, D., Fang, X., Elliott, S., Casey, T., Zheng, X., Li, J., Florian, L., & McCluskey, G. (2018). The relationships between violence in childhood and educational outcomes: A global systematic review and meta-analysis. *Child Abuse & Neglect*, *75*, 6–28. https://doi.org/10.1016/j.chiabu.2017.06.021

Gamazo, A., & Martínez-Abad, F. (2020). An Exploration of Factors Linked to Academic Performance in PISA 2018 Through Data Mining Techniques. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.575167

Gilbert, J.K., Boulter, C.J., & Elmer, R. (2000). Positioning Models in Science Education and in Design and Technology Education. In J.K. Gilbert & C.J. Boulter (Eds.), *Developing Models in Science Education* (pp. 3-17). Springer Netherlands. https://doi.org/10.1007/978-94-010-0876-1_1

Hall, M.A. (1999a). *Correlation-based Feature Selection for Machine Learning* [PhD Thesis]. The University of Waikato.

Hall, M.A. (1999b). *Feature selection for discrete and numeric class machine learning* [Working Paper]. Computer Science, University of Waikato. https://researchcommons.waikato.ac.nz/handle/10289/1033

Hall, M.A. (2000). *Correlation-based feature selection of discrete and numeric class machine learning* [Working Paper]. University of Waikato, Department of Computer Science. https://researchcommons.waikato.ac.nz/handle/10289/1024

Hanushek, E.A., & Woessmann, L. (2017). School Resources and Student Achievement: A Review of Cross-Country Economic Research. In M. Rosén, K. Yang Hansen, & U. Wolff (Eds.), *Cognitive Abilities and Educational Outcomes* (pp. 149–171). Springer International Publishing. https://doi.org/10.1007/978-3-319-43473-5_8

Harker, R. (2000). Achievement, Gender and the Single-Sex/Coed Debate. *British Journal of Sociology of Education*, *21*(2), 203–218. https://doi.org/10.1080/713655349

Hattie, J. (2005). The paradox of reducing class size and improving learning outcomes. *International Journal of Educational Research*, *43*(6), 387-425. https://doi.org/10.1016/j.ijer.2006.07.002

Hossin, M., & Sulaiman, M.N. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 01–11. https://doi.org/10.5121/ijdkp.2015.5201

Jan, A. (2015). Bullying in Elementary Schools: Its Causes and Effects on Students. *Journal of Education and Practice*, 15.

Karaboga, H.A., Gunel, A., Korkut, S.V., Demir, I., & Celik, R. (2021). Bayesian Network as a Decision Tool for Predicting ALS Disease. *Brain Sciences*, *11*(2), Article 2. https://doi.org/10.3390/brainsci11020150

Karakoç Alatlı, B. (2020). Investigation of Factors Associated with Science Literacy Performance of Students by Hierarchical Linear Modeling: PISA 2015 Comparison of

Turkey and Singapore. *TED Education and Science Magazine*. https://doi.org/10.15390/EB.2020.8188

Karataş, H., & Ergin, A. (2018). Üniversite Öğrencilerinin Başarı Odaklı Motivasyon Düzeyleri [Achievement-Oriented Motivation Levels of University Students]. *Hacettepe University Journal of Education*, 1–20. https://doi.org/10.16986/HUJE.2018036646

Kenekayoro, P. (2018). An Exploratory Study on the Use of Machine Learning to Predict Student Academic Performance: *International Journal of Knowledge-Based Organizations*, *8*(4), 67–79. https://doi.org/10.4018/IJKBO.2018100104

Kilic Depren, S. (2018). Prediction of Students' Science Achievement: An Application of Multivariate Adaptive Regression Splines and Regression Trees. *Journal of Baltic Science Education*, *17*(5), 887–903. https://doi.org/10.33225/jbse/18.17.887

Kilic Depren, S. (2020). Determination of the Factors Affecting Students' Science Achievement Level in Turkey and Singapore: An Application of Quantile Regression Mixture Model. *Journal of Baltic Science Education*, *19*(2), 247-260. https://doi.org/10.33225/jbse/20.19.247

Kiray, S.A., Gok, B., & Bozkir, A.S. (2015). Identifying the Factors Affecting Science and Mathematics Achievement Using Data Mining Methods. *Journal of Education in Science, Environment and Health*, *1*(1), 28. https://doi.org/10.21891/jeseh.41216

Kjærnsli, M., & Lie, S. (2004). PISA and scientific literacy: Similarities and differences between the nordic countries. *Scandinavian Journal of Educational Research*, *48*(3), 271–286. https://doi.org/10.1080/00313830410001695736

Korb, K.B., & Nicholson, A.E. (2010). *Bayesian Artificial Intelligence*. CRC Press.

Kustitskaya, T.A., Kytmanov, A.A., & Noskov, M.V. (2020). Student-at-risk detection by current learning performance indicators using Bayesian networks. *ArXiv:2004.09774 [Stat]*. http://arxiv.org/abs/2004.09774

Lee, J., & Shute, V.J. (2010). Personal and Social-Contextual Factors in K–12 Academic Performance: An Integrative Perspective on Student Learning. *Educational Psychologist*, *45*(3), 185–202. https://doi.org/10.1080/00461520.2010.493471

Levy, R. (2016). Advances in Bayesian Modeling in Educational Research. *Educational Psychologist*, *51*(3–4), 368–380. https://doi.org/10.1080/00461520.2016.1207540

Lima. (2014). Heuristic Discretization Method for Bayesian Networks. *Journal of Computer Science*, *10*(5), 869–878. https://doi.org/10.3844/jcssp.2014.869.878

Lytvynenko, V., Savina, N., Voronenko, M., Doroschuk, N., Smailova, S., Boskin, O., & Kravchenko, T. (2019). Development, Validation and Testing of the Bayesian Network of Educational Institutions Financing. *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, *1*, 412–417. https://doi.org/10.1109/IDAACS.2019.8924307

Margolis, E. (2001). *The Hidden Curriculum in Higher Education*. Psychology Press.

Marsland, S. (2015). *Machine Learning: An Algorithmic Perspective* (Second Edition). CRC press.

Martínez Abad, F., & Chaparro Caso López, A.A. (2017). Data-mining techniques in detecting factors linked to academic achievement. *School Effectiveness and School Improvement*, *28*(1), 39–55. https://doi.org/10.1080/09243453.2016.1235591

MEB. (2018). *Fen Bilimleri Dersi Öğretim Programı [Science Course Curriculum]*.. Talim ve Terbiye Kurulu Başkanlığı, Ankara. https://mufredat.meb.gov.tr/Dosyalar/201812312311937-FEN%20B%C4%B0L%C4%B0MLER%C4%B0%20%C3%96%C4%9ERET%C4%B0M%20PROGRAMI2018.pdf

MEB. (2019). *PISA 2018 Turkiye Ön Raporu [PISA 2018 Turkey Preliminary Report]*. Milli Eğitim Bakanlığı. http://www.meb.gov.tr/meb_iys_dosyalar/2019_12/03105347_PISA_2018_Turkiye_On_Raporu.pdf

Millán, E., Descalço, L., Castillo, G., Oliveira, P., & Diogo, S. (2013). Using Bayesian networks to improve knowledge assessment. *Computers & Education*, *60*(1), 436–447. https://doi.org/10.1016/j.compedu.2012.06.012

Muñoz-Merino, P.J., Molina, M.F., Muñoz-Organero, M., & Kloos, C.D. (2014). Motivation and Emotions in Competition Systems for Education: An Empirical Study. *IEEE Transactions on Education*, *57*(3), 182–187. https://doi.org/10.1109/TE.2013.2297318

Neapolitan, R.E. (2009). *Probabilistic methods for bioinformatics: With an introduction to Bayesian networks*. Morgan Kaufmann/Elsevier.

Nguyen, L., & Do, P. (2009). Combination of Bayesian Network and Overlay Model in User Modeling. In G. Allen, J. Nabrzyski, E. Seidel, G.D. van Albada, J. Dongarra, & P.M.A. Sloot (Eds.), *Computational Science – ICCS 2009* (pp. 5–14). Springer. https://doi.org/10.1007/978-3-642-01973-9_2

Nielsen, T.D., & Jensen, F.V. (2009). *Bayesian Networks and Decision Graphs*. Springer Science & Business Media.

Nojavan A., F., Qian, S.S., & Stow, C.A. (2017). Comparative analysis of discretization methods in Bayesian networks. *Environmental Modelling & Software*, *87*, 64–71. https://doi.org/10.1016/j.envsoft.2016.10.007

O'Connell, M. (2019). Is the impact of SES on educational performance overestimated? Evidence from the PISA survey. *Intelligence*, *75*, 41-47. https://doi.org/10.1016/j.intell.2019.04.005

Odell, B., Galovan, A.M., & Cutumisu, M. (2020). The Relation Between ICT and Science in PISA 2015 for Bulgarian and Finnish Students. *EURASIA Journal of Mathematics, Science and Technology Education*, *16*(6). https://doi.org/10.29333/ejmste/7805

OECD. (2019a). *PISA 2018 Assessment and Analytical Framework*. OECD. https://doi.org/10.1787/b25efab8-en

OECD. (2019b). *PISA 2018 Results (Volume I): What Students Know and Can Do*. OECD. https://doi.org/10.1787/5f07c754-en

OECD. (2019c). *PISA 2018 Results (Volume II): Where All Students Can Succeed*. OECD. https://doi.org/10.1787/b5fd1b8f-en

OECD. (2020). *Do boys and girls have similar attitudes towards competition and failure?* (PISA in Focus 105; PISA in Focus, Vol. 105). https://doi.org/10.1787/a8898906-en

Özdemir, E., Cansiz, M., Cansiz, N., & Üstün, U. (2019). Türkiye deki Öğrencilerin Fen Okuryazarlığını Etkileyen Faktörler Nelerdir PISA 2015 Verisine Dayalı Bir Hiyerarşik Doğrusal Modelleme Çalışması. *Hacettepe University Journal of Education*, 1–16. https://doi.org/10.16986/HUJE.2019050786

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: Networks of plausible inference* (Revised Second Printing). Morgan Kaufmann. https://doi.org/10.1016/B978-0-08-051489-5.50002-3

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, *41*(4), 1432-1462. https://doi.org/10.1016/j.eswa.2013.08.042

Ramírez-Noriega, A., Juárez-Ramírez, R., Leyva-López, J.C., Jiménez, S., & Figueroa-Pérez, J.F. (2021). A Method for Building the Quantitative and Qualitative Part of Bayesian Networks for Intelligent Tutoring Systems. *The Computer Journal*, *bxab124*. https://doi.org/10.1093/comjnl/bxab124

Rastrollo-Guerrero, J.L., Gómez-Pulido, J.A., & Durán-Domínguez, A. (2020). Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review. *Applied Sciences*, *10*(3), 1042. https://doi.org/10.3390/app10031042

Reichenberg, R. (2018). Dynamic Bayesian Networks in Educational Measurement: Reviewing and Advancing the State of the Field. *Applied Measurement in Education*, *31*(4), 335–

350. https://doi.org/10.1080/08957347.2018.1495217

Reilly, D., Neumann, D.L., & Andrews, G. (2019). Investigating Gender Differences in Mathematics and Science: Results from the 2011 Trends in Mathematics and Science Survey. *Research in Science Education*, *49*(1), 25–50. https://doi.org/10.1007/s11165-017-9630-6

Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Ropero, R.F., Renooij, S., & van der Gaag, L.C. (2018). Discretizing environmental data for learning Bayesian-network classifiers. *Ecological Modelling*, *368*, 391–403. https://doi.org/10.1016/j.ecolmodel.2017.12.015

Sağlam, A.Ç., & Aydoğmuş, M. (2016). Gelişmiş ve Gelişmekte Olan Ülkelerin Eğitim Sistemlerinin Denetim Yapıları Karşılaştırıldığında Türkiye Eğitim Sisteminin Denetimi Ne Durumdadır? [When the Supervision Structures of the Education Systems of Developed and Developing Countries are Compared, How is the Supervision of the Turkish Education System?]. *Uşak Üniversitesi Sosyal Bilimler Dergisi*, *9*(1), Article 1. https://doi.org/10.12780/uusbd.50788

Saini, M.K., & Goel, N. (2019). How Smart Are Smart Classrooms? A Review of Smart Classroom Technologies. *ACM Computing Surveys*, *52*(6), 130:1-130:28. https://doi.org/10.1145/3365757

Schleicher, A. (2019). PISA 2018: Insights and Interpretations. In *OECD Publishing*. OECD Publishing.

Sebastian, J., Moon, J.-M., & Cunningham, M. (2017). The relationship of school-based parental involvement with student achievement: A comparison of principal and parent survey reports from PISA 2012. *Educational Studies*, *43*(2), 123–146. https://doi.org/10.1080/03055698.2016.1248900

Sener, E., Karaboga, H.A., & Demir, I. (2019). Bayesian Network Model of Turkish Financial Market from Year-to-September 30th of 2016. *Sigma: Journal of Engineering & Natural Sciences / Mühendislik ve Fen Bilimleri Dergisi*, *37*(4), 1493–1507.

Sheldrake, R., Mujtaba, T., & Reiss, M.J. (2017). Science teaching and students' attitudes and aspirations: The importance of conveying the applications and relevance of science. *International Journal of Educational Research*, *85*, 167-183. https://doi.org/10.1016/j.ijer.2017.08.002

Shin, D., & Shim, J. (2021). A Systematic Review on Data Mining for Mathematics and Science Education. *International Journal of Science and Mathematics Education*, *19*(4), 639–659. https://doi.org/10.1007/s10763-020-10085-7

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: Visualizing classifier performance in R. *Bioinformatics*, *21*(20), 3940-3941. https://doi.org/10.1093/bioinformatics/bti623

Sinharay, S. (2006). Model Diagnostics for Bayesian Networks. *Journal of Educational and Behavioral Statistics*, *31*(1), 1–33.

Sinharay, S. (2016). An NCME Instructional Module on Data Mining Methods for Classification and Regression. *Educational Measurement: Issues and Practice*, *35*(3), 38–54. https://doi.org/10.1111/emip.12115

Sirin, S.R. (2005). Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research. *Review of Educational Research*, *75*(3), 417–453. https://doi.org/10.3102/00346543075003417

Sjøberg, S. (2019). The PISA-syndrome – How the OECD has hijacked the way we perceive pupils, schools and education. *Confero: Essays on Education, Philosophy and Politics*, *7*(1), 12–65.

Stearns, E., & Glennie, E.J. (2010). Opportunities to participate: Extracurricular activities' distribution across and academic correlates in high schools. *Social Science Research*, *39*(2), 296–309. https://doi.org/10.1016/j.ssresearch.2009.08.001

Sudrajad, K., Soemanto, Rb., & Prasetya, H. (2020). The Effect of Bullying on Depression, Academic Activity, and Communication in Adolescents in Surakarta: A Multilevel Logistic Regression. *Journal of Health Promotion and Behavior*, *5*(2), 79–86. https://doi.org/10.26911/thejhpb.2020.05.02.02

Suna, H.E., Tanberkan, H., & Özer, M. (2020). Changes in Literacy of Students in Turkey by Years and School Types: Performance of Students in PISA Applications. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *11*(1), 76-97. https://doi.org/10.21031/epod.702191

Tang, X., & Zhang, D. (2020). How informal science learning experience influences students' science performance: A cross-cultural study based on PISA 2015. *International Journal of Science Education*, *42*(4), 598–616. https://doi.org/10.1080/09500693.2020.1719290

Tatar, E., Tüysüz, C., Tosun, C., & Ilhan, N. (2016). Investigation of Factors Affecting Students' Science Achievement According to Student Science Teachers. *International Journal of Instruction*, *9*(2), 153–166.

Tingir, S., & Almond, R. (2017). Using Bayesian Networks to Visually Compare the Countries: An Example from PISA. *Journal of Education*, *4*(3), 11.

Topçu, M.S., Arıkan, S., & Erbilgin, E. (2015). Turkish Students' Science Performance and Related Factors in PISA 2006 and 2009. *The Australian Educational Researcher*, *42*(1), 117–132. https://doi.org/10.1007/s13384-014-0157-9

Topçu, M.S., Erbilgin, E., & Arikan, S. (2016). Factors Predicting Turkish and Korean Students' Science and Mathematics Achievement in TIMSS 2011. *EURASIA Journal of Mathematics, Science and Technology Education*, *12*(7). https://doi.org/10.12973/eurasia.2016.1530a

Torrecilla Sánchez, E.M., Olmos Miguélañez, S., & Martínez Abad, F. (2019). Explanatory factors as predictors of academic achievement in PISA tests. An analysis of the moderating effect of gender. *International Journal of Educational Research*, *96*, 111–119. https://doi.org/10.1016/j.ijer.2019.06.002

Üstün, U., Özdemir, E., Cansiz, M., & Cansiz, N. (2020). Türkiye'deki Öğrencilerin Fen Okuryazarlığını Etkileyen Faktörler Nelerdir? PISA 2015 Verisine Dayalı Bir Hiyerarşik Doğrusal Modelleme Çalışması [What are the Factors Affecting Students' Science Literacy in Turkey? A Hierarchical Linear Modeling Study Based on PISA 2015 Data]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *35*(3), Article 3.

van der Berg, S. (2008). How effective are poor schools? Poverty and educational outcomes in South Africa. *Studies in Educational Evaluation*, *34*(3), 145-154. https://doi.org/10.1016/j.stueduc.2008.07.005

Wachs, S., Bilz, L., Niproschke, S., & Schubarth, W. (2019). Bullying Intervention in Schools: A Multilevel Analysis of Teachers' Success in Handling Bullying from the Students' Perspective. *The Journal of Early Adolescence*, *39*(5), 642-668. https://doi.org/10.1177/0272431618780423

White, H. (2018). Small Class Size Has at Best a Small Effect on Academic Achievement. Plain Language Summary. In *Campbell Collaboration*. Campbell Collaboration. https://eric.ed.gov/?id=ED610283

Wong, T.-T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, *48*(9), 2839-2846. https://doi.org/10.1016/j.patcog.2015.03.009

Wößmann, L. (2005). Educational production in Europe. *Economic Policy*, *20*(43), 446–504. https://doi.org/10.1111/j.1468-0327.2005.00144.x

Xing, W., Li, C., Chen, G., Huang, X., Chao, J., Massicotte, J., & Xie, C. (2021). Automatic Assessment of Students' Engineering Design Performance Using a Bayesian Network Model. *Journal of Educational Computing Research*, *59*(2), 230-256. https://doi.org/10.1177/0735633120960422

Yang, Y., & Webb, G.I. (2002). A Comparative Study of Discretization Methods for Naive-Bayes Classifiers. *In Proceedings of PKAW 2002: The 2002 Pacific Rim Knowledge Acquisition Workshop*, 159–173.

Yip, D.Y., Chiu, M.M., & Ho, E.S.C. (2004). Hong Kong Student Achievement in OECD-PISA Study: Gender Differences in Science Content, Literacy Skills, and Test Item Formats. *International Journal of Science and Mathematics Education*, *2*(1), 91–106. https://doi.org/10.1023/B:IJMA.0000026537.85199.36

Yıldırım, S. (2012). Teacher Support, Motivation, Learning Strategy Use, and Achievement: A Multilevel Mediation Model. *The Journal of Experimental Education*, *80*(2), 150–172. https://doi.org/10.1080/00220973.2011.596855

Zhang, P. (1992). On the Distributional Properties of Model Selection Criteria. *Journal of the American Statistical Association*, *87*(419), 732-737. https://doi.org/10.1080/01621459.1992.10475275

Zwick, R., & Lenaburg, L. (2009). Using Discrete Loss Functions and Weighted Kappa for Classification: An Illustration Based on Bayesian Network Analysis. *Journal of Educational and Behavioral Statistics*, *34*(2), 190-200. https://doi.org/10.3102/1076998609332106

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 3, 434–453

https://doi.org/10.21449/ijate.1279152

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

Investigating the sources of differential item functioning: A sample critical thinking motivation scale

Fatma Betül Kurnaz[1,*], Hüseyin Yıldız[2]

[1]Karabuk University, Faculty of Letter, Department of Educational Sciences, Karabük, Türkiye
[2]Australian Council for Educational Research, Melbourne, Australia

ARTICLE HISTORY

Received: Apr. 07, 2023
Revised: July 17, 2023
Accepted: July 27, 2023

Keywords:

Differential item functioning,
Critical thinking,
Critical thinking motivation.

Abstract: Investigating the existence of items with differential item functioning (DIF) may provide more accurate comparisons of group differences in studies that aim to compare scores obtained in a test by groups with different characteristics. In the present study, a scale measuring critical thinking motivation that was adapted to the Turkish culture was applied to 817 participants, who were high school graduates, university students, and university graduates. The aim of the study was to examine whether the data collected from these participants had DIF or not. Hence, DIF analysis of the collected data was performed via the "lordif" function in the R "lordif" package. DIF was found to occur in twelve items, three of which were related to gender and nine to level of education. While it was revealed that the content of the items was the source of gender related DIF, the source of DIF related to level of education was found to be the language and expression of the items.

1. INTRODUCTION

The investigation of the psychometric properties of measurement tools measuring motivational factors related to such cognitive factors as success, intelligence, and critical thinking can facilitate understanding of the construct that is of cultural interest. In the development and adaptation of scales that measure affective factors related to cognitive features such as learning motivation, critical thinking dispositions, and beliefs about learning and knowing these cultural differences can provide important information to researchers developing or adapting these measurement tools. It is stated in the related literature that researchers need not only be well-informed, but also to provide explanations about the psychometric properties of developed or adapted measurement tools (Crocker & Algina, 1986). Adaptation of measurement tools developed in different cultures creates discussions on the problem that the construct expected to be measured via a measurement tool can show variations across cultures (Cole et al., 1993; Ferne & Rupp, 2007). Hence, it is stated in the relevant literature that the construct validity, item and test bias as well as cultural norms of measurement tools adapted to different cultures particularly need to be investigated when the aim is to make inter-cultural comparisons (Byrne et al., 2009).

*CONTACT: Fatma Betül Kurnaz ✉ betulkurnaz@gmail.com Karabuk University, Faculty of Letter, Department of Educational Sciences, Karabuk, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

As a result of the adaptation of a scale developed in one culture to another culture, experts may seek evidence that the original and adapted measurement forms ensure the equivalence of the construct measured, that the scale can reveal the difference between groups in a culture-independent manner, and that the effect of culture and language on the construct measured is reduced. For this reason, studies that provide evidence on how the results obtained from the application of the adapted scale represent the construct in the target culture gain importance. Such studies may require in-depth qualitative analyses of cultural characteristics as well as statistical evidence.

When a comparison needs to be made among the scores obtained from groups that have different characteristics, investigating the presence of differential item functioning (DIF) can enable more accurate comparisons regarding group differences (Galic et al., 2014). The present study examines the DIF and its sources in a scale measuring critical thinking motivation, which was developed in one culture and then adapted to the Turkish culture.

1.1. Critical Thinking and Motivation

Critical thinking as defined by Ennis (1996) is “reasonable, reflective thinking that is focused on deciding what to believe or do” (p. 166). French et al. (2014) also define critical thinking as “the conscious process a person does when s/he explores a situation or a problem from different perspectives” (p. 275). Critical thinking, therefore, enables an individual to solve a problem more effectively (Ennis, 1993) and also to produce more effective strategies when solving problems (Glevey, 2006), and thus facilitates lifelong learning skills (Halpern, 1998).

Ennis (1996) considers critical thinking dispositions as a component of critical thinking skills and emphasizes that critical thinking dispositions, such as “being open to alternatives”, should be accepted as part of the critical thinking skill. There are views in the literature supporting that critical thinking dispositions are essential for the use of critical thinking skills (Baron, 1985; Dewey, 1930; Ennis, 1991; Facione & Facione, 1992; McPeck, 1991; Paul, 1990; Perkins et al., 1993). Furthermore, it is claimed that motivation to think critically contributes to the use of critical thinking skills (Garcia & Pintrich, 1992; Ingle, 2007; Valenzuala et al., 2011). As reported in the literature motivation related beliefs and behaviors of both males and females are influenced by cultural stereotyping of gender roles (Meece et al., 2006). Studies on feelings of success and motivation have also revealed that males attribute their successes to their abilities; on the other hand, females attribute not their successes, but their failures to their abilities (Bar Tal, 1978, Crandall et al., 1965; Frieze, 1975). There are also views reported in the literature that, in areas culturally associated with gender, females are more disposed to experience learned helplessness when compared to males (Eccles et al., 1983, Farmer & Vispoel, 1990). On the other hand, a number of research findings also indicate that these gender related differences are not behavioral but only emerge in causal relationships (Eccles et al., 1983, Kloosterman, 1990; Parsons et al., 1984). Hence, it is important to examine affective factors related to cognitive skills in order to understand these constructs and their cultural associations.

With respect to characteristics regarding critical thinking, such as sustaining a discussion on a topic or refuting certain views, it is stated that females display a more accommodationist approach than males do. However, females are reported to display more behaviors than those of males in critically evaluating their own class performance (Feingold, 1994; Ruble et al., 1993). While some studies on critical thinking report gender differences (King et al., 1990; Serin et al., 2010), some others report no gender differences (Ersözlü & Arslan, 2009; McLean & Miller, 2010). French et al. (2012) claim that before such evaluations regarding these kinds of differences are made, it is important to examine measures for any indications of DIF.

While Ernst and Monroe (2004) stated that education has a positive impact on developing critical thinking, Tsui (2000) investigated how campus culture develops critical thinking and highlighted an increase in students' critical thinking skills and also dispositions in universities that support freedom of thinking and are run with a democratic understanding, whereas the condition in high school education where students are more passive and not made to engage actively in the learning process have a negative impact on the development of critical thinking. Taking such information into consideration together with other research findings and expert opinions, it can be said that the source of DIF in terms of the level of education variable could be language and expression.

Accordingly, in the present study, it has been considered that such a difference could emerge in tests measuring beliefs and perceptions related to cognitive skills such as critical thinking; thus, whether there were such gender and level of education related differences in the critical thinking, motivation test was investigated by means of DIF.

1.2. Differential Item Functioning

Differential item functioning emerges when individuals are at the same ability level but in different groups that have different probabilities of providing responses to items (Gierl et al., 1999). The concept of ability is defined in the Turkish Language Assosiation Updated Turkish Dictionary (n.d.) as an individual's attribute, capability, talent, or capacity to understand or to do. Based on this definition, it can be deduced that ability is more to do with the process of performing cognitive or psychomotor skills.

It may not be appropriate to use the concept of ability when defining DIF since when measuring affective features, the responses are based on individuals' self-reports, and there is no right or wrong behavior or response. Hence, as the scale in the present study measures an affective feature, the definition of DIF is operationalized as the differentiation in the response patterns given to some items by individuals at the same affective level but in different groups. Moreover, in the discussion on the findings obtained from DIF analyses, the concept critical thinking disposition level is used instead of ability level.

The presence of DIF in an item is believed to be a threat to construct validity (Jensen, 1980; Steinberg & Thissen, 2006). Thus, when DIF is found to be present in an item, it is recommended that the source should be investigated. This can be done by receiving expert opinions on the content of items with DIF in terms of, for example, conceptual or cultural features (Ateşok Deveci, 2008; Karakaya & Kutlu, 2012; Yıldırım & Büyüköztürk, 2018).

When studies on DIF and item bias in the related literature are examined, it is observed that DIF is mostly researched in tests measuring cognitive characteristics (e.g., French et al., 2014; Kurnaz & Kelecioğlu, 2008; Kurnaz Adıbatmaz & Yıldız, 2020; Maller, 2001; Stump et al., 2005; Yıldırım & Büyüköztürk, 2018), in national and international measurement tools (e.g., Altıntaş & Kutlu, 2019; Kalaycıoğlu & Kelecioğlu, 2011; Karakaya & Kutlu, 2012), and in studies on the development or adaptation of measurement tools (e.g., do Nascimento et al., 2021; Nielsen & Dammeyer, 2019). In recent years, the number of studies investigating DIF or item bias in measurement tools measuring affective characteristics (Gök et al., 2014; Garcia et al., 2021; Köse, 2015; Lau et al., 2020; Şengül Avşar & Emons, 2021; Usta, 2020) is becoming increasingly prevalent. It is believed that the present study will contribute to the literature in terms of DIF identification and the investigation of its sources based on data obtained from the administration of the measurement tool measuring an affective characteristic, namely critical thinking motivation.

In the Critical Thinking Motivation Scale used in the present study, the scores obtained from the items are evaluated with a mark ranging from 1 to 6: high scores indicate high critical thinking motivation levels. When DIF is found to be present in the items of the measurement tool, it is concluded that individuals at the same critical thinking motivation level but in different groups have a varying probability of providing responses to items. When this is the case, it is important that the items be examined for any expression or content that may be causing DIF. The findings of the present study can be instructive for researchers in two ways: first, if there are words and expressions that have an informative effect during the development or adaptation stage of a measurement tool, a finding can be generated on the discussion of how these can be eliminated; second, findings can be generated on whether results obtained from measurement tools create a difference stemming from items across the groups in terms of male and female scores or by level of education. In the measurement of cognitive or affective features, comparisons by gender and level of education are highly common; hence, the present study was designed to take into consideration the variables of gender and level of education.

In the present study, the responses to the following research questions were sought:

1. Do the items in the Critical Thinking Motivation Scale include DIF based on gender and level of education?
2. If there are items with DIF in the Critical Thinking Motivation Scale, how can the source of DIF in these items be accounted for?

2. METHOD

2.1. Study Group

In the present study, data were collected from 1050 individuals residing in various provinces in Türkiye and examined for univariate and multivariate outliers, while some part of the data were removed from the dataset in order to meet the fundamental statistical assumptions.

In total, data from 817 individuals were utilized in the DIF analysis. The age mean of the study group was 22.02±2.8 years. Of the participants, 47.5% were female, while 52.5% were male. The study group characteristics are presented in Table 1.

Table 1. *Study group characteristics.*

	Variable	Number	Percentage
Gender	Female	429	47.5
	Male	388	52.5
Province	İstanbul	92	11.3
	Ankara	92	11.3
	Karabük	80	9.8
	Konya	77	9.4
	Kastamonu	45	5.5
	Ağrı	39	4.8
	Mersin	36	4.4
	Afyon	32	3.9
	Bursa	46	5.6
	Çankırı	48	5.9
	Gaziantep	42	5.1
	Hatay	40	4.9
	Samsun	31	3.8
	Sakarya	44	5.4
	Other	73	8.9
Level of education	High school graduate	109	13.3
	University student	547	67.0
	University graduate	161	19.7

The data were collected from individuals residing in different provinces, namely İstanbul (11.3%), Ankara (11.3%), Konya (9.4%), and Karabük (9.8%). The collection of data from individuals living in different provinces is believed to increase the generalizability of the findings. In consideration of the measurement tool features, it was decided that the participants needed to be at least a high school graduate, which was set as a criterion in data collection. The study group was comprised of individuals who were high school graduates (n= 109, 13.3%), university students (n= 547, 67.0%), and university graduates (n=161, 19.7%).

2.2. Data Collection Tools

In the present study, the Critical Thinking Motivation Scale (Valenzuala Nieto & Saiz, 2011) adapted to the Turkish culture by Dönmez and Kaya (2016) was utilized. The Scale consisted of five subfactors, namely expectancy, attainment, intrinsic/interest value, utility, and cost and 19 items and the highest and lowest scores that could be obtained from the Scale were 114 and 19, respectively. The participants were expected to mark one of the six degrees of agreement in the Likert scale that they found most appropriate: (1 = “Strongly disagree”, 6 = “Strongly agree”), while the Scale did not have any items that required inverse marking.

The items in the Scale aimed to measure the participants’ expectations regarding critical and conscientious thinking (expectation) and the meaning they attributed to such thinking (value). The higher the total score obtained from the Scale was, the higher the participant’s critical thinking disposition (that is critical thinking expectation and value) was interpreted to be; conversely, the lower the total score of the participant was, the lower the participant’s critical thinking disposition (i.e. critical thinking expectation and value) was interpreted to be.

The scale was administered to 312 university students during its adaptation to the Turkish culture. The data collected from these participants were analyzed and the analysis results showed that all 19 items were categorized into five factors with eigenvalues values higher than 1 and they accounted for 67.91% of the total variance. The χ^2/df fit index value of the confirmatory factor analysis was 1.53. The NFI, CFI, and RMSEA were found to be 0.85, 0.94, and 0.58, respectively. Cronbach alpha coefficient of the scale was calculated between .73 and .85 for sub-dimensions and total score. These findings suggest that the research is valid at an acceptable level.

2.3. Data Collection and Analysis

The study was reported to be ethically appropriate in terms of ethical principles by the Karabük University Social and Human Sciences Research Ethics Committee (Decision number: E-78977401-050.02.04-49379). The items in the data collection tool and the questions found essential in the personal information form were used to develop an electronic Google form. This online form was sent to the participants, who voluntarily participated in the study.

The data were collected with the assistance of Karabük University students volunteering to contribute to the study. These students were asked to send the data collection form via Google forms to university students or high school graduates they knew. The collection of data via Google forms prevented the loss of data in the data set. Data were collected from 1050 individuals living in different provinces in Türkiye. However, during the stage of testing the fundamental assumptions, 233 data were removed from the data set after checking for the univariate and multivariate outliers.

It is recommended in the literature on scale adaptation that data obtained from the scale adapted should be checked for reliability in all the studies in which the scale is used. Hence, to check the reliability of the data obtained in the present study, the Cronbach alpha coefficient was calculated, and the reliability was found to be 0.88. The internal consistency of the sub-dimensions ranged between .76 and .80.

Prior to DIF analyses, unidimensionality and the normal distribution of the data were examined. To examine whether the data obtained from the measurement tool met the normality assumption, the skewness and kurtosis values were used. In the distribution, the skewness and kurtosis values were found to be 0.252 and -0.571, respectively; the standard error of skewness was calculated to be 0.086 and the standard error of kurtosis was 0.171. These values indicate that the distribution met the normality assumption (Büyüköztürk, 2021).

To examine the unidimensional outlier values in the distribution, *Z* standard scores were calculated for each item. All the items had *Z* standard scores ranging between 1.019 and -4.93. The unidimensional outliers in the distribution were eliminated, and after each outlier value was removed, the *Z* standard scores were recalculated for all the items and for all the participants. In the final data, the *Z* standard scores were found to range between 1.44 and -3.95. When the sample size is large, *Z* standard score that is ±3 is an expected condition. When this is the case, it is more appropriate to interpret the *Z* standard scores together with the mean, standard deviation, and the lowest and highest values (Tabachnick & Fidell, 2007).

On the other hand, multidimensional outliers were compared with the Mahalonobis distances (α=.001) and the critical chi-square value for K-1 degrees of freedom for all the items in the test of all the participants. The Mahalonobis distances ranged between the values of 1.11 and 113.6. At this stage, the data that showed deviation higher than the critical chi-square value was removed from the data set; subsequently, the *Z* score distributions and the Mahalonobis distances were reexamined. In the final data (N=817), the Mahalonobis distances were found to range between 42.2 and 1.72. The critical chi-square value for 18 degrees of freedom was 42.31. As there was no critical chi-square value exceeding the Mahalonobis distance, it could be concluded that there were no multidimensional outlier values in the data distribution (Mertler & Vannatta, 2005). On the other hand, kurtosis and skewness values were also calculated for all the items in order to check the multidimensional normality assumption, and these values were found to be between -1.2 and 0.89. It can therefore be said that each item is normally distributed separately and together.

After the removal of the unidimensional and multidimensional outlier values, which is essential for the administration of parametric tests that have multivariate data, other assumptions were tested. Exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) were performed with the data in order to examine unidimensionality in the distribution. The scatter plot obtained and the factors, the eigenvalues of the factors, and their contribution to the total variance are presented in Figure 1 and Table 2, respectively.

Figure 1. *Scatter plot.*

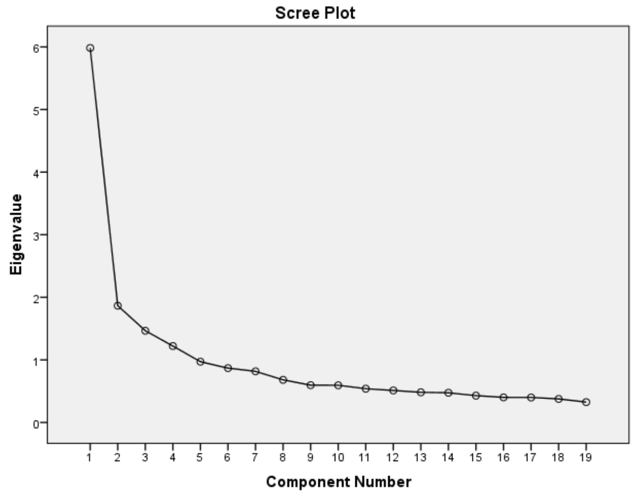


Table 2. *Factor eigenvalues and their contribution to the total variance based on the EFA results.*

Factor	Eigenvalue	Explained total variance
1	5.983	31.490
2	1.865	9.817
3	1.466	7.713
4	1.221	6.426
5	.971	5.111

It can be observed in the scatter plot in Figure 1 that while there is an abrupt fall after the slope of the first factor, the slope for the second and third factors has formed a plateau. The EFA results in Table 2 show that the difference between the eigenvalue of the first factor and the eigenvalue of the second value is higher than the differences between the eigenvalues of the other factors; the contribution of the first factor to the total variance is higher than the contribution of the other factors, which indicates that the unidimensionality assumption is met (Hambleton & Swaminathan, 1989). Meeting the unidimensionality assumption provides evidence for having met the local independence assumption (Hambleton et al., 1991).

The validity of the Critical Thinking Motivation Scale was checked with CFA for the sample group in this study and it was concluded that the measurement tool produced valid results (χ^2=590,415; χ^2/df=4,2; CFI=.92; GFI=.93; AGFI=.91; RMR=.045; NFI=.89; RMSEA=.06). Based on the evidence obtained as a result of the assumption checking analyses, it was concluded that the fundamental assumptions were met. The DIF analyses of the data collected were performed via the '*lordif*' function in the *R* '*lordif*' package (Choi et al., 2011). The lordif package was used because when DIF is identified in items that are scored across multiple categories and when there are more than two group variables, one of these variables is included in the model as a set of puppet variables. During the analysis process, the Generalized Partial Credit Model from the Item Response Theory was used (Muraki, 1992). In the Generalized Partial Credit Model, when DIF analysis is performed in the items to which weighted scoring is applied, the discriminatory parameters are also included in the model.

A form was developed to obtain expert opinions regarding the sources of DIF found to exist in some of the items; opinions were obtained from five measurement and evaluation experts, a child development expert who had worked on critical thinking, and a sociologist who had studied social classes and sexism. In the expert opinion form, an explanation of DIF was provided, the items with DIF and which groups these items favored were stated, and their opinions were asked about what the sources of the DIF could be. The results were interpreted and discussed based on these expert opinions.

3. RESULTS

The present study initially investigated whether there were items with DIF in the Critical Thinking Motivation Scale based on gender and level of education and then examined the sources of the items having DIF based on experts' opinions. Hence, this section is presented under two subtitles, which report results obtained from the analysis of DIF and results obtained from expert opinions regarding the sources of DIF.

3.1. The DIF Analysis Results

Whether or not the items displayed DIF based on gender and level of education was examined in the study and the results obtained are presented in Table 3.

Table 3. *DIF results based on the variables of gender and level of education.*

Item Number	Gender		Level of Education	
	Uniform (*p*)	Non-uniform (*p*)	Uniform (c^2)	Non-uniform (c^2)
1	0.065	0.422	0.017	**0.004***
2	0.934	0.962	**0.005***	0.651
3	0.011	0.105	**0.003***	0.575
4	0.325	0.863	**0.001***	0.130
5	**0.005***	0.021	0.692	0.447
6	0.034	0.357	0.313	0.515
7	**0.006***	0.820	0.646	0.851
8	0.199	0.379	0.472	0.748
9	0.746	0.234	**0.006***	0.025
10	0.237	0.311	**0.001***	0.557
11	0.932	0.745	0.057	0.682
12	0.502	0.844	**0.002***	0.104
13	0.308	0.986	**0.008***	0.980
14	0.071	0.562	**0.008***	0.308
15	**0.006***	0.929	0.790	0.669
16	0.021	0.288	0.090	0.773
17	0.084	0.150	0.651	0.064
18	0.197	0.038	0.612	0.415
19	0.887	0.720	0.618	0.661

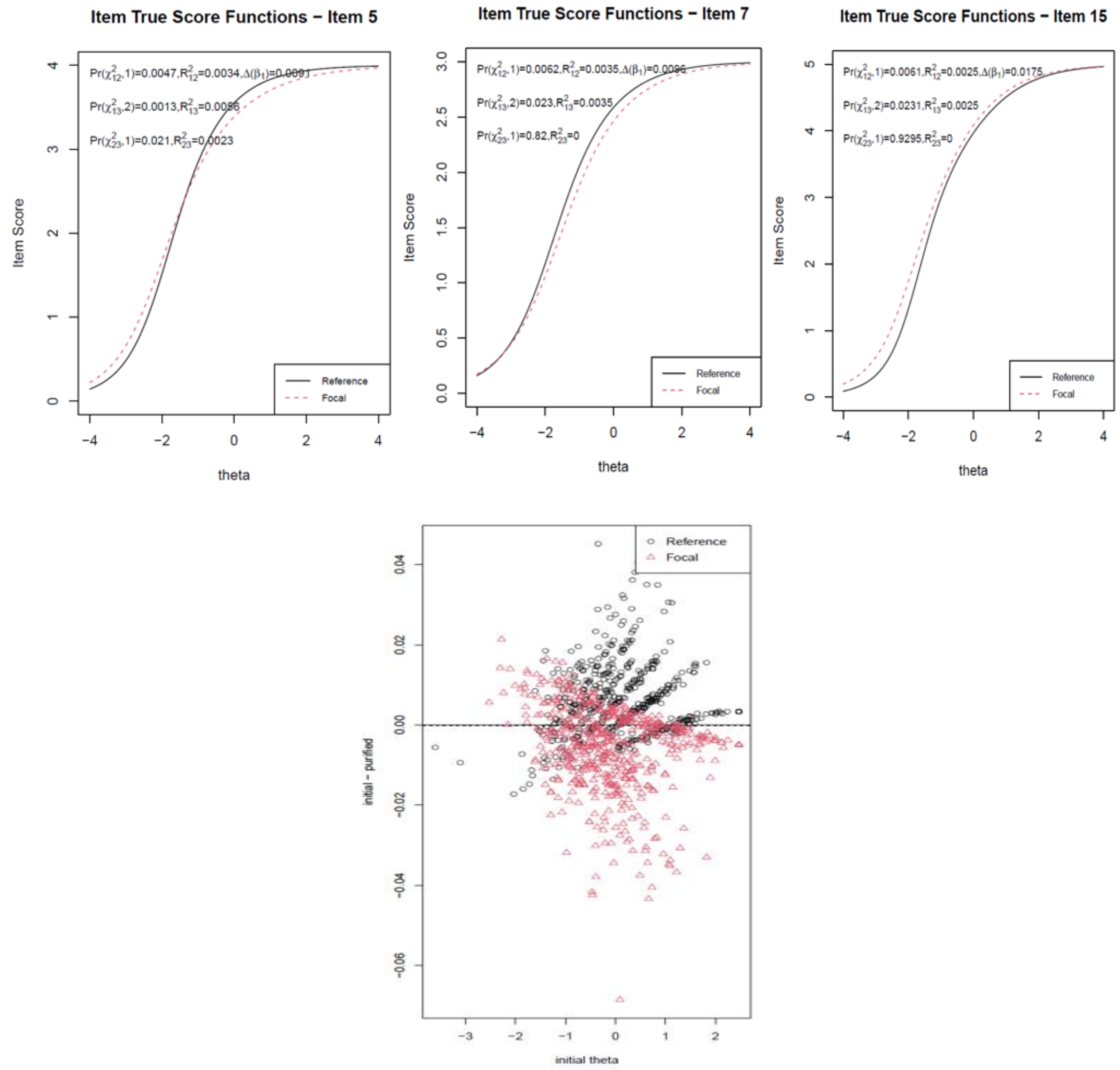
* Identification of DIF at .01 significance level

In the present study, the analyses based on the gender variable yielded three items with DIF, namely Items 5, 7, and 15. On the other hand, the analyses based on level of education yielded a total of nine items with DIF, namely Items 1, 2, 3, 4, 9, 10, 12, 13, and 14.

The scatter plot of the difference between the test characteristic curves of the items identified with DIF based on the gender variable and the predicted levels of critical thinking dispositions after the items with DIF were removed from the test is displayed in Figure 2.

When the test characteristic curves of the items with DIF are examined in Figure 2, it can be revealed that the items displaying DIF based on gender were in favor of female participants with low levels of critical thinking dispositions. In the scatter plot depicting the differences among the predicted critical thinking dispositions levels after items with DIF were removed from the test, the values on the y axis represent the difference between the predicted critical thinking disposition levels obtained from the entire scale and the critical thinking disposition levels after the items with DIF were removed from the test. It can be stated that individuals with a positive value on the vertical axis were influenced negatively from the items with DIF, while those with a negative value were positively influenced by the items with DIF. Accordingly, it can be said that items with DIF generally functioned in favor of female participants.

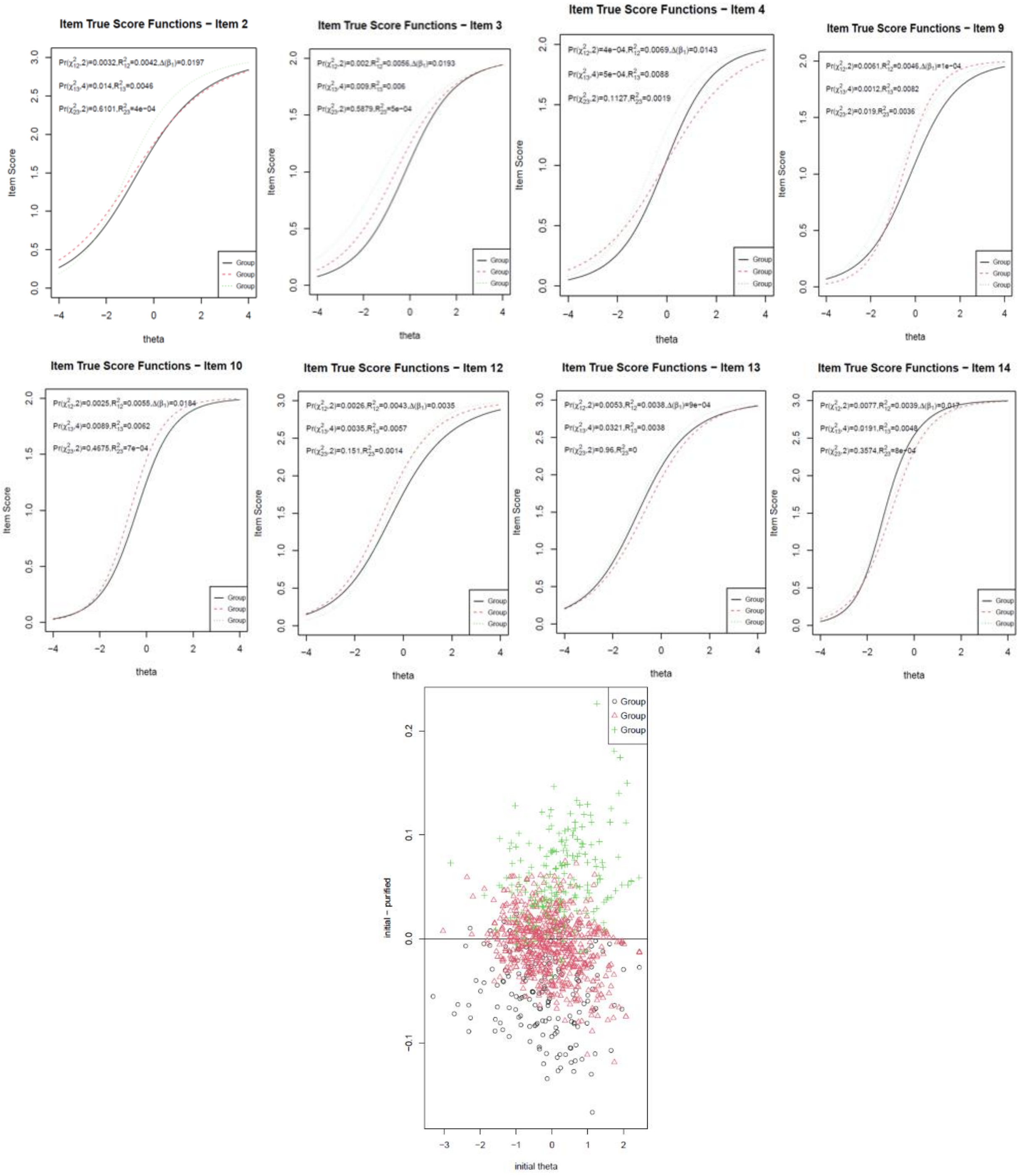
Figure 2. *The scatter plot of the difference between the test characteristic curves of the items identified with DIF based on the gender variable and the predicted levels of critical thinking dispositions after the items with DIF were removed from the test.*



The scatter plot of the difference between the test characteristic curves of the items identified with DIF based on the level of education variable and the predicted levels of critical thinking dispositions after the items with DIF were removed from the test is displayed in Figure 3. In the graphs, Group 1 represents individuals who are high school graduates and work in a job; Group 2 represents the university students; and Group 3 represents university graduates who have an occupation.

When the test characteristic curves of items identified as having DIF are examined in Figure 3, it can be observed that Items 1, 2, 3, 4, and 13 function in favor of university graduates, Items 9, 10 and 12 function in favor of university students, and Item 14 functions in favor of high school graduates. One other finding that was obtained was that results varied in items at low and high critical thinking disposition levels. It was generally observed that in items with DIF, the difference between university graduates and high school graduates was larger. Since the identification of the sources of DIF in the related items may provide important information to researchers who develop or adapt scales, the content of the items with DIF examined by the experts and the results obtained are provided in Figure 3.

Figure 3. *The scatterplot of the difference between the test characteristic curves of the items identified with DIF based on the level of education variable and the predicted levels of critical thinking dispositions after the items with DIF were removed from the test.*



3.2. Results on Expert Opinions on DIF Resources

The items identified as having DIF were examined in terms of item bias based on expert opinion. The experts were asked whether the items with DIF based on gender/level of education constituted a source of bias. The items with DIF by gender and level of education are presented in Table 4.

It was revealed that expert opinions had two foci as regards the source of DIF in items with DIF based on the gender factor. The first was that the ways of expression in some items in the measurement tool (e.g., reasoning correctly) could lead to DIF. Second, in Items 5, 7 and 12, expressions such as "…learning is important", "For me it is important to use my intellectual skills", …I like to think" could have increased women's inclination to provide "the response expected by the environment".

Table 4. *Items identified to have DIF.*

Variable	Item number	Sub- factor	Item with DIF	Group
Gender	5	Attainment	For me it is important to learn how to reason correctly.	In favor of women
	7	Attainment	For me it is important to use my intellectual skills.	
	15	Utility	I like to think critically.	
Level of education	1	Expectancy	Concerning reasoning correctly, I am better than most of my peers.	In favor of university graduates
	2	Expectancy	I am capable of understanding everything related to thinking in a rigorous way.	
	3	Expectancy	I am able to learn how to think in a rigorous way.	
	4	Expectancy	I am able to learn how to reason correctly better than most of my peers.	
	13	Utility	I like to reason properly before deciding about something.	
	9	Intrinsic value/ interest	Thinking critically will help me to become a good professional.	In favor of university students
	10	Intrinsic value/ interest	Thinking critically will be useful for my future.	
	12	Intrinsic value/ interest	Thinking critically is useful for other subjects and courses.	
	14	Utility	I like to learn things that will improve my way of thinking.	In favor of high school graduates

In items with DIF based on the subfactor of level of education, it was revealed that expert opinions regarding sources of DIF had three foci. The first opinion was that the expressions of some of the items in the measurement tool (e.g., reasoning correctly, being a good professional, how to think in a rigorous way) could be the source of DIF. The second was that being a university student or being a university graduate could increase individuals' motivation to think critically. The third opinion was the probability of high school graduates' refraining from answering items with high scores when the content was based on such expressions as being better or being a professional. Such findings are addressed in the discussion section in detail with samples from expert opinions.

According to expert opinions, the formation of DIF in three items (Items 5, 7, and 15) based on gender can be attributed to the fact that women with low critical thinking dispositions have a high tendency to meet societal expectations. Below are direct quotations from experts' views regarding this issue:

> *"that women need to develop correct reasoning skills to free themselves from the secondary position they are in when compared to men is a social reality. That women who do not learn to reason correctly will be eliminated from the system faster than men has been*

> *engrained into women's mind as a cultural code. Conversely, the errors that men make in society or their incorrect reasonings are tolerated more when compared to those of women."*

Expert A

> "*Regarding this topic, the metaphor of "leaking pipe" explains this topic in more detail. According to this approach, as a result of the challenges women face, they are eliminated within the process. Women who do not want to be eliminated must learn to think more accurately. For women, critical thinking is an important step to move out of the patriarchal system they are a part of. It is by this means that they can question the system and can struggle to raise themselves to the position they 'desire/deserve'*."

Expert D

Moreover, based on these findings, it can be highlighted that the adaptation of a scale to a new culture does not merely consist of psychometric calculations, and thus examining the cultural background of the measurement tool being adapted is important. In terms of level of education, the experts were of the common opinion that being a university student, or a university graduate could increase their motivation to think critically. Direct quotations from experts' opinions on the topic are provided as follows:

> *"…it reveals that not only education, but the university environment is also influential in the development of critical thinking. By creating a learning environment where students are encouraged to participate in discussions and debates on social and political topics, it appears that a campus culture with social and political awareness is conducive to development of critical thinking skills. The factor underlying the fact that university graduates evaluate the item with a high score when compared to high school graduates at the same ability level is not only about level of education but also the learning environment and the campus culture, which should not be disregarded."*

Expert A

> *"Thus, it could be that university graduates felt a higher need for thinking skills and the need to think. It is known that when compared to other people, those with a high need to think are more realistic in terms of their self-predictions. And when I look at the items here it seems that people were asked to make predictions about their own performance regarding critical thinking. University graduates could be more conscious about this as well."*

Expert B

> *"Reasoning correctly. "It doesn't look appropriate to the Turkish language structure to me. "Does it mean evaluating events accurately? How will inaccurate reasoning occur? These could stem from the unclarity of the expressions, from the university graduates' getting a different meaning from the item.*

Expert C

4. DISCUSSION and CONCLUSION

The accuracy of the evaluation of the results obtained from the administration of a measurement tool depends on the aim of the measurement tool in subject and in its technical adequacy (Glover & Albers, 2007). When a measurement tool developed in one culture is adapted to another culture, the linguistic and cultural differences between the respondents can substantially threaten the validity and the psychometric properties of the measurement tool (Hambleton et al., 2004). When measurement tools are adapted, in some circumstances, words or phrases used in the developed and adapted tools do not convey the same meaning either linguistically or culturally. When such a condition is present, the equivalence of the original and the adapted form is distorted, and the validity of the adapted tool becomes questionable (Poortinga, 1989).

The items and item content of adapted scales are expected to accurately reflect the differences between subgroups in the target culture. Examination of changing item function or item bias is a common way to investigate such differences. If the response behavior for an item varies between two individuals from the same culture who have the same level of the measured feature, and if this creates variance against or in favor of one of the groups, then this can cause wrong decisions to be made in between-group comparisons. DIF can provide crucial information to test developers or adaptors to identify such conditions and to investigate the sources of DIF.

In a study by Gallos (1995) it is reported that there is a significant relationship between critical thinking and gender, and that the reason underlying this is a learning environment that is in favor of males; it is also stated that females have more doubts than males have about their abilities/talents and intellectual competences; when females encounter failure, they more often impose the causes of failure upon themselves, while males do so on external conditions; and it was revealed that females are less likely than men to initiate small learning groups and to participate in these; however, when they are encouraged to do so, they are as successful as males.

Taking into consideration experts' opinions as well as the findings reported in the study by Gallos (1995), the reason why the items with DIF in the present study that are in favor of females at the same level in terms of the feature measured could be related to cultural codes and gender based cultural experiences. The non-existence of DIF in the other levels of the measured feature – that DIF only existed in low levels – could be attributed to the fact that women at low levels regarding the measured feature could have changed the meaning they derived from the items or caused a social acceptance error.

In items measuring affective features, the individual reads the items, attributes meaning to them, and then selects the item found most appropriate. As in maximum success tests, there is no response that is the most accurate nor an expected response. The responses are based on what the respondent finds appropriate. Hence, when interpreting the item, the individual is expected to remain completely independent of social norms or social doctrines; however, this may not be an easy task for test implementers or evaluators in real life. In this case, when writing items, many elements, such as social doctrines, collective subconscious, and culture need to be taken into consideration, and the feature measured through items should be freed of these contexts. To illustrate, in the fifth item (For me, reasoning correctly is important), a female respondent who has a low level of the measured feature can be disposed to select 'strongly agree' in an item to meet societal expectations; that is because she accepts the society's expectations of her to provide the correct response. If individuals in different groups (e.g., men and women) who have the same level of the measured trait understand the item or the meaning they attribute to the item changes, it can be said that the item does not represent the construct to the same degree in these groups (Davidov et al., 2014; Millsap, 2012).

Schwartz and Meyer (2010) state that all research areas are influenced by cultural practices (e.g., language, traditions), cultural values (e.g., individual versus group), and cultural identity (e.g., allegiance to a particular group). At the outset, it is important to examine how the motivation for critical thinking differs in the cultural context between men and women, as well as from those with higher to those with lower levels of education. In this respect, it is important to examine the psychological and sociological contexts of test development or adaptation processes and to examine what the meanings attributed to the language used in the items mean for individuals in different groups. The development or adaptation of a measurement tool is an effort to find the best meaning to represent the measured construct.

Kholberg (1973) stated that the majority of female participants displayed a moral tendency to be a 'good child' in terms of the responses given to conflict entailing questions in the moral development theory; Gilligan (1979) attributed this to cultural doctrines. In the phase of 'being a good child' in Kholberg's moral development theory, the individual tends to display behaviors accepted to be appropriate by the society in order to get others' approval. It would not be wrong to state that the stories that entail conflicts in Kholberg's theory requires critical thinking and critical evaluation. Hence, the results from Kholberg (1973) and Gilligan (1979) support the findings obtained from the present study.

When Tedesco was writing about her book titled *Women's Ways of Knowing* in 1991, she stated that women believed that language was not dependable, that they experienced difficulties in expressing their self-identity and preferred to remain silent, that women who possessed learned knowledge did not believe they could provide the correct responses, and that they would echo others' voices rather than express their own; she stated that apart from those whom they decide to be the same in terms of background, conditions or views, they were generally reluctant to share their inner world with others. Considering this, it can be stated that in items measuring females' affective features, there is an important cultural process, and that this cultural process should be carefully examined when creating items in a test or scale.

When Jensen (1980) explained the relationship between culture, language, and test bias, s/he explained culture sterility of a test as 'distance from culture' and stated that when a measure tool is translated into another language, it will have a different content and the meaning attributed to the items may vary. Considering that the groups responding to the items are from different subcultures in terms of gender, level of education etc., it may be important in terms of the construct validity of measurement tools to be reconstructed so that items with DIF convey the same meaning to all the subgroups.

Hambelton and Rogers (1995) stated that to prevent items in a test from creating bias in favor of/against a prevalent culture or subcultures, the following questions need to be answered:

(1) Does the item include words that express different meanings to different sub sociocultural groups or words that are unfamiliar to those subgroups?
(2) Does the item include words that are difficult to understand?
(3) Does the item include words that are peculiar to a certain region or words that are not used frequently across the country?

When this information and the expert opinions in the present study are examined in combination, it can be concluded that there may be content that causes DIF in the language and expressions of the items. It is possible to state that an examination of the items with DIF revealed that university graduates, when compared to the other education level participants but with the same level of the feature being measured, had more often marked the options that yielded higher scores in items such as "*…I am better than most of my peers*", "*…I find myself proficient*", and "*I like to reason before I decide about something*." As for the university students, they more often marked the higher end of the Likert scale in items when compared to the other participants with the same level of the feature being measured in items such as "*…it will help me become a good professional*", and "*…it will be helpful for my future*." On the other hand, high school graduates, when compared to the other participants with the same level of features being measured, seemed to mark the 'strongly agree' option more often in the item that read '*I like to learn things that will improve my way of thinking'*. The respondents' item response behaviors seem to be related to how they perceive themselves based on their level of education and what they expect from themselves based on their social status. This could indicate that when the content of items is interpreted, individuals create meaning based on their social status and what is expected of them; this can create a difference in the scores of individuals in different groups but with the same level of critical thinking disposition.

Lau et al. (2023) administered a scale measuring gelotophobia, gelotophilia, and catagelasticism to university students in Taiwan and Canada. The Canadian English version was adapted from the German version. English version was then adapted into Taiwanese Chinese. While there were no items with DIF in the data obtained from the Canada sample, five items with DIF were found in the Taiwan sample. Only one of these items had a significant level. Then, in the data collected from Canada, DIF was calculated for the subgroups defined as Chinese living in Canada and answering the English form. In the data obtained from the English form, it was determined that there was no DIF for this subgroup and the reason for the DIF in the item was explained by the meaning changes in the words during the translation process. These results obtained from the study of Lau et al. (2023) confirm the argument of this study. In the adapted tests, it can be said that the translation processes and the meanings of the items affect the power to represent the construct.

Osterlind (1983) and Jensen (1980) highlighted that DIF in items or item bias can be caused by external factors such as culture and environment. Accordingly, based on the results of the present study, it can be valid to say that there may be external bias causing DIF, but the language and expressions in the measurement tool also increase the probability of DIF in the related items.

Considering the results of the present study, it can be said that validity evidence based solely on translation processes and psychometric computations of the measurement tools adapted to the Turkish culture may not be sufficient. In data obtained from the administration of developed or adapted measurement tools, investigating DIF can also yield significant evidence regarding the validity of a scale. Furthermore, it can be argued that it is important to examine the nature of the impact of how items are understood in the sub cultural groups by receiving opinions of experts in such areas as sociology and psychology.

One of the limitations of this research is that most of the data collected in this study were from university students. It is not known in which direction increasing the number of high school students and high school graduates would change the results. Since the research is based on individuals' self-report, it is assumed that the participants answered the items sincerely and accurately and that their reading comprehension skills were at a similar level. The evidence of reliability and validity in the study confirms these assumptions.

Researchers can examine DIF in tools measuring different affective features. In achievement tests and tests measuring affective features, respondent behaviors will show variation based on the structure of the feature being measured. Hence, in tools measuring affective features, investigating DIF can lead to different results. In addition, two different tools measuring critical thinking and critical thinking motivation can be administered to the same group, and the scores obtained from the achievement test can be used as an external criterion. In this way, findings based on the relationship between real performance and the affective feature related to the performance can be obtained.

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**: Karabük University, 28/07/2021, 2021/07-05.

Authorship Contribution Statement

Fatma Betül Kurnaz: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing-original draft. **Hüseyin Yıldız**: Methodology, Software, Formal Analysis.

Orcid

Fatma Betül KURNAZ https://orcid.org/0000-0002-7042-2159
Hüseyin YILDIZ https://orcid.org/0000-0003-2387-263X

REFERENCES

Altıntaş, Ö., & Kutlu, Ö. (2019). Investigating differential item functioning of Ankara University examination for foreign students by recursive partitioning analysis in the Rasch model. *International Journal of Assessment Tools in Education, 6*(4), 602–616. https://doi.org/10.21449/ijate.554212

Ateşok Deveci, N. (2008). *Examination of Inter-university Board foreign language test in the frame of item bias*. [Doctoral dissertation, Ankara University]. https://tez.yok.gov.tr

Athman Ernst, J., & Monroe, M. (2004). The effects of environment-based education on students' critical thinking skills and disposition toward critical thinking. *Environmental Education Research, 10*(4), 507-522. https://doi.org/10.1080/1350462042000291038

Bar-Tal, D. (1978). Attributional analysis of achievement-related behavior. *Review of Educational Research, 48*(2), 259–271. https://doi.org/10.3102/00346543048002259

Baron, J. (1985). *Rationality and intelligence*. Cambridge University Press.

Büyüköztürk, Ş. (2021). *Sosyal bilimler için veri analizi el kitabı* [Data analysis handbook for social sciences]. PegemA.

Byrne, B.M., Oakland, T., Leong, F.T.L., Van De Vijver, F.J.R., Hambleton, R.K., Cheung, F. M., & Bartram, D. (2009). A critical analysis of cross-cultural research and testing practices: Implications for improved education and training in psychology. *Training and Education in Professional Psychology, 3*, 94-105. https://doi.org/10.1037/a0014516

Choi, S.W., Gibbons, L.E., & Crane, P.K. (2011). Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software, 39*(8), 1. https://doi.org/10.18637/jss.v039.i08

Cole, D.A., Maxwell, S.E., Avery, R., & Salas, E. (1993). Multivariate group comparisons of variable systems: MANOVA and structural equation modeling. *Psychological Bulletin, 114*, 174-184. https://doi.org/10.1037/0033-2909.114.1.174

Crandall, V.C., Katkovsky, W., & Crandall, V.J. (1965). Children's belief in their own control of reinforcement in intellectual-academic achievement situations. *Child Development, 36*, 91–109. https://doi.org/10.2307/1126783

Crane, P.K., Gibbons, L.E., Jolley, L., & Van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques: DIF detect and difwithpar. *Medical Care*, *44*(11), 115-123. https://www.jstor.org/stable/41219511

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Thomson Learning.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology, 40*, 55-75. https://doi.org/10.1146/annurev-soc-071913-043137

Dewey, J. (1930). *Human nature and conduct*. The Modern Library.

do Nascimento, C.D., Peter, W.F., Ribeiro, I.M., Moreira, B.S., Lima, V.P., Kirkwood, R.N., & Bastone, A.C. (2021). Cross-cultural validity of the animated activity questionnaire for patients with hip and knee osteoarthritis: A comparison between the Netherlands and Brazil, *Brazilian Journal of Physical Therapy, 25*(6):767-774. http://doi.org/10.1016/j.bjpt.2021.06.002

Dönmez, B., & Kaya, F. (2016). Eleştirel Düşünme Motivasyonu Ölçeği'nin Türkçe'ye uyarlanması [Turkish adaptation study of Critical Thinking Motivational Scale]. *HAYEF*

Journal of Education, 13-2(25), 159-173. https://dergipark.org.tr/tr/pub/iuhayefd/issue/24491/259590

Eccles, J.S., Adler, T.F., Futterman, R., Goff, S.B., Kaczala, C.M., & Meece, J.L. (1983). Expectancies, values and academic behaviors. In J.T. Spence (Ed.), *Achievement and Achievement Motives* (pp. 75–146). San Francisco Freeman.

Ennis, R.H. (1991). Critical thinking: A streamlined conception. *Teaching Philosophy, 14*, 5-25. http://doi.org/10.1057/9781137378057_2

Ennis, R.H. (1993). Critical thinking assessment. *Theory into Practice, 32*, 179-186. https://doi.org/10.1080/00405849309543594

Ennis, R.H. (1996). Critical thinking dispositions: Their nature and assessability. *Informal Logic*, *18*(2), 165-182. https://doi.org/10.22329/il.v18i2.2378

Ersözlü, A. N., & Arslan, M. (2009). The effect of developing reflective thinking on metacognitional awareness at primary education level in Turkey. *Reflective Practice, 10*, 683–695. https://doi.org/10.1080/14623940903290752

Facione, P.A., & Facione, N.C. (1992). *The California critical thinking dispositions inventory*. California Academic Press.

Farmer, H.S., & Vispoel, W.P. (1990). Attributions of female and male adolescents for real-life failure experiences. *Journal of Experimental Education*, *58*(2), 127–140. https://doi.org/10.1080/00220973.1990.10806529

Feingold, A. (1994). Gender differences in personality: A meta analysis. *Psychological Bulletin*, *116*, 429-456. https://doi.org/10.1037/0033-2909.116.3.429

Ferne, T., & Rupp, A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, *4*, 113-148. https://doi.org/10.1080/15434300701375923

French, B.F., Hand, B., Nam, J., Yen, H.J., & Vazquez, J.A.V. (2014). Detection of differential item functioning in the Cornell Critical Thinking Test across Korean and North American students. *Psychological Test and Assessment Modeling*, *56*(3), 275.

French, B.F., Hand, B., Therrien, W.J., & Vazquez, J.A.V. (2012). Detection of sex differential item functioning in the Cornell Critical Thinking Test. *Europen Journal of Psychological Assessment*, *28*(3), 201-207. http://doi.org/10.1027/1015-5759/a000127

Frieze, I.H. (1975). Women's expectations for and causal attributions of success and failure. In T. Mednick, S. Tangi, & L.W. Hoffman (Eds.), *Women and achievement. Social and motivational analysis.* (pp. 158–171). John Wiley and Sons.

Galic, Z., Scherer, K.T., & Leberton, J.M. (2014). Examining the measurement equivalence of the conditional reasoning test for aggression across U.S. and Croatian samples. *Psychological Test and Assessment Modeling, 56*, 195-216. http://darhiv.ffzg.unizg.hr/id/eprint/5547

Gallos, J.V. (1995). Gender and silence. *Collage Teaching*, *43*(3), 101-105. http://doi.org/10.1080/87567555.1995.9925525

Garcia, J.M., Gallagher, M.W., O'Bryant, S.E., & Medina, L.D. (2021). Differential item functioning of the Beck Anxiety Inventory in a rural, multi-ethnic cohort. *Journal of Affective Disorders, 293*, 36-42. http://doi.org/10.1016/j.jad.2021.06.005

Garcia, T., & Pintrich, P.R. (1992). The effect of PBL curriculum on students' motivation and self-regulation. Paper presented at The Biennial Conference of The European Association for Research on Learning and Instruction, Italy.

Gierl, M., Khaliq, S.N., & Boughton, K. (1999). Gender differential item functioning in mathematics and science: Prevalence and policy implications. In Improving Large-Scale Assessment in Education Symposium at the Annual Meeting of the Canadian Society for the Study of Education, Canada.

Gilligan, C. (1979). Woman's place in man's life cycle. *Harvard Educational Review, 49*, 431-446. https://doi.org/10.17763/haer.49.4.h136573541l3g463

Glevey, K.E. (2006). Promoting thinking skills in education. *London Review of Education, 4*(3), 291-302. http://doi.org/10.1080/14748460601044005

Glover, T.A., & Albers, C.A. (2007). Consideraitons for evaluating universal screening assessments. *Journal of School Psychology, 45*(2), 117-135. http://doi.org/10.1016/j.jsp.2006.05.005

Gök, B., Kabasakal, K.A., & Kelecioğlu, H. (2014). PISA 2009 öğrenci anketi tutum maddelerinin kültüre göre değişen madde fonksiyonu açısından incelenmesi [Analysis of attitude items in PISA 2009 student questionnaire in terms of differential item functioning based on culture]. *Journal of Measurement and Evaluation in Education and Psychology, 5*(1), 72-87. https://doi.org/10.21031/epod.64124

Halpern, D.F. (1998). Teaching critical thinking for transfer across domains. *American Psychologist, 53*, 449-455. https://doi.org/10.1037/0003-066X.53.4.449

Hambelton, R., & Rogers, J. (1995). Item bias review. *Practical Assessment, Research, and Evaluation, 4*(6), https://doi.org/10.7275/jymp-md73

Hambleton, R.K., & Rovinelli, R.J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10*(3), 287-302. https://doi.org/10.1177/014662168601000307

Hambleton, R.K., & Swaminathan, H. (1989). *Item response theory: Principles and applications*. Kluwer Nijhoff.

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

Ingle, C.O. (2007). *Predictors of critical thinking ability among college students*. [Doctoral dissertation]. Available from ProQuest Dissertations and Theses Database. (UMI No. 3263681).

Jensen, A.R. (1980). *Bias in mental testing*. Free Press.

Kalaycioğlu, D.B., & Kelecioğlu, H. (2011). Öğrenci Seçme Sınavı'nın madde yanlılığı açısından incelenmesi [Item Bias Analysis of the University Entrance Examination]. *Education and Science, 36*(161), 3-11.

Karakaya, İ., & Kutlu, Ö. (2012). Seviye Belirleme Sınavındaki Türkçe alt testlerinin madde yanlılığının incelenmesi [An Investigation of Item Bias in Turkish Sub Tests in Level Determination Exam]. *Education and Science, 37*(165). http://egitimvebilim.ted.org.tr/index.php/EB/article/view/1342

King, P.M., Wood, P.K., & Mines, R.A. (1990). Critical thinking among college and graduate students. *The Review of Higher Education, 13*, 167-186. http://doi.org/10.1353/rhe.1990.0026

Kholberg, L. (1973). Continuities and discontinuities in childhood and adult moral development revisited. In P.B. Baltes & L.R. Goutlet (Eds.), *Lifespan developmental psychology: Research and theory* (pp. 179-204). Academic Press.

Kloosterman, P. (2001). Attributions, performance following failure, and motivation in mathematics. In E. Fennema & G.C. Leder (Eds.), *Mathematics and gender*. Teachers College Press.

Köse, İ.A. (2015). PISA 2009 öğrenci anketi alt ölçeklerinde (Q32-Q33) bulunan maddelerin değişen madde fonksiyonu açısından incelenmesi [Investigation of items in PISA 2009 student questionnaire subscales (Q32-Q33) in terms of differential item functioning]. *Kastamonu Education Journal, 23*(1), 227-240. https://dergipark.org.tr/en/pub/kefdergi/issue/22600/241461

Kurnaz, F.B., & Kelecioğlu, H. (2008). Investigation of Peabody Picture Vocabulary Test from the point of item bias. *World Applied Sciences Journal, 3*(2), 231-239.

https://www.academia.edu/7282678/Investigation_of_Peabody_Picture_Vocabulary_Test_from_the_point_of_item_bias_peabody_picture_vocabulary_test

Kurnaz Adıbatmaz, F.B., & Yıldız, H. (2020). The Effects of distractors to differential item functioning in Peabody Picture Vocabulary Test. *Journal of Theoretical Educational Science, 13*(3), 530-547. https://dergipark.org.tr/tr/pub/akukeg/issue/54987/621581

Lau, C., Chiesi, F., Saklofske, D.H., Yan, G., & Li, C. (2020). How essential is the essential resilience scale? Differential item functioning of Chinese and English versions and criterion validity. *Personality and Individual Differences, 155*, 109666. http://doi.org/10.1016/j.paid.2019.109666

Lau, C., Swindall, T., Chiesi, F., Quilty, L.C., Chen, H.C., Chan, Y.C., ... & Torres-Marín, J. (2023). Cultural differences in how people deal with ridicule and laughter: Differential item functioning between the Taiwanese Chinese and Canadian English versions of the PhoPhiKat-45. *European Journal of Investigation in Health, Psychology and Education, 13*(2), 238-258. https://doi.org/10.3390/ejihpe13020019

Maller, S.J. (2001). Differential item functioning in the WISC-III: Item parameters for boys and girls in the national standardization sample. *Educational and Psychological Measurement, 61*(5), 793-817. https://doi.org/10.1177/00131640121971527

McLean, C.P., & Miller, N.A. (2010). Changes in critical thinking skills following a course on science and pseudoscience: A quasi-experimental study. *Teaching of Psychology, 37*, 85-90. https://doi.org/10.1080/00986281003626714

Mcpeck, J.E. (1990). Teaching critical thinking. Routledge.

Meece, J.L., Glienke, B.B., & Burg, S. (2006). Gender and motivation. *Journal of School Psychology, 44*(5), 351-373. https://doi.org/10.1016/j.jsp.2006.04.004

Mertler, C.A., Vannatta, R.A., & LaVenia, K.N. (2005). *Advanced and multivariate statistical methods: Practical application and interpretation*. Pyrczak. https://doi.org/10.4324/9781003047223

Millsap, R.E. (2012). *Statistical approaches to measurement invariance*. Routledge.

Muraki, E. (1992). *A generalized partial credıt model: Applicatıon of an em algorithm*. ETS Research Report Series. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Nielsen, T., & Dammeyer, J. (2019). Measuring higher education students' perceived stress: An IRT-based construct validity study of the PSS-10. *Studies in Educational Evaluation, 63*, 17-25. http://doi.org/10.1016/j.stueduc.2019.06.007

Osterlind, S.J. (1983). *Test item bias*. Sage.

Parsons, J., Adler, T.F., & Kaczala, C.M. (1984). Socialization of achievement attitudes and beliefs: Parental influences. *Child Development, 53*, 322-339. https://doi.org/10.2307/1128973

Paul, R.W. (1990). *Critical thinking: What every person needs to survive in a rapidly changing world*. Center for Critical Thinking and Moral Critique, Sonoma State University.

Perkins, D.N., Jay, E., & Tishman, S. (1993). Beyond abilities: A dispositional theory of thinking. *Merrill-Palmer Quarterly, 39*(1), 1-21. https://www.jstor.org/stable/23087298

Poortinga, Y.H. (1989). Equivalence of cross-cultural data: An overview of basic issues. *International Journal of Psychology, 24*(6), 737-756. https://doi.org/10.1080/00207598908247842

Ruble, D., Greulich, F., Pomerantz, E.M., & Gochberg, B. (1993). The role of gender-related processes in the development of sex differences in self-evaluation and depression. *Journal of Affective Disorders, 29*(1), 97-128. https://doi.org/10.1016/0165-0327(93)90027-H

Serin, Q., Serin, N.B., Saracaloğlu, A.S., & Ceylan, A. (2010). The examination of critical thinking styles of university students (TRNC Sample). *Procedia Social and Behavioral Sciences, 9*, 864–868. https://doi.org/10.1016/j.sbspro.2010.12.250

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: examples using item response theory to analyze differential item functioning. *Psychological Methods*, *11*, 402-415. https://doi.org/10.1037/1082-989X.11.4.402

Stump, T.E., Monahan, P., & Mchorney, C.A. (2005). Differential item functioning in the short portable mental status questionnaire. *Research on Aging*, *27*(3), 355-384. https://doi.org/10.1177/0164027504273784

Schwartz, S., & Meyer, I.H. (2010). Mental health disparities research: The impact of within and between group analyses on tests of social stress hypotheses. *Social Science & Medicine*, *70*(8), 1111-1118. https://doi.org/10.1016/j.socscimed.2009.11.032

Şengül Avşar, A., & Emons, W.H.M. (2021). A cross-cultural comparison of non-cognitive outputs towards science between Turkish and Dutch students taking into account detected person misfit. *Studies in Educational Evaluation*, *70*(101053). http://doi.org/10.1016/j.stueduc.2021.101053

Tabachnick, B.G., & Fidell, L.S. (2007). *Using multivariate statistics*. Pearson.

Tedesco, J. (1991). Women's ways of knowing/women's ways of composing. *Rhetoric Review*, *9*(2), 246-256. http://doi.org/10.1080/07350199109388931

Tsui, L. (2000). Effects of campus culture on students' critical thinking. *The Review of Higher Education*, *23*(4), 421-441. http://doi.org/10.1353/rhe.2000.0020

Turkish Language Assosiation (n.d.). Ability. In Updated Turkish Dictionary. Retrieved February 28, 2021. https://www.tdk.gov.tr/

Usta, H.G. (2020). Sınav kaygı ölçeği maddelerinin çeşitli yöntemlere göre değişen madde fonksiyonlarının incelenmesi [Analysis of Test Anxiety Scale items in terms of differential item functioning by different methods]. *Cumhuriyet International Journal of Education*, *9*(4), 1225-1242. https://doi.org/10.30703/cije.703337

Valenzuela, J., Nieto, A.M., & Saiz, C. (2011). Critical thinking motivational scale: A contribution to the study of relationship between critical thinking and motivation. *Journal of Research in Educational Psychology*, *9*(2), 823-848. http://repositorio.ual.es/bitstream/handle/10835/819/Art_24_588.pdf?sequence=1

Yıldırım, H., & Büyüköztürk, Ş. (2018). Using the Delphi Technique and focus-group interviews to determine item bias on the Mathematics Section of the Level Determination Exam for 2012. *Educational Sciences: Theory & Practice*, *18*(2), 447-470. http://doi.org/10.12738/estp.2018.2.0317

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 3, 454–481

https://doi.org/10.21449/ijate.1272517

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

The effect of student characteristics and socioeconomic status on mathematics achievement in Türkiye: Insights from TIMSS 2011-2019

Burçin Coşkun [1,*], **Kübra Karakaya Özyer** [2]

[1]Trakya University, Faculty of Education, Department of Educational Sciences, Edirne, Türkiye
[2]Eskişehir Osmangazi University, Faculty of Education, Department of Educational Sciences, Eskişehir, Türkiye

ARTICLE HISTORY

Received: Mar. 28, 2023
Revised: July 28, 2023
Accepted: Aug. 18, 2023

Keywords:
TIMSS,
Türkiye,
Mathematics Achievement,
8th-grade Students,
Multilevel Modelling.

Abstract: This study examines the factors affecting the mathematics achievement of 8th-grade students in Türkiye using data from the TIMSS in 2011, 2015, and 2019. The data were analysed with multilevel (two-level) modelling. The first level was the student, and the second level was the school. At the student level, such affective characteristics as self-confidence in learning mathematics, liking to learn mathematics, and value given to learning mathematics, as well as educational resources, namely at home, gender, and the frequency of speaking the language of the test at home, were taken into consideration. At the school level, the school's socioeconomic status was included in the model. The results showed that self-confidence in learning mathematics is the most important variable affecting students' mathematics achievement in all years. Besides, the school's socioeconomic status has the strongest effect on students' mathematics achievement, which has increased over the years. The study also showed that those students who performed higher achievement in TIMSS 2011, 2015, and 2019 are confident in learning mathematics, have many educational resources at home, frequently speak Turkish at home, and are from affluent schools. On the other hand, for TIMSS 2011 and 2019, female students were more successful than male students. The effect of liking to learn mathematics on achievement was negative and significant only for TIMSS 2015, while the effect of value given to learning mathematics was positive and significant only for TIMSS 2019. However, the effect size values of the variables showed that this effect was not significant in practice.

1. INTRODUCTION

The 21st century is a period in which technological advancements are rapidly developing. Due to changing job market conditions, new requirements are also emerging accordingly. The Program and Instruction for 21st Century Skills (P21) framework, which was first introduced in the United States but later accepted worldwide, emphasizes the importance of students having various knowledge, skills, and experiences to succeed in their careers and daily lives (Guo & Woulfin, 2016); within the P21 framework, critical thinking and problem-solving skills are particularly emphasized (Akgündüz et al., 2015). Türkiye has also been affected by these changes in the world, and efforts have been made to outline a new profile for Turkish students (EARGED, 2011) as well. Since mathematics lessons are quite effective in teaching critical

*CONTACT: Burçin Coşkun ✉ burcinermegan@gmail.com Trakya University, Faculty of Education, Department of Educational Sciences, Edirne, Türkiye

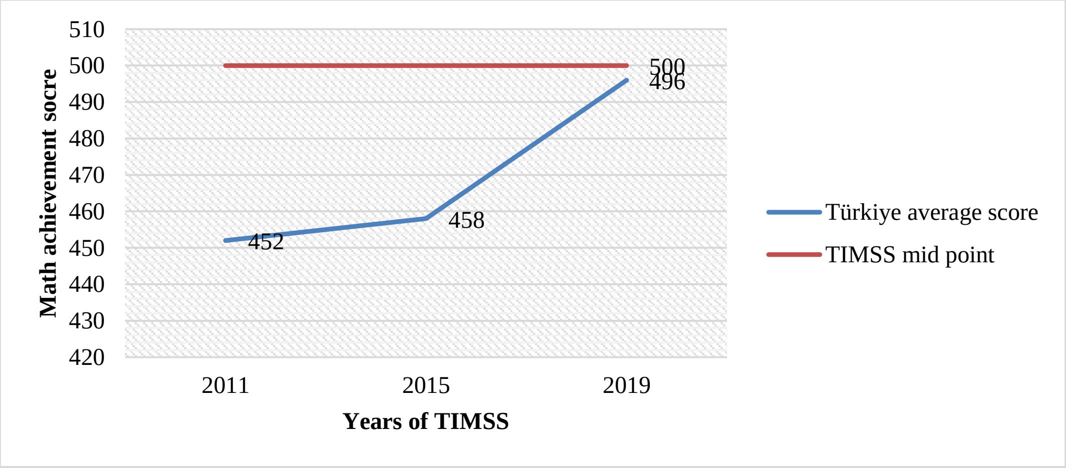
e-ISSN: 2148-7456 /© **IJATE** 2023

thinking and problem-solving skills (Mullis & Martin, 2017), Türkiye has taken steps to teach skills such as critical thinking, creative thinking, decision-making, and the use of technology in its mathematics curricula since 2005 (Dönmez & Dede, 2020). It is also important to measure students' mathematical skills and assess them internationally within certain standards. Examples of large-scale international studies in which the mathematical skills of students from many countries can be compared include PISA and the Trends in International Mathematics and Science Study (TIMSS).

TIMSS is the most comprehensive assessment that determines mathematics and science achievement trends of students (National Centre for Education Statistics, n.d.). TIMSS, which was first conducted in 1995, is repeated every four years and draws attention to the changes in the achievements of countries over time. In TIMSS, not only the participants' basic mathematical skills but also their problem-solving and reasoning skills are evaluated (Lee & Chen, 2019). These skills are evaluated in two parts, namely the learning domain and the cognitive domain (Yıldırım et al., 2013). Aside from mathematical skills, TIMSS also collects information about students' affective skills, home lives, and school lives. In particular, with student, school, and teacher questionnaires for 8th grade students, TIMSS endeavours to present all aspects of the educational environment in detail in which the student lives.

Türkiye did not participate in TIMSS in 1995 and 2003, while it participated only with eighth-grade students in 1999 and 2007. In 2011, 2015, and 2019, Türkiye participated with students at both levels (fourth and eighth grades) (Büyüköztürk et al., 2014; Polat et al., 2016; Suna et al., 2020). When looking at the average scores and country rankings for the last three TIMSS studies, there appears to be a quantitative increase (see Figure 1). However, when compared with 500 points, which is the midpoint of the assessment, the mathematics average scores for eighth-grade students in Türkiye remain at the middle level or below the average.

Figure 1. *TIMSS 2011, 2015, and 2019 mathematics average scores for eighth-grade students in Türkiye.*



According to its ranking among other participating countries, Türkiye lagged behind 57% of these countries in 2011, 62% in 2015, and 51% in 2019 (Büyüköztürk et al., 2014; Polat et al., 2016; Suna et al., 2020). In addition to examining the mathematics achievement of students in Türkiye from a global perspective, country-based research is also important to focus on the underlying causes of students' success or failure. After each TIMSS report was published, various scientific studies were conducted by taking into account the published data and trying to determine the variables affecting the mathematics achievement of students in Türkiye (e.g., about student characteristics: Abalı-Öztürk & Şahin, 2015; Çalışkan, 2014; Çavdar, 2015; Doğan & Barış, 2010; Sarı et al., 2017; about school characteristics: Akyüz, 2014; Aydın, 2015;

Coşkun & Karadağ, 2023; and about teacher characteristics: Yalcin et al., 2017). It is possible to group these variables in the context of students, teachers, and schools (Thomson et al., 2003). Among the variables studied in the student context are variables that reflect students' affective characteristics (attitudes toward mathematics, motivation levels to learn the lesson, and levels of self-confidence in learning the lesson), socioeconomic status, gender, the frequency of speaking the language of the test at home, and home educational resources.

When the education system in Türkiye is analysed, it is seen that students' acquisition of cognitive skills is much more important than their affective characteristics (Öztekin-Bayır & Tekel, 2021). However, studies showing the relationship between affective characteristics and cognitive skills emphasize the dangers of ignoring affective characteristics in determining students' academic achievement (Ferla et al., 2009; Khine et al., 2015; Leder & Forgasz, 2006; Ölçüoğlu & Çetin, 2016; Pajares & Miller, 1997; Wilkins & Ma, 2003; Wilkins, 2004). When these studies were examined, it was determined that the variables of self-confidence in learning, attitude towards the lesson (liking to learn the lesson), and the value given to the lesson significantly affected academic achievement (Arıkan, 2016; Barış, 2009; Coşkun & Karadağ, 2023; Doğan & Barış, 2010; Kaya, 2008; Louis & Mistele, 2012; Sarı et al., 2017; Yatağan, 2014). When the percentages of affective domain characteristics explaining academic achievement were examined, it was revealed that various studies found different results and that these percentages ranged between 12% and 20% (Chowa et al., 2015). While examining the effect of student characteristics on achievement, it is necessary to control the effect of some variables, especially socioeconomic status, on achievement. Socioeconomic status, gender, and ethnicity (frequency of speaking the language of the test at home) are the variables that explain most of the variance in student achievement (Coleman et al., 1966). In this study, the socioeconomic status of the student and the school were controlled to examine the effect of the variables expressing students' affective characteristics and home backgrounds on TIMSS 2011, 2015, and 2019 mathematics achievement. As a result, all students and schools had the same socioeconomic background (Martin et al., 2013).

1.1. Affective Domain

1.1.1. *Students' confidence in learning mathematics*

The first of the affective characteristics examined in TIMSS is self-confidence in mathematics, which can be defined as the student's belief in oneself in the mathematics class and seeing oneself as successful in the processing of this class (Demir & Kılıç, 2010). A student's self-confidence in learning the lesson means that he/she does not give up in any negative situations and feels sufficient motivation to correct that situation (Bandura et al., 2001). Both national (Abazaoğlu et al., 2015; Akyüz, 2014; Coşkun & Karadağ, 2023; Demir & Kılıç, 2010; Ertürk & Erdinç-Akan, 2018; Ölmez, 2020; Yalçın et al., 2017) and international (Arıkan et al., 2016; Chen, 2014; Ker, 2016; Papanastasiou, 2000; Wilson & Narayan, 2016; Yoshino, 2012) studies have revealed that students' feeling of self-efficacy while learning the lesson is related to their acquisition of the target behaviours of the lesson.

1.1.2. *Students liking to learn mathematics*

The attitudes and emotional states of students play an important role in the process of learning mathematics. The enjoyment of learning mathematics is directly relevant to their intrinsic motivation (Mullis et al., 2012). Enjoying learning mathematics, liking to do mathematics-related homework, and eagerly anticipating a math class all provide clues about students' intrinsic motivation (Hansford & Hattie, 1982). Some studies suggest that the latent variable of 'liking to learn mathematics' derived from the "Students Like Learning Mathematics Scale" (Mullis et al., 2020, p. 428) in TIMMS assessment is a variable that affects academic achievement (Belbase, 2013; Erşan, 2016; Khine et al., 2015; Liou, 2014; Tavşancıl & Yalçın, 2015; Yıldırım et al., 2013). According to these studies, individuals with high intrinsic motivation also tend to have high levels of mathematical achievement. However, some studies that compared different countries obtained different results; for example, Akyüz (2014) examined the effect of affective characteristics on mathematics achievement by analysing the TIMSS 2011 data from Singapore, Finland, the USA, and Türkiye's 8th-grade students: the study findings revealed that, ‘liking to learn mathematics’ was a significant variable for achievement in Singapore and the United States, but not in Türkiye. Coşkun and Karadağ (2023) found a negative relationship between students’ liking to learn mathematics and students' mathematics achievement. Similarly, Kara (2023), using the TIMSS 2019 data from Türkiye, found that as students' enjoyment of learning mathematics decreased, their mathematics achievement increased as well. Such contradictory results indicate that more research is required to understand the impact of students’ liking to learn mathematics. Placing greater emphasis on students' emotional states during the process of learning mathematics and increasing research in this area may therefore help to make the process of learning mathematics more effective.

1.1.3. *Students’ value given to learning mathematics*

Another affective characteristic associated with students' mathematics achievement is valuing mathematics learning. Value given to learning mathematics refers to students' belief that mathematics is important and will be useful in their future lives (Wigfield & Eccles, 2000). However, in the literature, valuing the lesson is also referred to as external motivation (Ryan & Deci, 2000). In other words, it is an extrinsic motivation source when a student thinks that the mathematics course is important and believes that it will be useful both in daily life and in work life (Wigfield & Eccles, 2000). Some studies have found that there is no relationship between the value given to learning mathematics and academic achievement (Arıkan et al., 2016; Yavuz et al., 2017), while other studies have shown that students who value mathematics use their cognitive skills more consciously and perform better in mathematics exams (Ker, 2016; Kim et al., 2013; Phan et al., 2010). Therefore, students' levels of value for learning mathematics can be an important variable for their mathematics achievement. Understanding the importance of mathematics class and being motivated in this regard can help students achieve higher academic success.

1.2. Gender

Various studies show that the most frequently studied variable when investigating factors affecting academic achievement is gender (Aydın, 2015; Karaca, 2018; Louis & Mistele, 2012). National and international studies have compared the achievements of male and female students and investigated the reasons for the differences (Aksu et al., 2017; Mullis et al., 2016). The effectiveness of the gender factor was also investigated in the studies on TIMSS mathematics achievement, and it was determined that there was a significant difference between male and female students (Aydın, 2015; Louis & Mistele, 2012; Martin et al., 2000). Several studies on gender differences have shown that male students have higher mathematics performance than that of female students (Işlak, 2020; Kılıç & Askin, 2013; Louis & Mistele, 2012; Mau & Lynn, 2010; Martin et al., 2000). Studies conducted in Türkiye also support similar results; for example, Kılıç and Askin (2013), based on TIMSS 2011 data, reported that male students have higher mathematics performance than female students have.

However, there are studies showing that female students outperform male students in TIMSS mathematics achievement (Aydın, 2015), while other studies demonstrate that gender is not an important variable in predicting students' mathematics achievement (Coşkun & Karadağ, 2023; Kaleli-Yılmaz & Hanci, 2015; Karaca, 2018; Lee & Kung, 2018; Mohammadpour & Abdul Ghafar, 2014; Sarouphim & Chartouny, 2017). Therefore, most studies suggest that the gender effect on how well students perform in mathematics depends on the TIMSS study year. As a result, it is seen that gender plays an important role in the process of investigating factors affecting academic achievement. However, further research is needed to arrive at a clear conclusion regarding the effect of gender on mathematics achievement.

1.3. Language of Test Spoken at Home

Another important factor that affects students' success is the language spoken at home. In the TIMSS study, similarity between the language of the test and the language used at home had a positive effect on students' achievement. For this reason, it is often questioned how often students speak the test language at home. Students who do not speak Turkish, the language of the test, at home are generally children of minority or immigrant families, which creates difficulties for students to understand and answer the test. The degree to which the language used at home and the language of the exam are similar is crucial to the learning process. The difference between the language spoken at home and the language on the test for children of minority or immigrant families creates some difficulties in the learning process (Lee, 2020). Since the families of these students are usually economically weak, the budget they allocate to their children may also be limited, which may prevent them from accessing new learning opportunities (Coleman, 1994; Portes & MacLeod, 1996). Some studies have shown a positive relationship between the frequency of speaking the language of the test at home and mathematics achievement (Chen, 2014; Ismail & Awang, 2008; Mohammadpour, 2013; Sevgi, 2009). However, in Sandoval-Hernández and Białowolski's (2016) study comparing five countries, a positive relationship was found between the frequency of speaking the test language at home and mathematics achievement in Taipei and Singapore, while a negative relationship was found in Hong Kong. Furthermore, for South Korea and Japan, there was no effect of the frequency of speaking the test language at home on mathematics achievement. In conclusion, the similarity between the language spoken at home and the language of the test is important for students' success. The fact that children of minority or immigrant families are less familiar with the language of the test can be an obstacle to the learning process. Therefore, education systems should create a favourable environment for students' success by considering language differences.

1.4. Socioeconomic Status

In a large-scale study conducted in 1966, Coleman et al. (1966) found that the most important factor affecting students' achievement was their socioeconomic status; the effect of schools on achievement was very small, though. Their study has been the basis for many subsequent studies. In TIMSS, socioeconomic status at the student level is determined by the variables of educational resources at home. The educational resources at home are determined by asking about the number of books in the student's home, whether the student has a computer, a room, and an internet connection, and the educational level of the parents. High scores of the student's answers to these questions indicate that he/she has access to a large number of educational resources at home (Sarı et al., 2017). The effect of this variable on achievement can be analysed at both the student and school levels. Since parents with high socioeconomic status can offer more educational opportunities to their children, students' achievement is expected to be higher (Broer et al., 2019; Mullis et al., 2020; Olatunde, 2010; Şirin, 2005). In another study, Chmielewski (2018) found a positive and significant relationship between socioeconomic status and student achievement in his study comparing 30 countries using data from international studies conducted over 50 years.

It is also possible to come across studies that found that the strongest variable affecting the mathematics achievement of 8th-grade students in TIMSS is the socioeconomic status of the student (Bos & Kuiper, 1999; Erşan, 2016; Kılıç & Askın, 2013; Mohammadpour & Abdul

Ghafar, 2012). Türkiye is a heterogeneous country in terms of socioeconomic status and as in the TIMSS Türkiye sample, there are students with very high and very low socioeconomic status (Büyüköztürk et al., 2014; Polat et al., 2016; Suna et al., 2020).

Focusing on the studies conducted in Türkiye, similar results were found to be valid for students in Türkiye and socioeconomic status was found to be effective in explaining student achievement (Akyüz, 2014; Acar-Güvendir, 2014; Bellibaş, 2016; Erdoğan &Acar-Güvendir, 2019; Gelbal, 2008; Kalaycıoğlu, 2015; Karaağaç, Cingöz & Gür, 2020; Özdemir, 2016; Özkan & Acar-Güvendir, 2014; Suna et al., 2020; Tomul & Savaşçı, 2012; Yetkiner-Özel et al., 2013).

However, in schools where the socioeconomic status of the school is high, students have access to more resources and opportunities. This can increase students' academic achievement and help them prepare for a better future. In addition, many studies using TIMSS data have included the school’s socioeconomic status in their research. These studies examined how the socioeconomic status of the school affected students' achievement in subjects such as mathematics and science. As a result, the effect of students' socioeconomic status on their achievement should be taken into consideration, and this factor should also be considered in the development of educational policies. Ersan and Rodrigez (2020), in their study with TIMSS 2015 Türkiye data, found that the effect of socioeconomic status is still strong at the school level when the effect of socioeconomic status within and between schools is separated.

1.5. Importance of Research

Although TIMSS results have caused various debates since 1995, they have pioneered educational reforms all over the world. The TIMSS study helped to determine the current educational situation and achievement trends by providing the opportunity to compare the mathematics and science achievement levels of Turkish students with those of students in other countries. However, systematic studies are needed to analyse these data well and transform them into educational policy. Since TIMSS data are obtained at different levels (student, school, and teacher levels), multilevel analysis methods should be preferred to minimize the error in the analysis of such data (Bryk & Raudenbush, 1992). In this study, student characteristics that are thought to affect Turkish students' mathematics achievement in TIMSS 2011, 2015, and 2019 were investigated with two-level modelling by controlling socioeconomic status at the student and school levels. When we look at the studies using Turkish TIMSS data in the literature on TIMSS data, it is generally seen that data from a single time are analysed (Akyüz, 2014; Tavşancıl & Yalçın, 2015). However, comparing the data of a country at different times may be more meaningful in showing the effect of the educational policies enacted by the country. In TIMSS 1999, unlike other assessments, students' views on liking and valuing mathematics were questioned through a single scale (the positive attitude toward mathematics scale). In the TIMSS 2007 assessment, student's attitudes towards the lesson were investigated on the scales of positive attitude towards the lesson, value given to learning the lesson, and confidence in learning the lesson. However, since Türkiye participated in this assessment only with 8th-grade students, there is no information on the index variable of educational resources at home. For this reason, the current study aims to draw attention to the changes in factors affecting the mathematics achievement of 8th-grade students in Türkiye over time by analysing TIMSS 2011, 2015, and 2019 data rather than data from a single time period. In this study, therefore, those factors affecting the mathematics achievement of 8th-grade students in TIMSS 2011, 2015, and 2019 are examined. The decision to compare TIMSS data in three different periods was taken because the scales used in these assessments are similar and the 8th grade is defined as the last grade of secondary school in the Turkish education system. This grade level is very important for high school selection and is related to high school success of students in the following years. In Türkiye, 8th-grade students are required to take the high school transition exam at the end of the year and with the score they get from this exam, they have the opportunity to be placed in various high schools. The high school entrance exam is the first large-scale and high-stakes exam for Turkish students and is taken very seriously by students as it directly affects their further education. For these reasons, the success levels of Turkish students in the 8th grade are related to their high school success in the following years (Özdemir & Gelbal, 2016).

In the current study, gender, the frequency of speaking the language of the test at home, and the educational resources at home were also studied while determining the affective characteristics that influenced students' mathematics achievement. Since previous studies have shown that most of the variation in student achievement is due to socioeconomic status, race, and gender (Chmielewski, 2018; Coleman et al., 1966; Hilton & Lee, 1988; Mullis et al., 2020), such variables were also included in the research model.

1.6. Research Objective

In this study, the effects of affective characteristics, gender, frequency of speaking the language of the test at home, educational resources at home, and the school's socioeconomic status on the eighth-grade students' mathematics achievement in TIMSS 2011, 2015, and 2019 were investigated. In this context, answers to the stated research problems were sought:

1. Does the mathematics achievement of eighth-grade students in TIMSS 2011, 2015, and 2019 vary between schools?
2. Which student variables (self-confidence in learning mathematics, liking to learn mathematics, value given to learning mathematics, gender, language of test spoken at home, and educational resources at home) have an effect on eighth-grade students' mathematics achievement in TIMSS 2011, 2015, and 2019, when students' socioeconomic status is controlled?

2. METHOD

2.1. Population and Sample

This study includes an analysis of Türkiye's TIMSS 2011 and 2015 and 2019 eighth-grade data. TIMSS uses a two-level random sampling design in which a school sample is first selected, and then all students in at least one classroom from these schools are sampled (LaRoche et al., 2020). Türkiye's TIMSS 2011, 2015, and 2019 eighth-grade population and sample sizes are given in Table 1. In this study, the missing data were removed from the datasets by the listwise elimination technique, since the missing data rates in the datasets did not exceed 5% (Garson, 2019).

Table 1. *Türkiye's TIMSS 2011, 2015, and 2019 population and sample sizes.*

	Population		Sample		Sample size after listwise elimination	
Years	School	Student	School	Student	School	Student
2011	17.621	1.198.697	239	6928	239	6850
2015	15.583	1.201.185	218	6079	218	5966
2019	16.179	1.158.547	181	4077	181	3930

One of the advantages of working with TIMSS data is that it provides weighting data at the student and school levels. Weighting is important to compensate for the negative effects of situations such as an unequal probability of being selected for sampling or not responding to questions (Von Secker & Lissitz, 1999). These weights are the inverse of the student's probability of being selected for the sample (LaRoche et al., 2016). In the study, student variables were weighted using Total Student Weight (TOTWGT), while no weighting was used for the school variables. In the analyses, the student-level variables were centered on the group mean, and the school-level variable was centered on the overall mean.

2.2. Data Collection Tools

The data collection tools of the current study consist of TIMSS 2011, 2015, and 2019 eighth-grade mathematics achievement tests and student questionnaires. TIMSS uses item response theory to describe student achievement on a scale representative of the entire assessment framework and to provide accurate measures of student proficiency distributions and trends (Foy & Yin, 2016). In addition, TIMSS calculates plausible values representing mathematics and science proficiency levels for all students to provide unbiased estimates of the relationship between student achievement and contextual variables (Foy et al., 2020). In the study, five plausible values calculated from the scores of students on mathematics achievement tests were used as dependent variables.

2.2.1. *Mathematics achievement test*

This section should indicate the study's design, the sampling, the data collection tools, and the data analysis. Clarification is essential in this part. This section should indicate the study's design, the sampling, the data collection tools, and the data analysis. Clarification is essential in this part.

The TIMSS mathematics tests are made based on comprehensive assessment frameworks that were made with the cooperation of participating countries. Each of the eighth-grade mathematics assessments is organized around two dimensions, namely the content dimension, which indicates the subject or content areas to be evaluated, and the cognitive dimension, which expresses the thinking processes that students can use while engaging with the content (Mullis et al., 2012). The TIMSS 2011, 2015, and 2019 eighth-grade mathematics tests are divided into four content areas: numbers (30%), algebra (30%), geometry (20%), and data and probability (20%). From the three cognitive domains (knowing, applying, and reasoning), TIMSS 2011 placed less emphasis on knowing and more on reasoning (Mullis et al., 2012). In TIMSS 2015, more emphasis was placed on knowing and applying it to the questions and less on reasoning (Gronmo et al., 2013). Most of the TIMSS 2019 mathematics items measure students' practice and reasoning skills (Mullis et al., 2020).

2.2.2. *Student questionnaire*

In TIMSS, the questionnaires are administered to students, teachers, and school administrators to learn more about their home, school, and classroom environments. The Student Questionnaire, administered to eighth-grade students, asks students about their home environment, availability of educational resources, and educational experiences related to learning mathematics and science at home and school and includes various scales about attitudes toward learning mathematics and science (Mullis & Fishbein, 2020). In this study, latent or observed variables from questionnaires of students were used as independent variables. Information on the independent variables is provided in the section that follows. Descriptive statistics for these variables are presented in Appendix.

The gender variable consists of two categories (Girl = 1 and Boy = 2); in this study, the gender variable is recoded as Girl = 0, and Male = 1.

The language of test spoken at home expresses how often students use the language of the TIMSS at home (1 = always, 2 = almost always, 3 = sometimes, and 4 = never). In this study, variable levels are reverse-coded.

The "educational resources at home" is a continuous variable created based on the level of agreement of the students with three statements, namely the number of books at home; the education level of the parents; and whether they have a room and/or internet connection of their own. The Cronbach Alpha reliability coefficient of the scale was calculated as 0.63 for TIMSS 2011, 0.62 for TIMSS 2015, and 0.64 for TIMSS 2019.

The 'liking to learn mathematics' is a continuous variable created according to the level of students' agreement with such statements as: I enjoy learning mathematics; I wish I did not have to study mathematics, and mathematics lessons are boring, etc. While the scale consists of five items for TIMSS 2011, it consists of nine items for TIMSS 2015 and 2019.

The 'value given to learning mathematics' is a continuous variable created based on the student's level of agreement with such statements as: I believe that learning mathematics will benefit me in my daily life; I need to be good at mathematics to attend the university of my choice; and to get the job I want, I need to be good at mathematics, etc. While the scale consists of six items for TIMSS 2011, it consists of eight items for TIMSS 2015 and TIMSS 2019. The Cronbach Alpha reliability coefficient of the scale was calculated as 0.75 for TIMSS 2011, 0.87 for TIMSS 2015, and 0.88 for TIMSS 2019.

The 'self-confidence in learning mathematics' is a continuous variable created based on the level of agreement of students with nine statements, such as: I am good at mathematics; I learn mathematics quickly; and my teacher says I am good at mathematics, etc. The Cronbach Alpha reliability coefficient of the scale was calculated as 0.87 for TIMSS 2011 and 2015 and 0.89 for TIMSS 2019.

The 'school's socioeconomic status' is a continuous variable that is calculated by taking the average of the *Home Educational Resources Scale* student scores for the entire school.

2.3. Data Analysis

The TIMSS data is organized hierarchically. According to TIMSS data, students are clustered in classes, classes are clustered in schools, and schools are clustered in nations. Hierarchical data cannot be analysed at a single level because clustering implies that individuals in one group will be increasingly similar to persons in other groups. Treating individuals as if they are separate from their social group leads to bias in analyses (Heck & Thomas, 2015). A key assumption of the linear regression model, the independence of the residuals, is relaxed by the multilevel modelling, which is an extension of that model (Snijders & Bosker, 2012). In the study, the analyses were carried out with the HLM 8 package software using the multilevel modelling method.

When the multilevel modelling assumptions were examined with the model created, it was concluded that the first-level and second-level errors were normally distributed, and there was no multicollinearity between the first-level variables. However, it was observed that the homogeneity of variance assumption at the first level analysed with the H test could not be achieved. In general, even though the violation of the homogeneity assumption does not significantly affect the estimation of the coefficient and standard errors, it is nevertheless advised to adopt the robust sandwich approach developed by White (1980) for parameter estimation in these circumstances (Raudenbush & Bryk, 2002). For this reason, robust estimation values produced by HLM using the sandwich estimation method were taken into account in the study (Raudenbush et al., 2011).

Because all students in a classroom are included in the sample in TIMSS Türkiye, classrooms represent schools. Thus, in the study, two-level modelling was done for a first-level student and a second-level school. In the study, firstly, the effect of school on TIMSS 2011, 2015, and 2019 in Türkiye's eighth-grade mathematics achievement was investigated. For this purpose, random effects of one-way ANOVA models (unconditional models) were created and analysed. Equation (1) is an expression of the unconditional model.

$$Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}. \tag{1}$$

Here, n_j stands for the number of students in the *j*th school when the total number of schools is *N*. Y_{ij} is the mathematics achievement score of the *i*th student in the *j*th school. u_{0j} is the school-level error, and r_{ij} is the student-level error (a random error related to the mathematics achievement score of the *i*th student in the *j*th school). The model divides the total variance into two independent components as shown in Equation (2): the first-level error variance $\hat{\sigma}^2$ and the second-level error variance $\hat{\tau}_{00}$ (Hox et al., 2018).

$$var(Y_{ij}) = \hat{\tau}_{00} + \hat{\sigma}^2 \tag{2}$$

Thus, the ratio of the second level variance to the total variance is calculated by the Intraclass Correlation Coefficient (ICC) Equation (3).

$$ICC = \frac{\hat{\tau}_{00}}{\hat{\tau}_{00}+\hat{\sigma}^2}. \tag{3}$$

In the study, random intercept regression models with more than one variable were created and analysed in order to investigate the effect of student characteristics on TIMSS 2011, 2015, and 2019 mathematics achievement by controlling socioeconomic status at the student and school levels (Snijder & Bosker, 2012). This model, in which only the slope coefficient of the constant term changes randomly between schools and the independent variables take place as fixed effects at the first and second levels, is expressed by Equation (4).

$$Y_{ij} = \gamma_{00} + \gamma_{p0}X_{pij} + \gamma_{0q}W_{qj} + u_{0j} + r_{ij}. \tag{4}$$

Here, X_{pij} represents p independent variables at the student level, and W_{qj} represents q independent variables at the school level (Hox et al., 2018).

After determining whether the variables had an effect on student achievement in the study, the effect sizes of the significant variables were investigated. Depending on the purpose of the study, the effect size may be the difference between the means, correlation, a standardized regression coefficient, odds ratio, explained variance ratio, etc. Since the total variance in multilevel modelling consists of two components, within-group ($\hat{\sigma}^2$) and between-groups ($\hat{\tau}_{00}$), the effect size can be calculated in three different ways in these models (Lou et al., 2021). In this study, effect sizes were calculated for the student level by dividing the estimated regression coefficients of the variables (γ_{p0}) by the student level standard deviation ($\hat{\sigma}$) of the unconditional model. The effect sizes for the school level were calculated by dividing the estimated regression coefficient of the variable (γ_{0q}) by the school level standard deviation of the unconditional model ($\sqrt{\hat{\tau}_{00}}$). The study used the effect size value ranges put forth by Rosenthal and Rosnow (1984). According to these value ranges, the effect size is considered large if it is higher than 0.5 standard deviation, medium if it is between 0.3 and 0.5 standard deviation, small if it is between 0.1 and 0.3 standard deviation, and practically insignificant if it is less than 0.1 standard deviation.

In addition, the variance rates (R^2) explained by the models created in the study were calculated by Equation (5) for student level and Equation (6) for school level. In these calculations, unconditional models were taken as reference (Raudenbush & Bryk, 2002).

$$R_1^2 = \frac{\sigma^2_{r_{ij}(unconditional\ model)} - \sigma^2_{r_{ij}(compared\ model)}}{\sigma^2_{r_{ij}(unconditional\ model)}} \tag{5}$$

$$R_2^2 = \frac{\sigma^2_{u_{0j}}(unconditional\ model) - \sigma^2_{u_{0j}}(compared\ model)}{\sigma^2_{u_{0j}}(unconditional\ model)}$$

3. RESULTS

3.1. School Effect on Mathematics Achievement in TIMSS 2011, 2015, and 2019

The analysis results of the unconditional models created to investigate whether there is a difference between schools in terms of the mathematics achievement of students in TIMSS 2011, 2015, and 2019 are given in Table 2. According to the results, the general mathematics achievement averages of the eighth-grade students increased over the years. Nonetheless, Türkiye performed below the TIMSS mean of 500 points in each of the three assessments.

Table 2. *Analysis results for unconditional models.*

Fixed effect		Coefficient	se	*t*	*df*	*p*
Average, γ_{00}	2011	450.64	4.25	106.06	238	0.00
	2015	455.51	4.58	99.53	217	0.00
	2019	490.88	5.18	94.75	180	0.00
Random effect		*sd*	Variance	χ^2	*df*	*p*
School level, u_0	2011	62.13	3859.77	3637.17	238	0.00
	2015	61.75	3812.79	3426.93	217	0.00
	2019	64.79	4197.50	2407.38	180	0.00
Student level, r	2011	92.04	8471.37			
	2015	84.65	7164.96			
	2019	87.64	7679.91			

The 95% confidence intervals for the general mathematics achievement averages of the assessments were calculated with the equation $\gamma_{00} \pm 1.96(\hat{\tau}_{00})^{1/2}$. According to the results, the TIMSS 2011 mathematics achievement scores of the students are between 328.87 and 573.41, the TIMSS 2015 mathematics achievement scores are between 334.48 and 576.54, and the TIMSS 2019 mathematics achievement scores are between 363.89 and 617.87 points.

According to the random effect estimates in Table 2, the differences in mathematics achievement between schools (TIMSS 2011: $\chi^2 = 3637.17; p < 0.05$, TIMSS 2015: $\chi^2 = 3426.93; p < 0.05$, TIMSS 2019: $\chi^2 = 2407.38\ ; p < 0.05$) are significant for all assessments. When the ICC value was computed using Equation (3), it was determined that the differences in achievement between schools accounted for 31% of the variance in students' mathematics achievement for TIMSS 2011 and 35% for TIMSS 2015 and 2019. According to the results, the school effect on students' TIMSS 2011, 2015, and 2019 mathematics achievement is sufficient to be examined with multilevel modelling (Musca et al., 2011). In addition, 69% of the variability in student achievement for TIMSS 2011 is due to differences between students (explained by student variables), while for TIMSS 2015 and 2019, 65% of this variability is due to differences between students (explained by student variables).

3.2. The Effect of Student and School Variables on TIMSS 2011, 2015, and 2019 Mathematics Achievement

According to the results in Table 3, the school's socioeconomic status has the strongest effect (TIMSS 2011: γ=34.98; *p*<0.01, TIMSS 2015: γ=41.51; *p*<0.01, TIMSS 2019: γ=43.92; *p*<0.01) on students' mathematics achievement for all TIMSS assessments. A one-unit increase in the school's socioeconomic status leads to an increase of approximately 35 points for TIMSS 2011, 42 points for 2015, and 44 points for 2019 in students' mathematics achievement. Over the years, the estimated coefficient value of the variable also increases. When the effect size values in Table 3 are examined, it is seen that the school's socioeconomic status has a large effect on students' mathematics achievement (Rosenthal & Rosnow, 1984). A one standard deviation increase in the variable is expected to have an effect of 0.56 standard deviation for TIMSS 2011, 0.67 for TIMSS 2015 and 0.68 standard deviation for TIMSS 2019 on students' mathematics achievement.

Table 3. *Random intercept regression models with multiple variables.*

Fixed effect		Coefficient	se	*t*	*df*	*p*	Effect size
Average, γ_{00}	2011	450.34	2.94	153.34	237	0.000	---
	2015	455.26	2.98	152.55	75	0.000	---
	2019	490.31	3.29	148.83	179	0.000	---
Level 2							
School's socioeconomic status, γ_{01}	2011	34.98	2.70	12.94	237	0.000*	0.56
	2015	41.51	2.63	15.80	216	0.000*	0.67
	2019	43.92	3.06	14.36	179	0.000*	0.68
Level 1							
Gender, γ_{10}	2011	-6.81	2.43	-2.80	122	0.006*	-0.07
	2015	-5.38	2.91	-1.85	19	0.080	---
	2019	-8.20	3.12	-2.63	84	0.010*	-0.09
Language of test spoken at home, γ_{20}	2011	10.35	2.69	3.85	65	0.000*	0.11
	2015	8.66	2.46	3.52	145	0.001*	0.10
	2019	7.67	3.05	2.51	98	0.014*	0.09
Educational resources at home, γ_{30}	2011	8.62	0.78	11.03	251	0.000*	0.09
	2015	7.54	0.90	8.36	24	0.000*	0.09
	2019	9.75	1.23	7.95	61	0.000*	0.11
Self-confidence in learning mathematics, γ_{40}	2011	21.44	0.73	29.26	1075	0.000*	0.23
	2015	21.77	0.86	25.40	25	0.000*	0.26
	2019	19.64	1.02	19.21	53	0.000*	0.22
Liking to learn mathematics, γ_{50}	2011	-0.53	0.91	-0.59	220	0.559	---
	2015	-4.15	0.96	-4.33	39	0.000*	-0.05
	2019	-2.02	1.27	-1.59	193	0.113	---
Value given to learning mathematics, γ_{60}	2011	0.71	0.75	0.96	95	0.342	---
	2015	-0.65	0.75	-0.86	140	0.391	---
	2019	1.84	0.91	2.03	63	0.047*	0.02
Random effect		sd	Variance	χ^2	*df*	*p*	
School level, u_0	2011	41.24	1700.72	2174.89	237	0.000	
	2015	35.67	1272.42	1731.21	216	0.000	
	2019	37.47	1403.80	1204.02	179	0.000	
Student level, r	2011	76.95	5921.30				
	2015	70.90	5027.15				
	2019	72.75	5293.10				

According to the coefficient estimations of student-level variables, self-confidence in learning mathematics has the strongest effect (TIMSS 2011: γ=21.44; *p*<0.01, TIMSS 2015: γ=21.77; *p*<0.01, TIMSS 2019: γ=19.64; *p*<0.01) on students' mathematics achievement. A one-unit increase in the self-confidence in learning mathematics leads to an increase of about 22 points on the TIMSS 2011 and 2015 and 20 points on the TIMSS 2019. A one standard deviation increase in self-confidence in learning mathematics is expected to have an effect of 0.23

standard deviation for TIMSS 2011, 0.26 for TIMSS 2015, and 0.22 for TIMSS 2019 on students' mathematics achievement, while these effect sizes are small.

Another affective variable, the effect of liking to learn mathematics on achievement, was found to be negative and significant for TIMSS 2015 (γ=-4.15; p<0.01). A one-unit increase in the variable decreases students' mathematics achievement by approximately four points. For TIMSS 2011 and TIMSS 2019, the effect of liking to learn the lesson on students' mathematics achievement is not significant. A one-standard deviation increase in the liking to learn mathematics is expected to result in a 0.05 standard deviation decrease in student achievement. The effect of value given to learning mathematics on students' achievement was found to be positive and significant (γ=1.84; p<0.05) only for TIMSS 2019. A one-unit increase in value given to learning mathematics leads to an increase of approximately 2 points in students' achievement. It is expected that a one standard deviation increase in the variable will cause an increase of 0.02 standard deviations in the TIMSS 2019 mathematics achievement scores of students. In practice, the effect sizes of both variables were not found to be significant.

The effect of educational resources at home on students' mathematics achievement (TIMSS 2011: γ=8.62; p<0.01, TIMSS 2015: γ=7.54; p<0.01, TIMSS 2019: γ=9.75; p<0.01), which is one of the variables expressing the home background of the students, is positive and significant for all assessments. A one-unit increase in educational resources at home leads to an increase of about 9 points on the TIMSS 2011, 8 points on the TIMSS 2015, and 10 points on the TIMSS 2019. Likewise, a one standard deviation increase in educational resources at home is expected to result in an increase in TIMSS mathematics achievement of 0.09 standard deviations in 2011 and 2015, and 0.11 standard deviations in 2019. It can be said that the effect size of the variable is practically insignificant for TIMSS 2011 and 2015, but small for TIMSS 2019.

The effect of student characteristics and gender on mathematics achievement was found to be negative and significant for TIMSS 2011 (γ = -6.81; p<0.01) and TIMSS 2019 (γ = -8.20; p<0.01). Female students scored approximately 7 points higher for TIMSS 2011 and 8.5 points higher for TIMSS 2019 than male students did. The effect of the variable on students' mathematics achievement is not significant for TIMSS 2015. Based on the effect size values for the gender variable, the average mathematics score of female students in TIMSS 2011 is 0.07 standard deviation higher than the average mathematics score of male students in TIMSS 2011 and 0.09 standard deviation higher than the average mathematics score in TIMSS 2019. These effect size values are not significant in practice.

The language of test spoken at home has a positive and significant effect on how well students perform in mathematics on all assessments. A one-unit increase in speaking the language of the test leads to an increase of about 11 points for TIMSS 2011, 9 points for TIMSS 2015, and 8 points for TIMSS 2019. The effect size of the language of the test was calculated as 0.11 for TIMSS 2011, 0.10 for 2015, and 0.09 for 2019. The effect size of the variable for TIMSS 2019 is not significant in practice, and for other years, the effect size on students' mathematics achievement is small.

Using Equation (6), we can say that the model with more than one independent variable and a constant term that changes randomly explains 56% of the variation at the school level for TIMSS 2011 and 67% of the variation for TIMSS 2015 and 2019. Several school variables that are not included in the model are expected to explain the unexplained amount of variance at the school level. When the explained variance rate at the student level with the model is calculated with Equation (5), it can be said that 30% of the variance at the student level is explained for the TIMSS 2011 and 2015 assessments and 31% for the 2019 one. It is expected that different student-level variables that were not included in the study would explain 70% of the student-level variation for TIMSS 2011 and 2015 and 69% for 2019.

4. DISCUSSION and CONCLUSION

International education studies provide important clues about the quality of education in different countries. Therefore, in attempt to improve the quality of education, Türkiye tries to improve its students' skills by making changes in national education policies based upon TIMSS results. In particular, the preparation of high school transition exam questions with skill-based questions aims to incorporate skills similar to those measured in TIMSS into students' learning.

However, research shows that Türkiye's TIMSS mathematics results are below about half of the results of other countries (Büyüköztürk et al., 2014; Polat et al., 2016; Suna et al., 2020). To solve this problem, it is important to investigate the underlying causes. In this study, student characteristics affecting the mathematics achievement of 8th grade students in Türkiye are analysed with a specific aim to improve the education system by designing educational policies according to the characteristics of students.

One of the most important characteristics associated with students' mathematics achievement is affective characteristics (Akyüz, 2014; Topçu et al., 2016) which are defined as students' self-confidence in learning the lesson, their liking the lesson, and their value given to learning the lesson in TIMSS. In the current study, only self-confidence in learning the lesson was found to be an important variable in students' mathematics achievement in all years. In other words, 8th grade students who are self-confident in mathematics receive higher scores on mathematics tests. When the effect size values were analysed, it was determined that self-confidence in learning mathematics was more effective than other affective variables. It can be said that this effect is small according to the effect size value ranges taken as references in the study (Rosenthal & Rosnow, 1984). A meta-analysis study conducted by Çiftçi and Yıldız (2019) revealed that student self-confidence has a moderate effect on academic achievement. Since Cohen's d effect size value ranges were taken as references in this study, they differed from the results of the current study. When other studies in the literature are examined, it is seen that there is a positive relationship between mathematics achievement and self-confidence in learning mathematics (Akyüz, 2014; Arıkan et al., 2016; Aydın, 2015; Chen, 2014; Coşkun & Karadağ, 2023; Demir & Kılıç, 2010; Kadijević, 2008; Ismail, 2009; Ismail & Awang, 2012; Lee & Chen, 2019; Lee & Stankov, 2018; Wang et al., 2023). Ismail (2009), in a study conducted with TIMSS 2003 data, stated that self-confidence in learning mathematics is the strongest variable explaining students' mathematics achievement. Similarly, Khine et al. (2015) designed a structural equation model explaining the mathematics achievement of students' affective characteristics with TIMSS 2011 data and revealed that the greatest contribution to mathematics achievement was due to self-confidence. This finding was also found in other large-scale studies other than TIMSS. For example, studies conducted with PISA data also found a positive and significant relationship between students' mathematics self-confidence and mathematics domain skills (Okatan & Tomul, 2020; Sarıer, 2021; Usta & Demirtaşlı, 2018). As a result, students with high self-confidence experience less anxiety and hesitate less because they are confident in themselves. Thus, they can benefit more from mathematical learning environments. Especially in view of the finding that the variable of students' self-confidence explains mathematics performance at a significant level in all years, a programme can be developed to make students self-confident in mathematics lessons, and whether this programme increases their mathematics performance can be tested with experimental research. For this reason, designing textbooks and lesson plans from easy to difficult can support students' self-confidence in learning mathematics.

The results of the analysis, based on TIMSS 2015 data, showed that the variable 'liking to learn mathematics' has a negatively significant effect on students' mathematics achievement. However, there is no such relationship for TIMSS 2011 and 2019 data. The results obtained for

TIMSS 2015 reveal that students who like mathematics have lower mathematics achievement. However, the effect size of the variable 'liking to learn mathematics' is not practically significant. Unlike the current study, Kara (2023) and Coşkun (2022) found a negative relationship between liking mathematics and students' mathematics achievement in studies conducted with TIMSS 2019 data. The fact that different student variables were also used in these studies may have caused this effect for TIMSS 2019. The results obtained in our study differ from some other studies in the literature. Previous studies have found a positive relationship between enjoyment of learning mathematics and mathematics achievement (Mohammadpour, 2012; Tavşancıl & Yalçın, 2015). Therefore, it is expected that mathematics achievement of secondary school students will increase with the increase in their level of liking mathematics. When the studies conducted on the TIMSS Türkiye sample were analysed, it was determined that some of them used a single-level correlational analysis. The use of different analysis models may therefore cause differences in the results. In this study, a multilevel analysis method was used, and school level variability was taken under control. In this study, the relationship between value given to learning mathematics learning as the last affective variable and students' mathematics achievement was examined. Valuing learning mathematics can be defined as students' belief that what they learn in mathematics lessons will benefit them in the future (Wigfield & Eccles, 2000). Considering the results of the current study, it was determined that the variable of value given to learning mathematics learning significantly explained students' mathematics achievement only in the 2019 data. Similarly, Yavuz et al. (2017), in their study comparing TIMSS 2007 and 2011 results, showed that there was no significant relationship between the value students placed on mathematics and mathematics achievement. In addition, Arıkan et al. (2016) analysed TIMSS 2007 and 2011 data not only within the scope of Türkiye but also tried to reveal the factors affecting the mathematics achievement of both Turkish and Australian students. According to the results of this study, the variable of value given to learning mathematics was not found to be related to achievement in all years in both countries. In light of the findings, the fact that students do not see mathematics as important does not prevent them from studying and succeeding in the course (Ivanova & Michaelides, 2022). When the effect size values in the current study are analysed, a non-significant effect can be mentioned. In 2018, some changes were made to the mathematics curriculum, and mathematics subjects started to be associated with daily life problems. This change may have been reflected in the TIMSS 2019 results.

Apart from affective characteristics, the gender factor also comes to the fore as a student characteristic. Considering the findings of the current study, a significant relationship was found between gender and mathematics achievement in all years except 2015. In other words, the mathematics achievement of female students is higher than the mathematics achievement of male students. The 2019 High School Entrance Exam (LGS) also yielded similar results. In the LGS mathematics subtest, female students scored higher than male students (Şensoy et al., 2019).

Aydın (2015) obtained similar results in his study with TIMSS 2011 data and showed that the mathematics achievement of female students was higher than that of male students. However, when the author analysed the effect size of the gender variable in his study, he stated that it had a small effect and was not practically significant. On the other hand, the results of the current study contradict some studies in the literature (Bassey et al., 2011; Butt & Dogar, 2014; Mohammadpour, 2013; Ross et al., 2012; Topal, 2021; Yayan & Berberoglu, 2004). Although there is a common belief that men are more successful in mathematics, some meta-analyses show that this belief is not true. Lindberg et al. (2010) summarized 242 studies conducted between 1990 and 2007 and found that gender had no significant effect on mathematics achievement.

The language of the test spoken at home was defined as the frequency of speaking Turkish. Considering that various ethnic groups live in Türkiye and that these ethnic groups preserve their own languages, it can be said that the language of the test spoken at home is important for students in Türkiye. Türkiye is constantly receiving immigrants from war-torn countries such as Afghanistan and Syria and is also a bridge between Europe and Asia. Therefore, there are many children of immigrants in the country (Yılmaz & Şekerci 2016). The results of the current study show that the frequency of students speaking Turkish at home has a significant effect on TIMSS 2011, 2015 and 2019 mathematics achievement. In other words, as the frequency of students speaking Turkish at home increases, their mathematics achievement also increases. Similar results have been obtained in other countries as well (e.g., Chinese Taipei, Hong Kong and Singapore) (Chen, 2014; Sandoval-Hernández & Białowolski, 2016). Looking at the effect sizes for each year, it can be concluded that having the same language spoken at home as the language of the test has a small effect on mathematics achievement. In the TIMSS study, there are skill-based questions in which students are expected to use their ability to understand the problem and produce an answer. Although the questions are designed for mathematical cognitive domain skills, it is also very important to use language skills such as reading comprehension since students cannot produce the correct answer if they do not understand what the question is asking. Therefore, it is necessary to improve the Turkish language skills of students whose mother tongue is not Turkish.

The number of books in the student's home, having a room of his/her own, the level of computer use at home, and the educational level of his/her parents (mother and father) are defined as the student's educational resources at home. Studies in the related literature show that educational resources at home are related to students' mathematics achievement and that students with high access to these resources have higher achievement (Akyüz, 2014; Acar-Güvendir, 2014; Koyuncu, 2021; Mullis et al., 2016; Oral & McGivney, 2013; Özer & Anıl, 2011; Topal, 2021; Topçu et al., 2016; Yayan & Berberoglu, 2004). The findings of the current study also yielded parallel results with those of the literature. The educational resources at home variable examined in the study emerged as an important determinant of mathematics achievement in all years. Although this variable was considered a control variable for socioeconomic status at the school level, the effect of educational resources at the student level still persisted. Therefore, it is important to increase the educational resources at students' homes. To this end, in Türkiye, between 2012 and 2015, 1,437,800 tablet computers were distributed to students under the FATIH project (http://fatihprojesi.meb.gov.tr/tablet_seti.html). Although this project was a step towards increasing educational resources at home, it was not sufficient on its own and was later shelved. According to the findings of the current study, it can be said that the FATIH project did not make a difference in the TIMSS results, since the effect of educational resources at home on academic achievement emerged in all years. Therefore, more comprehensive and effective ways to increase students' educational resources at home should be sought. In order to eliminate the inequalities arising from the educational opportunities at home, various practices can be carried out in the classroom or at school. For example, enriching the library corner in the classroom or making the computer lab available to students outside class hours can be among the steps to be taken for equal opportunity.

The school's socioeconomic status was identified as the factor that most influenced the achievement of eighth grade students in mathematics. Considering the effect sizes of the variables in the multivariate model in all years, it is seen that the school's socioeconomic status is in the first place. The study reveals that the school's socioeconomic status has a positive relationship with academic achievement. These findings indicate that schools with students with higher socioeconomic status have higher mathematics achievement. Especially in Türkiye, studies conducted by Arifoğlu (2019) and Gustafsson et al. (2018) confirm that school's socioeconomic status has a significant effect on students' mathematics achievement.

Gustafsson et al. (2018) compared TIMSS 2011 data from 50 different countries and found that the mathematics achievement of eighth grade students in Türkiye was related to school socioeconomic status. Similarly, Arifoğlu (2019) examined the factors affecting the mathematics achievement of both fourth and eighth grade students using TIMSS 2015 data from Türkiye. As a result of the study, it was found that the school's socioeconomic status was a significant variable affecting mathematics achievement for both grade levels. These results indicate that the academic achievement of economically disadvantaged students in Türkiye should be lower than expected. This situation is the main indicator of non-compliance with the principle of equal opportunities in education (Coleman et al., 1966). When the effect sizes on the basis of years are analysed, it is seen that the effect size increases as we progress from 2011 to 2019. This means that the achievement gap between economically strong and economically weak students has increased over time. This gap in students' achievement persists into adulthood and increases the economic imbalance in society. Therefore, national and local policies should be developed for disadvantaged students, and learning opportunities in schools be improved.

4.1. Conclusion

The Turkish education system has undergone radical changes since 2003 which are based on the impact of comprehensive international studies (TIMSS, PISA, and PIRLS). However, when the results of the present study are analysed, it is observed that the variables affecting student achievement have not changed in the last 13 years, which raises a serious question mark about the effectiveness of the reform policies. The results reveal that factors such as students' self-confidence in mathematics, access to educational resources at home, the language spoken at home being Turkish, and the school's socioeconomic status are determinants of their academic achievement.

In this context, there are concerns about the adequacy of the interventions made for educational reform. The fact that the factors affecting student achievement have remained relatively constant suggests that the reforms have not contributed sufficiently to achievement. In particular, access to educational resources at home plays a significant role in students' mathematics achievement. In addition, the language of the test spoken at home has a significant impact on mathematics achievement and also the school's socioeconomic status is a critical factor determining student achievement.

In conclusion, although it is difficult to give a clear answer to the extent to which the reforms in the Turkish education system have contributed to student achievement, factors such as access to educational resources at home, the language of the test spoken at home, and the school's socioeconomic status seem to play a decisive role in student achievement. Therefore, these factors should be taken into consideration when determining educational policies.

4.2. Limitations and Suggestions

There are some limitations in the present study. Firstly, this study was designed within the scope of a relational model. For this reason, the findings should not be interpreted as a cause-and-effect relationship. Next, the current study focused only on the mathematics achievement of 8th grade students in TIMSS data. Therefore, interested researchers can compare the results obtained from this study with the results obtained from 4th grade students by working with their data. In addition, this study has shown that educational resources at home is an important variable; however, which of the variables under the index variable of educational resources at home, such as the number of books in the home, having an individual room, having access to a computer, and parental education levels, is more important is beyond the scope of the current study. For this reason, it is necessary to determine which of these variables is more important in order to develop an education policy in line with the results obtained. Especially with the

COVID-19 pandemic, students' access to education from home has become more critical, and it has become difficult to ensure the principle of equal opportunity in education (Özer & Suna, 2020; Özer et al., 2020). Considering the possibility that the effects of this unexpected situation may be reflected in future TIMSS data, the model used in the current study should be tested again with TIMSS 2023 data.

Declaration of Conflicting Interests and Ethics

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

Authorship Contribution Statement

Burcin Coşkun: Investigation, Resources, Methodology, Visualization, Software, Formal Analysis, and Writing-original draft. **Kübra Karakaya Özyer**: Literature Review, Methodology, Supervision, and Writing-original draft.

Orcid

Burcin Coşkun https://orcid.org/0000-0002-6974-4353
Kübra Karakaya Özyer https://orcid.org/0000-0002-0208-7870

REFERENCES

Abalı-Öztürk, Y., & Şahin, Ç. (2015). Matematiğe ilişkin akademik başarı-özyeterlilik ve tutum arasındaki ilişkilerin belirlenmesi [Determining the relationships between academic achievement, self-efficacy and attitudes towards maths]. *The Journal of Academic Social Science Studies*, *31*(1), 343-366. http://dx.doi.org/10.9761/JASSS2621

Abazaoğlu, İ., Yıldızhan, Y., & Yatağan, M. (2015). Science and mathematics education in Turkey. *İlköğretim Online*, *14*(2), 557-573. https://doi.org/10.17051/io.2015.77272

Acar-Güvendir, M. (2014). Öğrenci başarılarının belirlenmesi sınavında öğrenci ve okul özelliklerinin Türkçe başarısı ile ilişkisi [Student and school characteristics' relation to Turkish achievement in student achievement determination exam]. *Eğitim ve Bilim, 39*(172), 163-180.

Akgündüz, D., Aydeniz, M., Çakmakçı, G., Çavaş, B., Çorlu, M.S., Öner, T., & Özdemir, S. (2015). *STEM eğitimi Türkiye raporu* [*STEM education Turkiye report*]. Scala Basım.

Aksu, G., Guzeller, C.O., & Eser, M.T. (2017). Analysis of maths literacy performances of students with Hierarchical Linear Modeling (HLM): The case of PISA 2012 Turkey. *Education & Science*, *42*(191), 247-266. https://doi.org/10.15390/EB.2017.6956

Akyüz, G. (2014). The effects of student and school factors on mathematics achievement in TIMSS 2011. *Education and Science*, *39*(172), 150-162. https://egitimvebilim.ted.org.tr/index.php/EB/article/view/2867

Arıkan, S. (2017). TIMSS 2011 verilerine göre Türkiye'deki ev ödevi ve matematik başarısı arasındaki ilişki [The relationship between homework and mathematics achievement in Turkey according to TIMSS 2011]. *International Journal of Eurasia Social Sciences*, *8*(26), 256-276. https://www.researchgate.net/publication/315726438_TIMSS_2011_VERILERINE_GORE_TURKIYE'DEKI_EV_ODEVI_VE_MATEMATIK_BASARISI_ARASINDAKI_ILISKI

Arıkan, S., van de Vijver, F.J.R., & Yağmur, K. (2016). Factors contributing to mathematics achievement differences of Turkish and Australian students in TIMSS 2007 and 2011. *EURASIA Journal of Mathematics*, *Science and Technology Education*, *12*(8), 2039-2059.

Arifoğlu, A. (2019). *Öğrenci başarısına okul etkisinin araştırılması: TIMSS 2015 Türkiye verisine göre çok düzeyli bir analiz* [*Investigating the effect of school on student* *achievement: A multilevel analysis based on TIMSS 2015 Turkey data*] [Doctoral Disseration]. Hacettepe University.

Aydın, M. (2015). *Öğrenci ve okul kaynaklı faktörlerin TIMMS matematik başarısına etkisi* [*The effect of student and school-related factors on TIMMS mathematics achievement*]. [Unpuplished Doctoral Disseration]. Necmettin Erbakan University.

Bandura, A., Barbaranelli, C., Caprara, G.V., & Pastorelli, C. (2001). Self-efficacy beliefs as shapers of children's aspirations and career trajectories. *Child Development*, *72*(1), 187-206. https://doi.org/10.1111/1467-8624.00273

Barış, F. (2009). *TIMSS-R ve TIMSS-2007 sınavlarının öğrenci başarısını yordayan değişkenler açısından incelenmesi* [*Investigation of TIMSS-R and TIMSS-2007 exams in terms of variables predicting student achievement*][Master's Thesis]. Hacettepe University.

Bassey, S.W., Joshua, M.T., & Asim, A.E. (2011). Gender differences and mathematics achievement of rural senior. *Mathematics Connection*, 1-8.

Belbase, S. (2013). Images, anxieties, and attitudes toward mathematics. *International Journal of Education in Mathematics, Science and Technology*, *1*(4), 230-237.

Bellibas, M.S. (2016). Who are the most disadvantaged? Factors associated with the achievement of students with low socio-economic backgrounds. *Educational Sciences: Theory and Practice*, *16*(2), 691-710. https://doi.org/10.12738/estp.2016.2.0257

Bos, K., & Kuiper, W. (1999). Modelling TIMSS data in a European comparative perspective: Exploring influencing factors on achievement in mathematics in grade 8. *Educational Research and Evaluation*, *5*(2), 157-179. https://doi.org/10.1076/edre.5.2.157.6946

Broer, M., Bai, Y., & Fonseca, F. (2019). *Socioeconomic inequality and educational outcomes: Evidence from twenty years of TIMSS*. Springer Nature.

Bryk, A.S., & Raudenbush, S.W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Sage Publications.

Butt, I.H., & Dogar, A.H. (2014). Gender disparity in mathematics achievement among the rural and urban high school students in Pakistan. *Pakistan Journal of Social Sciences*, *34*(1), 93-100.

Büyüköztürk, Ş., Çakan, M., Tan, Ş., & Atar, H.Y. (2014). *TIMSS 2011 ulusal matematik ve fen raporu: 8. Sınıflar* [*TIMSS 2011 national math and science report: Grade 8*]. Yenilik ve Eğitim Teknolojileri Genel Müdürlüğü, Milli Eğitim Bakanlığı. https://timss.meb.gov.tr/meb_iys_dosyalar/2022_03/07135958_TIMSS-2011-8-Sinif.pdf

Chen, Q. (2014). Using TIMSS 2007 data to build mathematics achievement model of fourth graders in Hong Kong and Singapore. *International Journal of Science and Mathematics Education*, *12*, 1519-1545. https://doi.org/10.1007/s10763-013-9505-x

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

Chowa, G.A., Masa, R.D., Ramos, Y., & Ansong, D. (2015). How do student and school characteristics influence youth academic achievement in Ghana? A hierarchical linear modeling of Ghana YouthSave baseline data. *International Journal of Educational Development*, *45*, 129-140. https://doi.org/10.1016/j.ijedudev.2015.09.009

Cingöz, Z.K., & Gür, B. (2020). The effect of economic, social and cultural status on academic achievement a comparison of PISA 2015 and TEOG 2017 results. *İnsan ve Toplum*, *10*(4), 247-288. https://doi.org/10.12658/M0563

Coleman, J.S. (1994). Family, school, and social capital. In T. Husen & T. N. Postlethwaite (Eds.), *International encyclopedia of education* (2nd ed., pp. 2272–2274). Pergamon Press.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D., & York, R.L. (1966). *Equality of educational opportunity*. Government Printing Office. https://files.eric.ed.gov/fulltext/ED012275.pdf

Coşkun, B., & Karadağ, E. (2023). The effect of student and school characteristics on TIMSS 2015 science and mathematics achievement: The case of Türkiye. *Journal of Pedagogical Research*, *7*(1), 203-227. https://doi.org/10.33902/JPR.202318875

Çalışkan, M. (2014). Bir derse yönelik duyuşsal giriş özelliklerinin belirlenmesi: Bir ölçme modeli önerisi [Determination of affective entry characteristics for a particular course: A measurement model suggestion]. *Kastamonu Eğitim Dergisi*, *22*(1), 57-68. https://dergipark.org.tr/en/pub/kefdergi/issue/22603/241532

Çavdar, D. (2015). *TIMSS 2011 matematik başarısının öğrenci ve öğretmen özellikleri ile ilişkisi* [*The relationship between student and teacher characteristics and mathematics achievement in TIMSS 2011*] [Master's Thesis]. Gazi University.

Çiftçi, Ş.K., & Yıldız, P. (2019). The effect of self-confidence on mathematics achievement: The meta-analysis of Trends in International Mathematics and Science Study (TIMSS). *International Journal of Instruction*, *12*(2), 683-694. https://doi.org/10.29333/iji.2019.12243a

Demir, I., & Kılıç, S. (2010). Using PISA 2003, examining the factors affecting students' mathematics achievement. *Hacettepe University Journal of Education*, *38*(38), 44-54.

Doğan, N., & Barış, F. (2010). Tutum, değer ve özyeterlik değişkenlerinin TIMSS-1999 ve TIMSS-2007 sınavlarında öğrencilerin matematik başarılarını yordama düzeyleri [The prediction levels of attitude, value and self-efficacy variables on students' mathematics achievement in TIMSS-1999 and TIMSS-2007 exams]. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *1*(1), 44-50. https://dergipark.org.tr/tr/pub/epod/issue/5808/77253

Dönmez, S.M.K., & Dede, Y. (2020). Ortaöğretime geçiş sınavları matematik sorularının matematiksel yeterlikler açısından incelenmesi [Analysis of mathematics questions in the exams for transition to secondary education in terms of mathematical proficiency]. *Baskent University Journal of Education*, *7*(2), 363-374. https://buje.baskent.edu.tr/index.php/buje/article/view/327

Eğitimi Araştırma ve Geliştirme Dairesi Başkanlığı. [EARGED]. (2011). *MEB 21. yy öğrenci profili* [*MoNE 21st century student profile*]. https://www.meb.gov.tr/earged/earged/21.%20yy_og_pro.pdf

Erdoğan, E., & Acar-Güvendir, M. (2019). Uluslararası öğrenci değerlendirme programında öğrencilerin sosyoekonomik özellikleri ile okuma becerileri arasındaki ilişki [The relationship between students socioeconomic attributes and their reading skills in Programme For International Student Assessment]. *Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi*, *20*, 493-523. https://doi.org/10.17494/ogusbd.548530

Ersan, O., & Rodriguez, M.C. (2020). Socioeconomic status and beyond: A multilevel analysis of TIMSS mathematics achievement given student and school context in Turkey. *Large-scale Assessments in Education*, *8*, 1-32. https://doi.org/10.1186/s40536-020-00093-y

Erşan, Ö. (2016). *TIMSS 2011 sekizinci sınıf öğrencilerinin matematik başarılarını etkileyen faktörlerin çok düzeyli yapısal eşitlik modeliyle incelenmesi* [*Investigating the factors affecting the mathematics achievement of TIMSS 2011 eighth grade students with multilevel structural equation modeling*] [Master's Thesis]. Hacettepe University.

Ertürk, Z., & Erdinç-Akan, O. (2018). TIMSS 2015 matematik başarısını etkileyen değişkenlerin yapısal eşitlik modeli ile incelenmesi [The investigation of the variables effecting TIMSS 2015 mathematics achievement with SEM]. *Ulusal Eğitim Akademisi Dergisi*, *2*(2), 14-34. https://doi.org/10.32960/uead.407078

Ferla, J., Valcke, M., & Cai, Y. (2009). Academic self-efficacy and academic self-concept: Reconsidering structural relationships. *Learning and Individual Differences*, *19*(4), 499-505. https://doi.org/10.1016/j.lindif.2009.05.004

Foy, P., & Yin, L. (2016). Scaling the TIMSS 2015 achievement data. In M.O. Martin, I.V.S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS*. http://timss.bc.edu/publications/timss/2015-methods/chapter-13.html

Foy, P., Fishbein, B., von Davier, M., & Yin, L. (2020). Implementing the TIMSS 2019 scaling methodology. In M.O. Martin, M. von Davier, & I.V.S. Mullis (Eds.), *Methods and Procedures: TIMSS 2019 Technical Report* (pp.12.1-12.146). https://timssandpirls.bc.edu/timss2019/methods/chapter-12.html

Garson, G.D. (2019). *Multilevel modeling: Applications in STATA®, IBM® SPSS®, SAS®, R, & HLM*. SAGE Publications.

Gelbal, S. (2008). The effect of socio-economic status of eighth grade students on their achievement in Turkish. *Egitim ve Bilim*, *33*(150), 1-13.

Gronmo, L.S., Lindquist, M., Arora, A., & Mullis, I.V.S. (2013). TIMSS 2015 mathematics framework (Chapter 1). In I.V.S. Mullis & M.O. Martin (Ed), *TIMSS 2015 Assessment Frameworks* (pp. 11-27). http://timssandpirls.bc.edu/timss2015/frameworks.html

Guo, J., & Woulfin, S. (2016). Twenty-first century creativity: An investigation of how the partnership for 21st century instructional framework reflects the principles of creativity. *Roeper Review*, *38*(3), 153-161.

Gustafsson, J.E., Nilsen, T., & Hansen, K.Y. (2018). School characteristics moderating the relation between student socio-economic status and mathematics achievement in grade 8. Evidence from 50 countries in TIMSS 2011. *Studies in Educational Evaluation*, *57*, 16-30. https://doi.org/10.1016/j.stueduc.2016.09.004

Hansford, B.C., & Hattie, J.A. (1982). The relationship between self and achievement/perfor mance measures. *Review of Educational Research*, *52*(1), 123-142.

Heck, R.H., & Thomas, S.L. (2015*). An introduction to multilevel modeling techniques: MLM and SEM approaches using Mplus*. Routledge.

Hilton, T.L., & Lee, V.E. (1988). Student interest and persistence in science: Changes in the educational pipeline in the last decade. *The Journal of Higher Education*, *59*(5), 510-526. https://doi.org/10.1080/00221546.1988.11780210

Hox, J.J., Moerbeek, M., & Van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications*. Routledge.

Ismail, N.A. (2009). Understanding the gap in mathematics achievement of Malaysian students. *The Journal of Educational Research*, *102*(5), 389-394. https://doi.org/10.3200/JOER.102.5.389-394

Ismail, N.A., & Awang, H. (2008). Differentials in mathematics achievement among eighth-grade students in Malaysia. *International Journal of Science and Mathematics Education*, *6*, 559-571. https://doi.org/10.1007/s10763-007-9109-4

Işlak, O. (2020). *Prediction of mathematics achievement of students attending TIMSS 2015 according to student, family and school variables* [Doktoral dissertation]. Burdur Mehmet Akif Ersoy University.

Ivanova, M., & Michaelides, M.P. (2022). Motivational components in TIMSS 2015 and their effects on engaging teaching practices and mathematics performance. *Studies in Educational Evaluation*, *74*, 101-173. https://doi.org/10.1016/j.stueduc.2022.101173

Kadijević, Đ. (2008). TIMSS 2003: Relating dimensions of mathematics attitude to mathematics achievement. *Zbornik Instituta za Pedagoska İstrazivanja*, *40*(2), 327-346.

Kalaycioglu, D.B. (2015). The influence of socioeconomic status, self-efficacy, and anxiety on Mathematics achievement in England, Greece, Hong Kong, the Netherlands, Turkey, and the USA. *Educational Sciences: Theory and Practice*, *15*(5), 1391-1401. https://doi.org/10.12738/estp.2015.5.2731

Kaleli-Yılmaz, G., & Hanci, A. (2016). Examination of the 8th grade students' TIMSS mathematics success in terms of different variables. *International Journal of Mathematical Education in Science and Technology*, *47*(5), 674-695. https://doi.org/10.1080/0020739X.2015.1102977

Kara, M. (2023). *TIMSS 2019 matematik başarısını açıklayan değişkenlerin çok düzeyli yapısal eşitlik modeli ile incelenmesi* [*Examining the variables explaining TIMSS 2019 Mathematics achievement with multilevel structural equation modeling*] [Master's Degree Thesis]. Hacettepe Univesity.

Karaca, F. (2018). *Sekizinci sınıf öğrencilerinin TIMSS matematik başarılarının bazı değişkenler açısından incelenmesi: Eskişehir ili örneği* [*Investigation of TIMSS mathematics achievement of eighth grade students in terms of some variables: The case of Eskisehir province*] [Master's Thesis]. Eskişehir Osmangazi University.

Kaya, S. (2008). *The effects of student-level and classroom-level factors on elementary students' science achievement in five countries* [Doktoral dissertation]. The Florida State University.

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

Khine, M.S., Al-Mutawah, M., & Afari, E. (2015). Determinants of affective factors in mathematics achievement: Structural equation modeling approach. *Journal of Studies in Education*, *5*(2), 199-211. https://www.researchgate.net/publication/276294926_Determinants_of_Affective_Factors_in_Mathematics_Achievement_Structural_Equation_Modeling_Approach

Kilic, S., & Askin, Ö.E. (2013). Parental influence on students' mathematics achievement: The comparative study of Turkey and best performer countries in TIMSS 2011. *Procedia-Social and Behavioral Sciences*, *106*, 2000-2007.

Kim, S.J., Park, J.H., Park, S.W., & Kim, S.S. (2013). *The effects of school and students' educational contexts in Korea, Singapore, and Finland using TIMSS 2011*. 5th IEA International Research Conference, Singapore. http://www.iea.nl/fileadmin/user_upload/IRC/IRC_2013/Papers/IRC-2013_Kim_etal.pdf

Koyuncu, İ. (2021). TIMSS international benchmarks of eighth graders in mathematics: A correspondence analysis study. *International Electronic Journal of Elementary Education*, *14*(2), 179-194.

LaRoche, S., Joncas, M., & Foy, P. (2016). Sample design in TIMSS 2015. In M.O. Martin, I.V.S. Mullis, & M. Hooper (Eds.), *Methods and procedures: TIMSS 2015 technical report*. http://timss.bc.edu/publications/timss/2015 -methods/chapter-3.html

LaRoche, S., Joncas, M., & Foy, P. (2020). Sample design in TIMSS 2019. In M.O. Martin, M. Von Davier, & I.V.S. Mullis (Eds.), *Methods and procedures: TIMSS 2019 technical report*. https://timssandpirls.bc.edu/timss2019/methods/chapter-3.html

Leder, G.C., & Forgasz, H.J. (2006). Affect and mathematics education: PME perspectives. In *Handbook of research on the psychology of mathematics education* (pp. 403-427). Brill.

Lee, C.Y., & Kung, H.Y. (2018). Math self-concept and mathematics achievement: Examining gender variation and reciprocal relations among junior high school students in Taiwan. *Eurasia Journal of Mathematics, Science and Technology Education*, *14*(4), 1239-1252. https://doi.org/10.29333/ejmste/82535

Lee, J. (2020). *Non-cognitive characteristics and academic achievement in Southeast Asian countries based on PISA 2009, 2012, and 2015* (Working Paper No. 233). OECD Education.

Lee, J., & Chen, M. (2019). Cross-country predictive validities of non-cognitive variables for mathematics achievement: Evidence based on TIMSS 2015. *EURASIA Journal of Mathematics, Science and Technology Education*, *15*(8), 2-16. https://doi.org/10.29333/ejmste/106230

Lee, J., & Stankov, L. (2018). Non-cognitive predictors of academic achievement: Evidence from TIMSS and PISA. *Learning and Individual Differences*, *65*, 50-64. https://doi.org/10.1016/j.lindif.2018.05.009

Lindberg, S.M., Hyde, J.S., Petersen, J.L., & Linn, M.C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin*, *136*(6), 1123–1135. https://doi.org/10.1037/a0021276

Liou, P.Y. (2014). Investigation of the big-fish-little-pond effect on students' selfconcept of learning mathematics and science in Taiwan: Results from TIMSS 2011. *Asia-Pacific Education Research*, *23*(3), 769-778. https://doi.org/10.1007/s40299-013-0152-3

Louis, R.A., & Mistele, J.M. (2012). The differences in scores and self-efficacy by student gender in mathematics and science. *International Journal of Science and Mathematics Education*, *10*, 1163-1190. https://doi.org/10.1007/s10763-011-9325-9

Luo, W., Li, H., Baek, E., Chen, S., Lam, K.H., & Semma, B. (2021). Reporting practice in multilevel modeling: A revisit after 10 years. *Review of Educational Research, 91*(3), 311-355.

Martin, M.O., Mullis, I.V.S., & Foy, P. (2013). TIMSS 2015 assessment design. In I.V.S. Mullis, & M.O. Martin, (Eds.), *TIMSS 2015 assessment frameworks* (pp. 85–100). TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2015/frameworks.html

Martin, M.O., Mullis, I.V.S., Gonzales, E.J., Gregory, K.D., Smith, T.A., Chrostowski, S.J. et al. (2000). *TIMSS 1999: International science report*. International Study Center.

Mau, W.C., & Lynn, R. (2010). Gender differences in homework and test scores in mathematics, reading and science at tenth and twelfth grade. *Journal Psychology, Evolution & Gender*, *2*, 119-125.

Mohammadpour, E. (2012). A multilevel study on trends in Malaysian secondary school students' science achievement and associated school and student predictors. *Science Education*, *96*(6), 1013-1046. https://doi.org/10.1002/sce.21028

Mohammadpour, E. (2013). A three-level multilevel analysis of Singaporean eighth-graders science achievement. *Learning and Individual Differences*, *26*, 212-220. https://doi.org/10.1016/j.lindif.2012.12.005

Mohammadpour, E., & Abdul Ghafar, M.N. (2014). Mathematics achievement as a function of within-and between-school differences. *Scandinavian Journal of Educational Research*, *58*(2), 189-221. https://doi.org/10.1080/00313831.2012.725097

Mullis, I.V.S., & Martin, M.O. (2017). *TIMSS 2019 assessment frameworks*. International Association for the Evaluation of Educational Achievement.

Mullis, I.V.S., & Fishbein, B. (2020). Updating the TIMSS 2019 instruments for describing the contexts for student learning. In M. O. Martin, M. von Davier, & I.V.S. Mullis (Eds.), *Methods and Procedures: TIMSS 2019 Technical Report*. https://timssandpirls.bc.edu/timss2019/methods/chapter-2.html

Mullis, I.V.S., Martin, M.O., Foy, P., Kelly, D.L., & Fishbein, B. (2020). *TIMSS 2019 international results in mathematics and science*. https://timss2019.org/reports/

Mullis, I.V.S., Martin, M., & Loveless, T. (2016). 20 years of TIMSS. *Trends in International Mathematics and Science Study*. http://timssandpirls.bc.edu/timss2015/internationalresults/timss2015/wp-content/uploads/2016/T15-20-years-of-TIMSS.pdf

Mullis, I.V.S., Martin, M.O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. TIMSS & PIRLS International Study Center, Boston College.

Musca, S.C., Kamiejski, R., Nugier, A., Méot, A., Er-Rafiy, A., & Brauer, M. (2011). Data with hierarchical structure: impact of intraclass correlation and sample size on type-I error. *Frontiers in Psychology*, *2*(74), 1-6. https://doi.org/10.3389/fpsyg.2011.00074

National Center for Education Statistics. (n. d.). *Overview*. https://nces.ed.gov/timss/overview.asp

Okatan, Ö., & Tomul, E. (2021). Uluslararası öğrenci başarılarını değerlendirme programı'na (PISA) göre Türkiye'deki öğrencilerin matematik başarıları ile ilişkili değişkenlerin incelenmesi [Investigation of interrelated variables with students success in mathematics according to Programme For International Student Assessment (PISA)]. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi*, *57*, 98-125. https://dergipark.org.tr/en/pub/maeuefd/issue/60050/663150

Olatunde, Y.R. (2010). Socio-economic background and mathematics achievement of students in some selected senior secondary schools in Southwestern Nigeria. *Pakistan Journal of Social Sciences*, *7*(1), 23-27.

Oral, I., & McGivney, E. (2013). *Student performance in math and science in Turkey and determinants of success*. Education Reform Initiative.

Ölçüoğlu, R., & Çetin, S. (2016). TIMSS 2011 sekizinci sınıf öğrencilerinin matematik başarısını etkileyen değişkenlerin bölgelere göre incelenmesi [The investigation of the variables that affecting eight grade students' TIMSS 2011 math achievement according to regions]. *Journal of Measurement and Evaluation in Education and Psychology*, *7*(1), 202-220. https://doi.org/10.21031/epod.34424

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

Özdemir, A., & Gelbal, S. (2016). Predictive power of primary and secondary school success criterion on transition to higher education examination scores. *Journal of Measurement and Evaluation in Education and Psychology*, *7*(2), 309-334.

Özdemir, C. (2016). Equity in the Turkish education system: A multilevel analysis of social background influences on the mathematics performance of 15-year-old students. *European Educational Research Journal*, *15*(2), 193-217. https://doi.org/10.1177/1474904115627159

Özer, Y., & Anıl, D. (2011). Examining the factors affecting students' science and mathematics achievement with structural equation modeling. *Hacettepe University Journal of Education*, *41*, 313-324.

Özer, M., & Suna, H. E. (2020). *Covid-19 salgını ve eğitim* [*Covid-19 pandemic and education*]. In M. Şeker, A. Özer & C. Korkut (Eds.), *Küresel toplumun anatomisi: İnsan ve toplumun geleceği* (p. 171-192). Tuba Publication.

Özer, M., Suna, H.E., Aşkar, P., & Çelik, Z. (2020). The impact COVID-19 school closures on educational inequalities. *Insan ve Toplum*, *10*(4), 217-246. http://doi.org/10.12658/M0611

Özkan, Y.Ö., & Acar-Güvendir, M. (2014). Socioeconomic factors of students' relation to mathematic achievement: comparison of PISA and ÖBBS. *International Online Journal of Educational Sciences*, *6*(3), 776-789. http://dx.doi.org/10.15345/iojes.2014.03.020

Öztekin-Bayır, Ö., & Tekel, E. (2021). Problems of Turkish education system and suggested solutions: What do pre-service teachers think?. *Journal of Pedagogical Research*, *5*(1), 275-292. http://dx.doi.org/10.33902/JPR.2021167894

Pajares, F., & Miller, M.D. (1997). Mathematics self-efficacy and mathematical problem solving: Implications of using different forms of assessment. *The Journal of Experimental Education*, *65*(3), 213-228. https://doi.org/10.1080/00220973.1997.9943455

Papanastasiou, C. (2000). Internal and external factors affecting achievement in mathematics: Some findings from TIMSS. *Studies in Educational Evaluation*, *26*(1), 1-8.

Phan, H., Sentovich, C., Kromrey, J., Dedrick, R., & Ferron, J. (2010). *Correlates of mathematics achievement in developed and developing countries: An analysis of TIMSS 2003 eighth-grade Mathematics scores*. American Educational Research Association, Colorado, USA.

Polat, M., Gönen, E., Parlak, B., Yıldırım, A., & Özgürlük, B. (2016). *TIMSS 2015 ulusal matematik ve fen bilimleri ön raporu 4. ve 8. Sınıflar* [*TIMSS 2015 national math and science preliminary report grades 4 and 8*]. Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü, Milli Eğitim Bakanlığı. https://timss.meb.gov.tr/meb_iys_dosyalar/2022_03/07135609_TIMSS_2015_Ulusal_Rapor.pdf

Portes, A., & MacLeod, D. (1996). Educational progress of children of immigrants: The roles of class, ethnicity, and school context. *Sociology of Education*, *69*(5), 255-275. https://doi.org/10.2307/2112714

Raudenbush, S.W., Bryk, A.S., Cheong, Y.F., Congdon, R.T., & Du Toit, M. (2011). *HLM 7: Hierarchical linear and nonlinear modeling*. Scientific Software International.

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Rosenthal, R., & Rosnow, R. (1984). *Essentials of behavioral research: Methods and data analysis.* McGraw-Hill.

Ross, J.A., Scott, G., & Bruce, C.D. (2012). The gender confidence gap in fractions knowledge: Gender differences in student belief-achievement relationships. *School Science and Mathematics*, *112*(5), 278-288. https://doi.org/10.1111/j.1949-8594.2012.00144.x

Ryan, R.M., & Deci, E.L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*(1), 54–67.

Sandoval-Hernández, A., & Białowolski, P. (2016). Factors and conditions promoting academic resilience: A TIMSS-based analysis of five Asian education systems. *Asia Pacific Education Review*, *17*(3), 511-520.

Sarı, M.H., Arıkan, S., & Yıldızlı, H. (2017). 8. sınıf matematik akademik başarısını yordayan faktörler-TIMSS 2015 [Factors predicting mathematics achievement of 8th graders in TIMSS 2015]. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 8(3), 246-265. https://doi.org/10.21031/epod.303689

Sarıer, Y. (2021). PISA uygulamalarında Türkiye'nin performansı ve öğrenci başarısını yordayan değişkenler [Turkey's performance in PISA applications and variables predicting student's success]. *Türkiye Sosyal Araştırmalar Dergisi*, *25*(3), 905-926. https://dergipark.org.tr/en/download/article-file/1167361

Sarouphim, K.M., & Chartouny, M. (2017). Mathematics education in Lebanon: Gender differences in attitudes and achievement. *Educational Studies in Mathematics*, *94*, 55-68. https://doi.org/10.1007/s10649-016-9712-9

Sevgi, S. (2009). *The connection between school and student characteristics with mathematics achievement in Turkey* [Master's thesis]. Middle East Technical University.

Snijders, T.A., & Bosker, R.J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage.

Suna, H.E., Şenso, S., Parlak, B., & Özdemir, E. (2020). *TIMSS 2019 Türkiye ön raporu* [*TIMSS 2019 Turkey preliminary report*]. Directorate of Measurement, Evaluation and Testing Services, Ministry of National Education. http://www.meb.gov.tr/meb_iys_dosyalar/2020_12/10173505_No15_TIMSS_2019_Turkiye_On_Raporu_Guncel.pdf

Suna, H.E., Tanberkan, H., Gür, B., Perc, M., & Özer, M. (2020). Socioeconomic status and school type as predictors of academic achievement. *Journal of Economy Culture and Society, 61*, 41-64. https://doi.org/10.26650/JECS2020-0034

Şensoy, S., Suna, H.E., Tanberkan, H., Eroğlu, E., & Altun, Ü. (2019). *2019 first placement results within the scope of the 2019 high school entrance system* (Report no 8). Ministry of National Education. https://mtegm.meb.gov.tr/meb_iys_dosyalar/2019_07/23104940_LGS_2019_yerlestirme_22temmuz.pdf

Şirin, S.R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417-53. https://doi.org/10.3102/00346543075003417

Tavsancil, E., & Yalcin, S. (2015). A determination of Turkish student's achievement using hierarchical linear models in Trends in International Mathematics-Science Study (TIMSS) 2011. *The Anthropologist, 22*(2), 390-396. https://doi.org/10.1080/09720073.2015.11891891

Thomson, S., Lokan, J., Lamb S., & Ainley, J. (2003). *Lessons from the third international mathematics and science study*. TIMSS Australia Monograph Series. Australian Council for Educational Research. https://research.acer.edu.au/timss_monographs/9/

Tomul, E., & Savasci, H.S. (2012). Socioeconomic determinants of academic achievement. *E ducational Assessment, Evaluation and Accountability*, *24*(3), 175-187. https://doi.org/10.1007/s11092-012-9149-3

Topal, H. (2021). Variable selection via the adaptive elastic net: Mathematics success of the students in Singapore and Turkey. *Journal of Applied Microeconometrics*, *1*(1), 41-55. https://doi.org/10.53753/jame.1.1.04

Topçu, M.S., Erbilgin, E., & Arıkan, S. (2016). Factors predicting Turkish and Korean students' science and mathematics achievement in TIMSS 2011. *Eurasia Journal of Mathematics Science and Technology Education, 12*(7), 1711-1737. https://doi.org/10.12973/eurasia.2016.1530a

Usta, H.G., & Demirtaşlı, R.N. (2018). PISA 2012 matematik okuryazarlığı üzerine uluslararası bir karşılaştırma: Türkiye ve Finlandiya [An international comparison according to PISA 2012 mathematical literacy Turkey and Finland]. *Electronic Turkish Studies*, *13*(11), 1389-1420. http://dx.doi.org/10.7827/TurkishStudies.13377

Wang, F., King, R.B., & Leung, S.O. (2023). Why do East Asian students do so well in mathematics? A machine learning study. *International Journal of Science and Mathematics Education, 21*(3), 691-711. https://doi.org/10.1007/s10763-022-10262-w

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817–838. https://doi.org/10.2307/1912934

Wigfield, A., & Eccles, J.S. (2000). Expectancy-value theory of achievement motivation. *Con temporary Educational Psychology, 25*(1), 68-81. https://doi.org/10.1006/ceps.1999.1015

Wilkins, J.L., & Ma, X. (2003). Modeling change in student attitude toward and beliefs about mathematics. *The Journal of Educational Research, 97*(1), 52-63. https://doi.org/10.1080/00220670309596628

Wilson, K., & Narayan, A. (2016). Relationships among individual task self-efficacy, self-regulated learning strategy use and academic performance in a computer-supported collaborative learning environment. *Educational Psychology*, *36*(2), 236-253. https://doi.org/10.1080/01443410.2014.926312

Yalcin, S., Demirtasli, R.N., Dibek, M.I., & Yavuz, H.C. (2017). The effect of teacher and student characteristics on TIMSS 2011 mathematics achievement of fourth-and eighth-grade students in Turkey. *International Journal of Progressive Education*, *13*(3), 79-94. https://eric.ed.gov/?id=EJ1159910

Yatağan, M. (2014). *Fen ve teknoloji dersi öğretim programının öğrenci ve öğretmen özelliklerine göre değerlendirilmesi: TIMSS 2007 ve 2011 verileri ile bir durum analizi* [*Evaluation of science and technology curriculum according to student and teacher characteristics: A case analysis with TIMSS 2007 and 2011 data*] [Unpublished Doctoral Dissertation]. Gazi University.

Yavuz, H.Ç., Demirtaşlı, R.N., Yalçın, S., & Dibek, M.İ. (2017). Türk öğrencilerin TIMSS 2007 ve 2011 matematik başarısında öğrenci ve öğretmen özelliklerinin etkileri [The effects of student and teacher level variables on TIMSS 2007 and 2011 mathematics achievement of Turkish students]. *Eğitim ve Bilim*, *42*(189), 27-47. https://doi.org/10.15390/EB.2017.6885

Yayan, B., & Berberoglu, G. (2004). A re-analysis of the TIMSS 1999 mathematics assessment data of the Turkish students. *Studies in Educational Evaluation*, *30*(1), 87-104.

Yetkiner-Özel, Z.E., Özel, S., & Thompson, B. (2013). SES-related mathematics achievement gap in Turkey compared to European Union countries. *Education & Science*, *38*(170), 179-193.

Yıldırım, H.H., Yıldırım, S., Ceylan, E., & Yetişir, M.İ. (2013). *Türkiye perspektifinden TIMSS 2011 sonuçları* [*TIMSS 2011 results from Turkey's perspective*]. Pelin Ofset Tipo Matbaacılık.

Yılmaz, F., & Şekerci, H. (2016). Ana dil sorunsalı: Sınıf öğretmenlerinin deneyimlerine göre ilkokul öğrencilerinin yaşadıkları sorunlar [Native language problematic: Problems faced by primary school students according to the experiences of classroom teachers]. *Eğitimde Nitel Araştırmalar Dergisi*, *4*(1), 47-63. https://dergipark.org.tr/en/pub/enad/issue/32029/354378

Yoshino, A. (2012). The relationship between self-concept and achievement in TIMSS 2007: A comparison between American and Japanese Students. *Int Rev Educ*, *58*, 199-219. https://doi.org/10.007/s11159-012-9283-7

APPENDIX

Descriptive Statistics for Dependent and Independent Variables

	Dependent variables	Year	n	Mean	se	Minimum value	Maximum value
PV1	First plausible value	2011	6850	450.79	108.58	105.73	839.23
		2015	5966	457.16	102.91	77.00	773.03
		2019	3930	491.92	106.95	128.39	871.37
PV2	Second plausible value	2011	6850	450.05	110.01	93.32	845.22
		2015	5966	457.92	103.38	30.88	780.64
		2019	3930	493.21	107.35	115.80	866.83
PV3	Third plausible value	2011	6850	449.27	111.70	59.20	875.19
		2015	5966	456.91	104.66	54.71	808.36
		2019	3930	492.63	108.59	100.62	888.37
PV4	Fourth plausible value	2011	6850	449.10	110.42	44.54	917.68
		2015	5966	454.63	107.13	69.46	794.79
		2019	3930	490.41	110.01	91.89	862.05
PV5	Fifth plausible value	2011	6850	450.03	110.61	95.53	840.44
		2015	5966	457.88	104.98	55.51	785.09
		2019	3930	491.69	107.86	117.26	838.33
	Independent variables						
Student level							
Student characteristics	Gender	2011	6850	0.50	0.50	0.00	1.00
		2015	5966	0.51	0.50	0.00	1.00
		2019	3930	0.50	0.50	0.00	1.00
	Language of the test spoken at home	2011	6850	3.67	0.73	1.00	4.00
		2015	5966	3.68	0.71	1.00	4.00
		2019	3930	3.62	0.77	1.00	4.00
	Educational resources at home	2011	6850	8.35	2.07	4.32	14.02
		2015	5966	9.11	1.90	4.23	13.88
		2019	3930	9.39	1.79	4.55	13.52
Affective characteristics	Self-confidence in learning mathematics	2011	6850	9.72	2.20	3.18	15.82
		2015	5966	9.75	2.29	3.20	15.93
		2019	3930	9.81	2.35	3.28	15.85
	Liking to learn mathematics	2011	6850	10.24	1.99	5.04	13.47
		2015	5966	10.26	1.98	4.97	13.98
		2019	3930	10.33	1.94	5.09	13.85
	Value given to learning mathematics	2011	6850	9.98	1.99	3.41	13.71
		2015	5966	10.06	2.10	3.00	13.65
		2019	3930	10.07	2.08	3.04	13.48
School level			*N*				
	School's socioeconomic status	2011	239	8.30	1.35	4.84	13.09
		2015	218	9.05	1.23	6.19	12.47
		2019	181	9.34	1.23	6.39	12.54

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 3, 482–495

https://doi.org/10.21449/ijate.1263531

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

Adaptation of compulsive sport consumption scale into Turkish culture: CSCS-T

Murat Aygün[1,*], **Sait Çüm**[2]

[1]Ardahan University, Faculty of Sport Sciences, Department of Sport Management, Ardahan, Türkiye.

[2]Dokuz Eylul University, Buca Faculty of Education, Department of Educational Measurement and Evaluation, Izmir, Türkiye.

ARTICLE HISTORY

Received: Mar. 10, 2023
Revised: Aug. 10, 2023
Accepted: Aug. 23, 2023

Keywords:

Compulsive sport consumption,
Scale adaptation,
Parallel analysis,
Validity,
Reliability.

Abstract: Consuming sports products and services incessantly without being able to restrain oneself is characterized as compulsive sports consumption. The aim of this study is to adapt the Compulsive Sport Consumption Scale (CSCS) developed in English by Aiken et al. (2018) into Turkish utilizing a scientific scale adaptation process. The CSCS consists of six items and is graded on a seven-point Likert scale ranging from strongly disagree to strongly agree. Higher CSCS levels are affiliated with psychological and behavioral constructs related to the effects of sports consumption, such as time, money, coping, and psychological and behavioral neglect. The scale has been tailored via a group of English and Turkish linguists, sports scientist, and psychometrist. Parallel analysis has been performed on account of inspecting the dimensionality of the scale, and many statistics such as unidimensional congruence, explained common variance, mean of item residual absolute loadings, and robust fit statistics have been used. In accordance with parallel analysis, the scale was unidimensional, and all other statistics supported that as well. The unidimensional adapted scale (CSCS-T) explained approximately 83% of the total variance. Additionally, internal consistency, composite reliability, and test-retest reliability have been examined to determine the measurement's reliability. Cronbach's Alpha was .958, McDonald's Omega was .958, and Pearson's product-moment correlation coefficient was .923 in the wake of the test-retest application. All of the findings propound that when investigating compulsive over-participation in sports consumption in Turkish-speaking populations, the CSCS-T can be used to acquire valid and reliable measures.

1. INTRODUCTION

It is known that contemporary western societies use sports and various social resources for individuals' lifestyles and identity achievement (Wheaton, 2000). In the historical process, sports were regarded as worthless in terms of economy until the 1970s. Nonetheless, investments by dint of economic support led to an increase in the value of sports after the 1970s (Lera-Lopez & Rapun-Garate, 2007). In the 21st century, the fact that sports are one of the most substantial economic resources in the world has induced societies to benefit from sports predominately. The proliferation of technology and the escalating competition have led to an

*CONTACT: Murat Aygün ✉ aygunmurat06@gmail.com 🖃 Ardahan University, Faculty of Sport Sciences, Department of Sport Management, Ardahan, Türkiye.

e-ISSN: 2148-7456 /© **IJATE** 2023

increase and diversification in consumption elements. This has culminated individuals into engage with sports consumption, which is covered the consumption phenomenon.

Numerous studies on sports consumption have been found in the literature. The first studies examined socio-demographic variables such as gender, age, and income status in relation to sports consumption in the 1970s and 1980s (Lera-Lopez & Rapun-Garate, 2007; Armstrong & Peretto Stratta, 2004), as well as the factors influencing sports participation (Hansen & Gauthier, 1989). Studies mainly included (a) relational structures including concepts such as trust, commitment and closeness, (b) interaction and media involved in providing communication, (c) demographic factors including variables such as age and gender among sports consumption and sports organization (Kim & Trail, 2011). These variables and structural differences are beneficial in understanding the individual's participation in sports and interpreting the relationship.

In addition to the factors affecting participation, this phenomenon is a form of hedonic consumption (Hopkinson & Pujari, 1999). Consumption is a substantial part of sports, which can be motivated by hedonic pleasure, supports emotional and cognitive states (Kempf, 1999) and provides sports consumers a wide perspective thanks to the moral excellence of sports (Jang et al., 2020). The ascending focus on fans, social media tools, sponsorship revenues, and advertisements have increased the economic viability of sports and ensured its presence in the sports market owing to the development of sports.

A significant issue of sports marketing is the content and necessity of sports consumption. The fact that sports products embody both physical and non-monetary services causes sports consumption behavior directly or indirectly (Yoshida & Nakazawa, 2016). Sports consumption impresses emotions, behavioral outcomes and motivation (Jang et al., 2021). Nevertheless, autonomous or controlled motivation is thought to be a factor in identifying the requirements in sports consumption (Kim & Mao, 2021). Consequently, sports consumption encapsulates all activities that individuals do with active or passive participation (Koning, 2009) in order to consume sports products and services immediately or later. Aside from the requirements that arise for the occurrence of sports consumption or the other factors that influence it, the economy plays an important role in sports consumption.

Low costs have come under the reasons for preferring sports consumption (Kim and Mao, 2021). The dramatic upswing in sports consumption in recent years has brought competition (Trail et al., 2003). The economic sustainability of sports and the increase in consumption have triggered more effective use of social media tools this is why the consumption of sports products and services is directly related to individuals who have easy access to media tools.

The internet is particularly used as a notable market and marketing tool in the realization of sports consumption. Therefore, it is a worthy part of sports consumption (Hur et al., 2007; Seo & Green, 2008; Kim & Trail, 2011). Smartphones (Chan-Olmsted & Xiao, 2019; Ha et al., 2017), participation in sports and culture (Mehus, 2005), income status (Thibaut et al., 2014), media (Koronios et al., 2020; Chan-Olmsted & Kwak, 2020) and environmental factors (Fink et al., 2002) seem to be outstanding components in determining the level of sports consumption with developing technology. Being a technology-enabled society today has made it easier for us to provide immediate and continuous access to sports. This has led individuals to be unable to stop themselves and to have an active role in sports consumption (Aiken et al., 2018). Factors such as inability to stop oneself, physical and psychological dependence, loss of control cause compulsive behavior (Ronald et al., 1987). Individuals who exhibit this behavior place a high value on the appearance of products (Trautmann-Attmann & Johnson, 2009). Factors influencing consumption, such as the convenience of online shopping and the ease of access to products, may also contribute to an increase in compulsive behaviors (Huang et al., 2022).

Different products or consumption requirements will help temporarily alleviate mental problems such as stress, apprehension, anxiety and depression.

The inability of individuals to prevent or stop them or to engage in uncontrolled sports consumption is called “compulsive sports consumption”. It is critical to make compulsive sports consumption measurable for researchers studying the sports industry and human behaviors interested in sports. The literature review revealed that some researchers attempted to measure sports consumption motivation (Cottingham et al., 2014; Seo & Green, 2008; Trail & James, 2001) whereas only Aiken et al. (2018) addressed compulsive sports consumption.

1.1. The Present Study

This is a scale adaptation study that arose from the need to investigate the compulsiveness of sport consumption, which affects a large number of people, including Turkish-speaking populations.

It is possible to use a measurement tool developed in one language in another, but translation alone is insufficient, and even considering it sufficient leads to serious scientific errors. To accomplish this, it is necessary to culturally adapt the scale and obtain evidence for the scale's validity and reliability through studies conducted with target culture samples. Furthermore, scale adaptation is a collaborative effort that necessitates the collaboration of experts in the field, psychometrists, and linguists.

There are some well-known sources in the literature that describe the scale adaptation processes (Hambleton & Patsula, 1999; Hambleton, Meranda & Spielberger, 2005). Taking these contexts into account, we followed the scale adaptation steps outlined below in our study.

- Deciding whether it is more useful to develop a new scale or adapt an existing scale.
- Requesting Permission for Adaptation.
- Choosing highly qualified translators.
- Translation and adaptation of the scale into the target language.
- Feedback application of the adapted version of the scale on a small group.
- Analyzing linguistic equivalence.
- Applying the scale to a larger group that can represent the target group and obtaining evidence of the scale's validity and reliability.
- Examining test-retest reliability.

The Compulsive Sport Consumption Scale (CSCS), developed in English by Aiken et al. (2018), was aimed to be adapted into Turkish with scientific accuracy by following the predetermined steps in this study.

2. METHOD

During the adaptation process, both theoretical and field studies were conducted with 12 experts and 521 participants. The participants were distributed as follows: nine in the small group application, 66 in the linguistic equivalence application, 409 in the large group application, and 37 in the test-retest reliability application.

2.1. Description of CSC Scale Original Form

The CSCS consists of six items and uses a 7-point Likert scale from *strongly disagree* to *strongly agree*. Higher levels of CSCS were linked to psychological and behavioral constructs such as past and current sport participation, as well as the consequences of sport consumption (i.e., time, money, coping, and psychological and behavioral neglect). CSCS is capable of classifying and distinguishing compulsive sport consumers from less compulsive sport consumers. The studies demonstrated that CSCS-identified compulsive sport consumers spent a disproportionate amount of time and money on sport and experienced more negative consequences as a result of their participation. Confirmatory factor analysis (CFA) was used to evaluate the six-item CSCS's unidimensionality. Results of the CFA indicated a good fit of the model to the data ($\mathcal{X}^2$/df = 2.71, CFI = .99, TLI = .99, SRMR = .04, RMSEA = .064). The results of the scale development study demonstrated that the one-dimensional CSCS has adequate reliability and internal consistency. Cronbach's alpha (= .94) was greater than .70 (Nunnally, 1978), the average variance extracted (AVE = .72) was greater than .50, and composite reliability (CR = .84) was greater than .60. (Fornell & Larcker, 1981). In terms of criterion validity, positive correlations were found between sport fan related construct dimensions and the CSCS, as expected. Higher levels of CSCS correspond to higher levels of identification, sporting event orientation, and obsessive and harmonious passion. The majority of correlations were moderately significant (Aiken et al., 2018).

CSCS contains the following items: (1) Much of my life centers around the consumption of sport., (2) I think about sport all the time., (3) I find it difficult to stop watching, reading, or talking about sport., (4) The urge to consume sport is strong. I can't help myself from doing this activity., (5) Consuming sport is something I cannot live without., (6) I am completely taken with sport consumption.

2.2. Deciding on the Adaptability of the Scale

After recognizing the need for a Turkish scale to measure compulsive sports consumption, we had to decide whether it would be more appropriate to develop a new scale or adapt an existing one. During our literature review, we came across a scale called CSCS, which was developed in English to measure this construct. The adaption of the English scale was accompanied by undeniable benefits, given our easy access to English linguists and our capacity to assess linguistic equivalence with individuals fluent in both languages. Furthermore, the aforementioned scale was developed accurately and in accordance with scientific processes. The scale has sufficient evidence of validity and reliability. Additionally, its small number of items makes it simple to use and apply. All these arguments were persuasive in favor of adapting this scale instead of developing a new one.

Some measurement tools may be inappropriate for cultural adaptation as the expressions in the scale items are not fully understood or perceived differently by respondents from the target culture. Before beginning the scale adaptation study, we conducted a process that included theoretical discussions on the adaptability of the scale into Turkish with a team of one psychometrist, one English and one Turkish linguist, and one sport scientist in order to avoid problems such as difficulty in understanding and structural differentiation caused by intercultural differences. At length of the process, it was agreed that the expressions in the scale items are not foreign to Turkish culture and that the scale will be comprehensible if the concept of sports consumption is explained in the scale instructions. The measured structure was expected to be validated in the Turkish sample, and it was decided that it could be adapted.

2.3. Requesting Permission for Adaptation

To avoid breaking any ethical rules, each of the three researchers who developed the scale was contacted via e-mail, and permission to adapt the scale was obtained.

2.4. Translation of Scale

A team of twelve experts was assembled to translate the scale, including eight English and two Turkish linguists, a sports scientist, and a psychometrist. Eight English linguists were divided into four two-person groups. In each group, one linguist translated the scale's original English form into Turkish (forward translation), and the other linguist translated the Turkish form back into English (back translation). Each group discussed the differences between the back-translated and original forms before finalizing the translation. As a result, four different Turkish forms were obtained from four different groups. Twelve experts then gathered to compare these four forms and reconcile some of their differences. At the end of the process, the translation was finalized, reaching a final form with unanimous agreement among all experts involved.

2.5. Small Group Application

The Turkish Form of the Compulsive Sport Consumption Scale (CSCS-T) was carried out face to face to 9 people aged 25-35. Participants were asked if they clearly understood the scale's items and instructions. All participants concur that all of the scale's expressions are comprehensible and that no correction is required.

2.6. Linguistic Equivalence Application

We administered the English and Turkish versions of the scale to 66 university students who were fluent in both languages, with a two-week interval between the two paper-pencil applications. The paired samples t-test was utilized to compare the means of total scores, and the Wilcoxon signed-rank test was utilized to compare the medians of item scores between the applications.

2.7. Large Group Application

Data were collected from 409 participants, 248 (60.6%) male and 161 (39.4%) female, ranging in age from 13 to 59 years. Additionally, 202 (49.4%) were university students studying at the faculty of sports sciences, 33 (8.1%) were university students studying in other departments, 132 (32.2%) individuals were from various professions, the majority of whom were teachers, and 42 (10.3%) were K-12 students. Furthermore, 235 (57.5%) of them declared that they actively participate in sports, while 174 (42.5%) did not. The scale was administered face-to-face to 118 university students and online to the remaining participants. The potential problems such as failure to complete the test were not encountered during the online application due to the fact that the scale consists of six items and can be completed in a matter of minutes

2.8. Analyzing Data from Large Group Application

A parallel analysis was performed to observe if it was also achieved in the target culture because of the fact that the scale's unidimensionality was established in the original culture. Parallel analysis method is utilized in exploratory factor analysis to determine the number of factors. Many researchers propose the parallel analysis since it provides more accurate results in many conditions than other methods, and it is also thought to be the best method for identfying the number of factors (Silverstein, 1987; Williams et al., 2010; Zwick & Velicer, 1986; Hayton et al., 2004). Polychoric correlation matrix was used for parallel analysis, and the optimal implementation procedure was used to determine the number of dimensions, with robust unweighted least squares (RULS) factor extraction method. The number of bootstrap samples was set at 500, the maximum number of iterations at 1000, and the convergence value was set at 0.00001. When the factorability of the items was examined, it was discovered that Bartlett's statistic = 2669.6 (df = 15; *p* = .000000) and the Kaiser-Meyer-Olkin (KMO) value was .924. The results indicate that the correlation matrix factorability was very good. In this regard, we continued the factor analysis and reported all of the other findings in the study's following sections. Additional evidences for unidimensionality were also investigated, including unidimensional congruence, explained common variance, mean of item residual absolute loadings, and robust fit statistics. Furthermore, Cronbach's Alpha coefficient for internal consistency and McDonald's Omega coefficient for composite reliability were calculated to demonstrate the reliability of the measurements taken with the adapted scale. And at last, the graded response model, one of the polythomous item response theory models, was used to estimate the discrimination and category difficulties parameters of the items, which were then reported along with the test information function. Exploratory factor analysis and parallel analysis were conducted using the FACTOR software (Version 12.04.02), while analyses based on item response theory were performed using the R ltm package (Rizopoulos, 2006).

2.9. Additional Application for Reliability

To obtain evidence for the stability of the scale scores, we applied the scale to 37 university students (Male=21, Female=16) with an interval of two weeks. The Pearson's product-moment correlation coefficient was employed to calculate the correlation between the scores obtained from the test-retest, and the paired samples *t*-test was utilized to ascertain the existence of statistically significant difference in the means of the total scores. Furthermore, the Wilcoxon signed-rank test was applied to compare the medians of item scores across the test-retest applications.

3. RESULTS

3.1. Linguistic Equivalence

The comparison of the means of total scores derived from the two applications, which aimed to examine linguistic equivalence, is presented in Table 1.

Table 1. *Results of the paired samples t-test for comparison of total scores (linguistic equivalence).*

Form	M	SD	*t*	*df*	*p*
English	18.33	8.611	-.960	65	.343
Turkish	19.77	9.600			

Upon examination of Table 1, it is evident that there exists no statistically significant difference between the means of total scores obtained from the applications of the English and Turkish versions of the scale. Additionally, the medians of item scores between the applications were compared, and the significance of the differences are provided in Table 2.

Table 2. *Results of the wilcoxon signed-rank test for comparison of item scores (linguistic equivalence).*

Item	Test Statistic	*SE*	*p*
1	770.000	109.007	.457
2	602.500	105.985	.568
3	542.500	90.892	.982
4	653.500	102.579	.876
5	834.500	121.008	.763
6	636.500	108.446	.628

As observed in Table 2, there is no statistically significant differentiation among the item scores of the two different language versions. Consequently, the accumulated evidence lends support to the successful establishment of linguistic equivalence.

3.2. Construct Validity

Table 3 shows the polychoric correlation matrix upon which the parallel analysis is based. All of the inter-item correlation coefficients were found to be positive and high.

Table 3. *Polychoric correlation matrix.*

	1	2	3	4	5	6
1	1					
2	.848	1				
3	.784	.850	1			
4	.739	.782	.801	1		
5	.782	.865	.837	.814	1	
6	.703	.768	.719	.763	.798	1

The original unidimensional structure of the scale was investigated to determine whether it was preserved in the target culture. The eigenvalues of the factors and their explained variances were shown in Table 4.

Table 4. *Explained variance based on eigenvalues.*

Factor	Eigenvalue	Proportion of Variance
1	4.955	.826
2	.334	.056
3	.253	.042
4	.200	.034
5	.147	.024
6	.112	.019

As demonstrated in the table, the eigenvalues of the first factor was 4.955, which accounts for approximately 83% of the variance. There was no other component with an eigenvalue greater than one, indicating that structure was unidimensional in the target culture. The number of dimensions advised by parallel analysis was one as well.

In addition, more evidence for the existence of a unidimensional structure was obtained and presented in Table 5.

Table 5. *Additional evidences for unidimensionality.*

	Value	95% Confidence Intervals	
		Lower Bound	Upper Bound
UniCo	.995	.989	.998
ECV	.937	.922	.960
MIREAL	.203	.161	.238

A UniCo (Unidimensional Congruence) value greater than .95 indicates that the data can be treated as essentially unidimensional. A value of ECV (Explained Common Variance) greater than .85 indicates that the data is essentially unidimensional. A MIREAL (Mean of Item Residual Absolute Loadings) value less than .30 suggests that the data can be treated as essentially unidimensional (Ferrando & Lorenzo-Seva, 2018). When the values and confidence intervals in Table 5 are compared to the criteria mentioned above, it is clear that strong evidence exists for the existence of unidimensional structure.

Furthermore, we examined robust fit statistics utilized in the exploratory factor analysis to assess the fit of the unidimensional structure with the data, and these findings have been recorded in Table 6.

Table 6. *Robust fit statistics.*

Index	RMSEA	NNFI	CFI	GFI	AGFI
Statistic	.042	.997	.998	.999	.999

RMSEA values of .05 or less are commonly considered to be indicative of good fit. Greater CFI values indicate that the target model fits the data better than the baseline, with values of .95 or higher typically used to identify models that fit the data well. Similarly, the closer the NNFI, CFI, and AGFI statistics are to 1.00, the better the fit (Hu & Bentler, 1999; Schermelleh-Engel et al., 2003; Finch, 2020). In this study, the unidimensional structure was found to fit the data well, as shown in Table 5. Besides, this one dimension's (factor) generalized H-Index was 0.963. The H index assesses how well a set of items represents a factor. H values greater than .80 indicate a well-defined latent variable that is more likely to be stable across studies (Hancock & Mueller, 2000). Furthermore, Table 7 shows the item-level assessments.

Table 7. *Item-level assessments.*

Item	Mean	MSA	Factor Loading
1	4.858	.940	.863
2	4.631	.936	.935
3	4.778	.919	.900
4	4.381	.893	.873
5	4.560	.929	.930
6	3.824	.931	.833

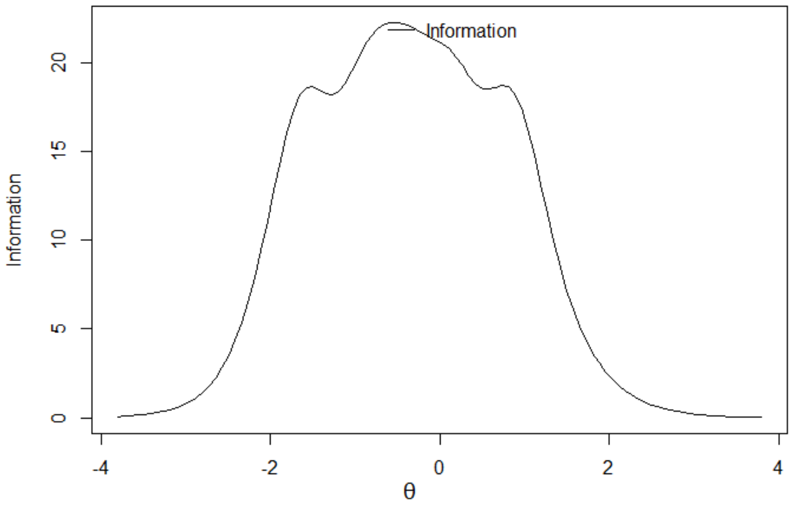
If the Measure of Sampling Adequacy (MSA) value is less than .50, it indicates that the item does not measure the same domain as the remaining items in the pool and should be removed (Lorenzo-Seva & Ferrando, 2021). Table 7 shows that all MSA values are greater than the criterion. Furthermore, factor loadings for all items were high. These were enthusiastic findings for item-level assessments. Table 8 also includes *a* and *b* parameters derived from item response theory parameterizations of the items.

Table 8. *Item response theory parameters.*

Item	*a*	*b1*	*b2*	*b3*	*b4*	*b5*	*b6*
1	2.864	-5.179	-2.896	-2.268	-1.336	-0.399	2.115
2	4.210	-6.893	-3.641	-2.800	-1.592	0.610	3.691
3	3.367	-5.995	-3.188	-2.450	-1.269	0.185	2.413
4	3.224	-5.270	-2.652	-1.636	-0.467	0.778	3.561
5	4.118	-6.333	-3.445	-2.268	-1.015	0.626	3.586
6	2.477	-3.618	-1.439	-0.606	0.410	1.733	3.859

The *a* parameters in Table 8 indicate item discrimination and the *b* parameters indicate category difficulties. According to the findings, all items had very high discrimination (Baker, 2001). Considering category difficulties, it is necessary to have a much higher level of compulsive sport consumption in order to agree and strongly agree with the sixth item. Figure 1 depicts the test information function, which displays how much information the scale explains at θ level.

Figure 1. *Test information function.*



In the test information function (Figure 1), the levels of individuals in the compulsive sport consumption trait (θ) range between -4 and 4. We discovered that obtaining CSCS-T measurements on people with θ levels between approximately -1.5 and 1 yields the most accurate results.

3.3. The Power of CSCS-T to Distinguish Between Groups

As another evidence of construct validity, we examined how effectively the CSCS-T distinguishes individuals with different levels of the compulsive sport consumption. We divided the participants into two groups: those who participate in active sports and those who do not, based on the assumption that those who participate in active sports consume more sports-related things. The mean scores of these two groups from the CSCS-T were compared using the independent samples *t*-test, and the results showed that the difference was statistically significant, as demonstrated in Table 9.

Table 9. *Results of the independent samples t-test for comparison of total scores.*

Group	M	*SD*	*t*	*df*	*p*
Active	30.09	10.307	-6.752	361.242	.000
Passive	22.90	10.885			

The *a* parameters in Table 8 indicate item discrimination and the *b* parameters indicate category difficulties. According to the findings, all items had very high discrimination (Baker, 2001). Considering category difficulties, it is necessary to have a much higher level of compulsive sport consumption in order to agree and strongly agree with the sixth item. Figure 1 depicts the test information function, which displays how much information the scale explains at which θ level.

3.4. Reliability

Cronbach's alpha was 0.958, which was used to determine the internal consistency of CSCS-T. Likewise, McDonald's Omega coefficient for composite reliability (CR) was calculated as 0.958. As a result of the test-retest application to determine the stability of the measurements, Pearson’s product-moment correlation coefficient was calculated as .923, indicating a strong positive correlation supporting the test-retest reliability. Besides this, the paired samples *t*-test found no statistically significant difference between test and retest mean scores (t=-1.205, p>0.05).

Additionally, the medians of item scores from both the test and retest applications were compared, and the significances of the differences are provided in Table 10.

Table 10. *Results of the wilcoxon signed-rank test for comparison of item scores (test-retest).*

Item	Test Statistic	*SE*	*p*
1	188.000	43.540	.730
2	193.500	45.999	.602
3	186.500	48.317	.341
4	194.500	48.399	.432
5	148.500	41.322	.327
6	190.500	50.773	.257

As depicted in Table 10, there exists no statistically significant differentiation among item scores obtained from the test and retest applications. As a result, the gathered body of evidence supports the establishment of reliability in terms of stability.

4. DISCUSSION and CONCLUSION

The purpose of the study is to adapt the Compulsive Sport Consumption Scale developed by Aiken et al. (2018) to Turkish culture. We began the process by discussing whether the structure measured by the CSC scale and the expressions used in the scale exist in Turkish culture, and thus, whether it is appropriate to adapt the scale to the target culture. After deciding to adapt the scale, we got permission from the scale's developers, and then, completed the translation process with a team of English and Turkish linguists, sports scientist, and psychometrist.

Given the unidimensional nature of the original scale version, our prediction was that the same unidimensionality would also hold true for the target culture. Nonetheless, as the structure doesn't strictly adhere to a rigid psychological theory even within its original cultural context, initiation was carried out through the utilization of exploratory factor analysis to unveil the representation of the structure within the target culture. Upon analysis, it was concluded that the scale had a unidimensional structure in the target culture as well, based on the eigenvalues. Confirmatory factor analysis was not pursued in this context, as the nonexistence of a complicated structure that included relations between scale items and multiple factors rendered such investigation unnecessary. The important issue, in this case, was the gathering of new evidence supporting the structure's unidimensionality. From this perspective, we performed parallel analysis and used many statistics such as unidimensional congruence, explained common variance, mean of item residual absolute loadings, and robust fit statistics. The scale was unidimensional, according to parallel analysis, and all other statistics supported the scale's unidimensionality. Moreover, regarding fit statistics, the scale’s unidimensional structure fits the study data well. Aiken et al. (2018) discovered that the six-item single-factor structure in the original form of scale explained 69% of the total variance. Unidimensional CSCS-T, on the other hand, explained approximately 83% of the total variance. All the item factor loadings in the adapted scale are greater than those in the original structure. On a side note, the generalized H-Index of this single factor was .963. The H index measures how well a set of items represents a factor, and this value indicated that CSCS-T would be highly stable across different studies.

To assess the reliability of the CSCS-T measurements, internal consistency, composite reliability, and test-retest reliability were examined. Cronbach's alpha and McDonald's Omega coefficients were both calculated as .958. Aiken et al. (2018) reported Cronbach's alpha coefficient as .94 and McDonald's Omega coefficient as .84. The measurements are reliable because these values are greater than .70 (Nunnally, 1978) for both the original and adapted forms. Additionally, the results of the test-retest applications indicated the stability of both the total and item scores across time.

All of the results prove that the CSCS-T (Appendix) can be used to obtain valid and reliable measurements when compulsively over-involvement in people's sports consumption is being investigated. It should be noted, however, that the study's inability to reach a larger sample size can be considered a limitation. The study's inability to reach a larger sample size can be counted as a limitation. This scale's target group is not restricted to a specific age or occupational group. It was not possible to reach all population subgroups in a single study. Further research can be conducted in this area, with additional studies to be conducted with groups of varying characteristics. It can be also suggested that a criterion validity study be conducted using scales measuring similar or distinct structures in Turkish.

Acknowledgments

We would like to express our gratitude to the English linguists: Prof. Dr. Turan Paker, Assist. Prof. Dr. Jasmin Esin Duraner Dikmen, Assist. Prof. Dr. Mehmet Erguvan, Assist. Prof. Dr. Olcay Şener, Tuncay Tezcan, Emre Kaya, Ayşe Ece Derelioğlu, Devrim Ulaş Arslan, and Ömer Faruk Akova. Additionally, our appreciation goes to the Turkish linguists: Assoc. Prof. Dr. Özden Ölmez Ceylan and Eflak Malgaca Vargel for their invaluable contribution to the translation process of the scale.

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**: Ardahan University, 16.02.2023, E-67796128-000-2300006274.

Authorship Contribution Statement

Murat Aygün: Investigation, Data Collection, Methodology, Visualization, and Writing the Introduction of the Manuscript. **Sait Çüm**: Methodology, Visualization, Data Analysis, Validation, and Discussion of the Results. All authors reviewed the results and approved the final version of the manuscript.

Orcid

Murat Aygün https://orcid.org/0000-0002-7636-8325
Sait Çüm https://orcid.org/0000-0002-0428-5088

REFERENCES

Aiken, K.D., Bee, C., & Walker, N. (2018). From passion to obsession: Development and validation of a scale to measure compulsive sport consumption. *Journal of Business Research*. *87*, 69-79. https://doi.org/10.1016/j.jbusres.2018.02.019

Armstrong, K.L., & Peretto Stratta, T.M. (2004). Market analyses of race and sport consumption. *Sport Marketing Quarterly*, *13*(1), 7-16.

Baker, F.B. (2001). *The basics of item response theory*. MD: ERIC Clearinghouse on Assessment and Evaluation.

Chan-Olmsted, S., & Kwak, D.H. (2020). Fantasy sport usage and multiplatform sport media consumption behaviors. *Sport Marketing Quarterly*, *29*(3), 204-214. http://doi.org/10.32731/SMQ.293.092020.04

Chan-Olmsted, S., & Xiao, M. (2019). Smart sports fans: Factors influencing sport consumption on smartphones. *Sport Marketing Quarterly*, *28*(4), 181-194. http://doi.org/10.32731/SMQ.284.122019.01

Cottingham, M., Phillips, D., Hall, S.A., Gearity, B.T., & Carroll, M.S. (2014). Application of the motivation scale for disability sport consumption: An examination of intended future consumption behavior of collegiate wheelchair basketball spectators. *Journal of Sport Behavior*, *37*(2), 117.

Faber, R.J., O'Guinn, T.C., & Krych, R. (1987). Compulsive consumption. *Advances in Consumer Research*, *14*, 132-135.

Finch, W.H. (2020). Using fit statistic differences to determine the optimal number of factors to retain in an exploratory factor analysis. *Educational and Psychological Measurement* , *80*(2), 217-241. https://doi.org/10.1177/0013164419865

Fink, J.S., Trail, G.T., & Anderson, D. (2002). Environmental factors associated with spectator attendance and sport consumption behavior: Gender and team differences. *Sport Marketing Quarterly*, *11*(1), 8-19.

Ha, J.P., Kang, S.J., & Kim, Y. (2017). Sport fans in a "smart sport" (SS) age: Drivers of smartphone use for sport consumption. International *Journal of Sports Marketing and Sponsorship*, *18*(3), 281-297. http://doi.org/110.1108/IJSMS-08-2017-093

Hancock, G.R., & Mueller, R.O. (2000*). Rethinking construct reliability within latent variable systems.* In R. Cudek, S.H.C. duToit & D.F. Sorbom (Eds.), Structural equation modeling: Present and future (pp. 195-216). Scientific Software.

Hansen, H., & Gauthier, R. (1989). Factors affecting attendance at Professional sport events. *Journal of Sport Management*, *3*(1), 15-32. https://doi.org/10.1123/jsm.3.1.15

Hayton, J.C., Allen, D.G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor Analysis: A tutorial on parallel analysis. *Organizational Research Methods*, *7*(2), 191-205. https://doi.org/10.1177/1094428104263675

Hopkinson, G.C., & Pujari, D. (1999). A factor analytic study of the sources of meaning in hedonic consumption. *European Journal of Marketing*, *33*(3/4), 273-290. https://doi.org/10.1108/03090569910253053

Hu, L.-I., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55. https://doi.org/10.1080/10705519909540118

Huang, H.L., Chen, Y.Y., & Sun, S.C. (2022). Conceptualizing the internet compulsive-buying tendency: What we know and need to know in the context of the covıd-19 pandemic. *Sustainability*, *14*, 1549. https://doi.org/10.3390/su14031549

Hur, Y., Ko, Y.J., & Valacich, J. (2007). Motivation and concerns for online sport consumption. *Journal of Sport Management*, *21*, 521-539. https://doi.org/10.1123/jsm.21.4.521

Jang, W., Wu, L., & Wen, J. (2021). Understanding the effects of different types of meaningful sports consumption on sports consumers' emotions, motivations, and behavioral intentions. *Sport Management Review*, *24*(1), 46-68. https://doi.org/10.1016/j.smr.2020.07.002

Kempf, D.S. (1999). Attitude formation from product trial: Distinct roles of cognition and affect for hedonic and functional products. *Psychology & Marketing*, *16*(1), 35–50. https://doi.org/10.1002/(SICI)1520-6793(199901)16:1<35::AID-MAR3>3.0.CO;2-U

Kim, M.J., & Mao, L.L. (2021) Sport consumers motivation for live attendance and mediated sports consumption: A qualitative analysis. *Sport in Society*, *24*(4), 515-533. https://doi.org/10.1080/17430437.2019.1679769

Kim, Y.K., & Trail, G. (2011). A Conceptual framework for understanding relationships between sport consumers and sport organizations: A relationship quality approach. *Journal of Sport Management*, *25*(1), 57-69. https://doi.org/10.1123/jsm.25.1.57

Koning, R.H. (2009). Sport and measurement of competition. *De Economist*, *157*, 229-249. https://doi.org/10.1007/s10645-009-9113-x

Koronios, K., Travlos, A., Douvis, J., & Papadopoulos, A. (2020). Sport, media and actual consumption behavior: An examination of spectator motives and constraints for sport media consumption. *EuroMed Journal of Business*, *15*(2), 151-166. https://doi.org/10.1108/EMJB-10-2019-0130

Lera-Lopez, F., & Rapun-Garate, M. (2007). The demand for sport: Sport consumption and participation models. *Journal of Sport Management*, *21*, 103-122. https://doi.org/10.1123/jsm.21.1.103

Lorenzo-Seva (2021). *SOLOMON: A method for splitting a sample into equivalent subsamples in factor analysis.* Behavior Research Method, in press.

Mehus, I. (2005). Distinction through sport consumption: Spectators of soccer, basketball, and ski-jumping. *International Review for the Sociology of Sport*, *40*(3), 321-333. http://doi.org//10.1177/1012690205060159

Nunnally, J.C. (1978). *Psychometric Theory*. McGraw-Hill.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1-25. https://doi.org/10.18637/jss.v017.i05

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*, *8*(2), 23-74.

Seo, W.J., & Green, B.C. (2008). Development of the motivation scale for sport online consumption. *Journal of Sport Management*, *22*, 82-109. https://doi.org/10.1123/jsm.22.1.82

Silverstein, A.B. (1987). Note on the parallel analysis criterion for determining the number of common factors or principal components. *Psychological Reports*, *61*(2), 351-354. https://doi.org/10.2466/pr0.1987.61.2.351

Thibaut, E., Vos, S., & Scheerder, J. (2014). Hurdles for sports consumption? The determining factors of household sports expenditures. *Sport Management Review*, *17*, 444-454. http://doi.org/10.1016/j.smr.2013.12.001

Trail, G., Fink, J.S., & Anderson, D.F. (2003). Sport spectator consumption behavior. *Sport Marketing Quarterly*, *12*(1), 8-17.

Trautmann-Attmann, J., & Johnson, T.W. (2009). Compulsive consumption behaviours: investigating relationships among binge eating, compulsive clothing buying and fashion orientation. *International Journal of Consumer Studies*, *33*(3), 267-273. https://doi.org/10.1111/j.1470-6431.2009.00741.x

Wheaton, B. (2000). Just do it: Consumption, commitment, and identity in the windsurfing subculture. *Sociology of Sport Journal*, *17*(3), 254-274. https://doi.org/10.1123/ssj.17.3.254

Williams, B., Onsman, A., & Brown, T. (2010). Exploratory factor analysis: A five-step guide for novices. *Australasian Journal of Paramedicine*, *8*(3), 1-13. https://doi.org/10.33151/ajp.8.3.93

Yoshida, M., & Nakazawa, M. (2016) Innovative sport consumption experience: An Empirical test in spectator and participant sports. *Journal of Applied Sport Management, 8*(1), 1-21. https://doi.org/10.18666/JASM-2016-V8-I1-6024

Zwick, W.R., & Velicer, W.F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*(3), 432-442. https://doi.org/10.1037/0033-2909.99.3.432

APPENDIX: Compulsive Sport Consumption Scale – Turkish version

Kompulsif Spor Tüketimi Ölçeği

Sayın Yanıtlayıcı,

Ölçeği yanıtlamadan önce lütfen bu yönergeyi dikkatle okuyunuz.

Bu ölçek spor tüketimi ile ilgili davranışlarınızı belirlemek amacıyla hazırlanmıştır. Spor ürünlerini satın almak, kiralamak, spor faaliyetlerini izlemek, dinlemek ve spor gıdalarının tüketimini de kapsayan spor ile ilgili birçok davranış **spor tüketimi** olarak ifade edilmektedir.

Ölçekte yanıtlaması yaklaşık beş dakika sürecek altı madde bulunmaktadır. Adınızı yazmanıza veya kimliğinizi belirtecek herhangi bir ifade eklemenize gerek yoktur. Yanıtlarınız hiçbir kişi veya kurumla paylaşılmayacak yalnızca bilimsel araştırma amacıyla kullanılacaktır. Maddeleri okuduktan sonra içtenlikle aklınıza ilk gelen seçeneği işaretleyiniz ve lütfen yanıtsız madde bırakmayınız.

Katkılarınız için teşekkür ederiz.

Aşağıdaki ifadelere ne derece katılıp/katılmadığınızı belirtiniz.	Kesinlikle Katılmıyorum	Katılmıyorum	Zannederim Katılmıyorum	Kararsızım	Zannederim Katılıyorum	Katılıyorum	Kesinlikle Katılıyorum
1. Spor tüketimi hayatımda çok önemli bir yer tutar.							
2. Spor aklımdan hiç çıkmaz.							
3. Sporla ilgili bir şeyler izlemekten, okumaktan veya konuşmaktan kendimi alıkoyamıyorum.							
4. İçimde spora dair şeyler tüketmeye yönelik çok güçlü bir istek var ve buna engel olamıyorum.							
5. Hayatımda spor ve spora dair şeyler olmadan yaşayamam.							
6. Kendimi tamamen spor tüketimine kaptırmış durumdayım.							

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 3, 496–506

https://doi.org/10.21449/ijate.1278728

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

A bibliometric journey into research trends in curriculum field: Analysis of two journals

Suat Kaya [1,*]

[1]Ağrı İbrahim Çeçen University, Faculty of Education, Department of Educational Sciences, Ağrı, Türkiye

ARTICLE HISTORY

Received: Apr. 7, 2023
Revised: Aug. 18, 2023
Accepted: Aug. 24, 2023

Keywords:
Bibliometric analysis,
Co-occurrence analysis,
Curriculum studies,
Democratic curriculum reform.

Abstract: The field of curriculum is characterized by its porous and evolving boundaries, which are influenced by ongoing shifts in sociological, technological, scientific, and political domains. Given this dynamic context, the field necessitates continuous advancements to address these evolving trends. Consequently, its scope and prevailing research foci are subject to change, thereby shaping curricular adaptations. The primary objective of this study is to delineate the contemporary scope of curriculum studies by examining prevailing topics of discourse. Research articles published in two selected journals—Curriculum Inquiry and Journal of Curriculum Studies—were analyzed to achieve this. These journals were chosen for their alignment with the study's objective and were employed as primary data sources. A bibliometric analysis was conducted on data harvested from these publications, utilizing descriptive statistics through the Web of Science (WoS) system as an initial analytical step. Subsequently, VOSviewer software was employed for advanced bibliometric analyses. The study's findings offer both visual and descriptive insights into how the thematic focus within curriculum studies has shifted over time. Notably, recent discussions within the field underscore the exigency for democratic curriculum reforms. Moreover, the issues addressed by the selected journals closely align with current societal challenges.

1. INTRODUCTION

The relentless advancements in technology, science, and communication necessitate an urgent global recalibration of educational paradigms for nations. This imperative arises from the recognition that formal education remains the most productive conduit for disseminating knowledge and skills that can mitigate societal stagnation. Unlike incidental, informal education which occurs ubiquitously in social interactions (Dewey, 2004), formal education is often institutionalized in schools. In these settings, curricula serve as the operative mechanisms for instructional delivery (Oliva, 1997). Therefore, the dynamism of curricula becomes a pivotal factor in shaping and advancing societal progress (Kaya, 2018).

“The education system is a social institution which should be expected to change along with other institutions. It would be more surprising, not to say disturbing, if the education system were to stand still while all else changed” (Kelly, 2004, p.1). In other words, education “does not possess a reality apart from the time, place, and mores in which it exists” (Ornstein &

*CONTACT: Suat KAYA ✉ sukaya@agri.edu.tr Ağrı İbrahim Çeçen University, Faculty of Education, Department of Educational Sciences, Ağrı, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

Hunkins, 2004, p. 133), so “it is important to continuously reappraise and revise existing curricula” (Ornstein & Hunkins, 2004, p. 150).

Many scholars and professionals in the world of education such as Dwayne Heubner has “ascribed ambiguity and a lack of precision to the term curriculum, therefore as highlighted by Elizabeth Vallance, “the curriculum field is by no means clear; as a discipline of study and as a field of practice, curriculum lacks clean boundaries” (as cited in Oliva, 1997). While this fluidity enriches the curriculum landscape, it simultaneously poses challenges for researchers seeking to precisely delineate its scope. The singular certainty regarding curriculum studies is its pressing need for constant revision to accommodate emergent global trends. The primary objective of this research is to scrutinize contemporary topics within curriculum studies with the aim of defining its evolving scope. A quintessential approach to conceptualizing a field of study involves systematically examining related scholarly output, as each discipline is responsible for periodically reassessing its contributions (Staton-Spicer & Wulff, 1984). Echoing the assertions by Cohen, Manion, and Morrison (2007), analyses of research within a given discipline provide invaluable insights for aspiring scholars in that field. Moreover, understanding the current landscape and prevailing research trends offers distinct advantages for scholars, not only in guiding their research trajectories but also in enhancing their academic publishing endeavors (Lee et al., 2009). Studies that map out these research trends effectively serve as pivotal benchmarks for future scholarly undertakings within the field (Chang et al., 2010).

2. METHOD

There are many ways such as literature review, content analysis, meta-analysis, meta-synthesis etc. to analyze the research trends in a field. These analysis methods can include a limited quantity of research studies, so bibliometric analysis was utilized as it can be used to analyze huge numbers of research studies conducted in a field (Zupic & Cater, 2015). It can be used to find out and understand the relationships between studies (Zupic & Cater, 2015); the trends, status, and possible gaps in a particular field (Romanelli et al., 2018); and the content of a particular domain (Fahimnia et al., 2015; Hallinger & Suriyankietkaew, 2018). Bibliometric studies also help journal editors review past publications, devise new policies, and make decisions (Zupic & Cater, 2015).

2.1. Data Collection

There are two main approaches while preparing data set in bibliometric analysis: searching by using selected keywords or phrases and then identifying studies on detailed readings, which is generally used in studies that focus on a specific subject, while the second approach is to select one or more journals and include all the studies published here or the studies determined as a result of the examinations in the analysis (Zupic & Cater, 2015). The second approach was adopted in this study by selecting two journals publishing research about education and curriculum field.

As shown in Table 1, the selection process started with the analysis of journals relevant to the “curriculum” keyword in the master journal list in WOS database, which is “the most common source of bibliographic data” (Zupic & Cater, 2015, p.14). The search was refined to only the journals indexed in Social Sciences Citation Index. After analyzing their aims and scopes, two out of six journals were selected: Curriculum Inquiry [CI] and Journal of Curriculum Studies [JCS] as they focused on general issues in education related to the curriculum field rather than a specific topic included by other journals such as “Language, culture and Curriculum” or “Medical Education”. The main aim of selected journals, on the other hand, was to publish research dealing with contemporary issues, problems, topics and trends in education specifically related to the curriculum field (CI, 2023; JCS, 2023). Both journals are published by Taylor & Francis, while The Ontario Institute for Studies in Education, in Canada collaborates with Taylor & Francis for publishing CI.

Table 1. *Criteria for selection process of the journals and publications.*

Criteria	Value
1. Data Source	1. WOS Database
2. Search Terms	2. "Curriculum"
3. Selected Journals	3. *Curriculum Inquiry* and *Journal of Curriculum Studies*
4. Citation Index	4. SSCI
5. Document Type	5. Articles and Review Articles
6. Excluded Documents	6. Correction, Addition, Letter, Proceeding Papers, Discussion, Bibliographical-Item, Item about an individual and Note
7. Number of Articles	7. 2484 (CI:895; JCS:1589)

The Web of Science (WoS) Core Collection database was accessed upon selecting the target journals. The initial search query consisted of the Boolean expression "Curriculum Inquiry" AND "Journal of Curriculum Studies" specified within the "Publication Title" field. This preliminary search yielded a corpus of 3,901 documents. Articles published in the year 2023 were subsequently omitted, given that the year was not yet complete, to ensure data validity. After that, additional filtering was conducted to exclude specific document types, namely "Correction," "Addition," "Letter," "Proceeding Paper," "Discussion," "Biographical-Item," "Item About an Individual," and "Note." Following these refinements, a final dataset comprising 2,484 articles, spanning the years 1998 to 2022, remained available for analysis.

As a matter fact, two journals were analyzed individually first, but the analysis resulted in similar topics leaving no room to discuss the field much. When the two were combined; however, the analysis resulted in a vivid journey of curriculum field as portrayed in the discussion part.

2.1. Data Analysis

Data pertaining to the temporal distribution, geographic origin, contributing authors, and affiliating institutions of studies published in the selected journals were subject to descriptive statistical analysis via the Web of Science (WoS) platform. Subsequently, bibliometric evaluation was conducted using VOSviewer software. Among various bibliometric analysis methods—such as citation analysis, co-citation analysis, bibliographic coupling analysis, and co-author analysis—co-occurrence analysis was specifically chosen in alignment with the study's objective: to scrutinize contemporary topics within the curriculum field with the intent to delineate its scope. Co-occurrence analysis involves linking keywords that appear concurrently in a document's title, abstract, or keyword list (Zupic & Cater, 2015). This method was employed to identify thematic clusters, emerging trends, and salient topics relevant to the curriculum field. The underlying rationale for utilizing co-occurrence, or co-word analysis, is the presupposition that frequent co-occurrence of terms within a corpus implies thematic or conceptual relatedness (Zupic & Cater, 2015). In summary, this refined bibliometric methodology aimed to answer the following research question:

- What are the prevailing trends and topics in the field of curriculum studies?

3. FINDINGS

3.1. Descriptive Findings

Figure 1 outlines the annual distribution of articles published in the selected journals. The data reveal that the inaugural year, 1998, saw the publication of over 40 articles, establishing a foundational volume of work. Subsequent observations confirm that the annual count of published articles has consistently remained above this initial threshold of 40. Additionally, the figure indicates periodic fluctuations in the annual publication rate, culminating in a zenith in the year 2019. Post-2019, however, the data exhibit a discernible downward trend in the number of articles published in these academic outlets.

Figure 1. *Distribution of publications by year.*

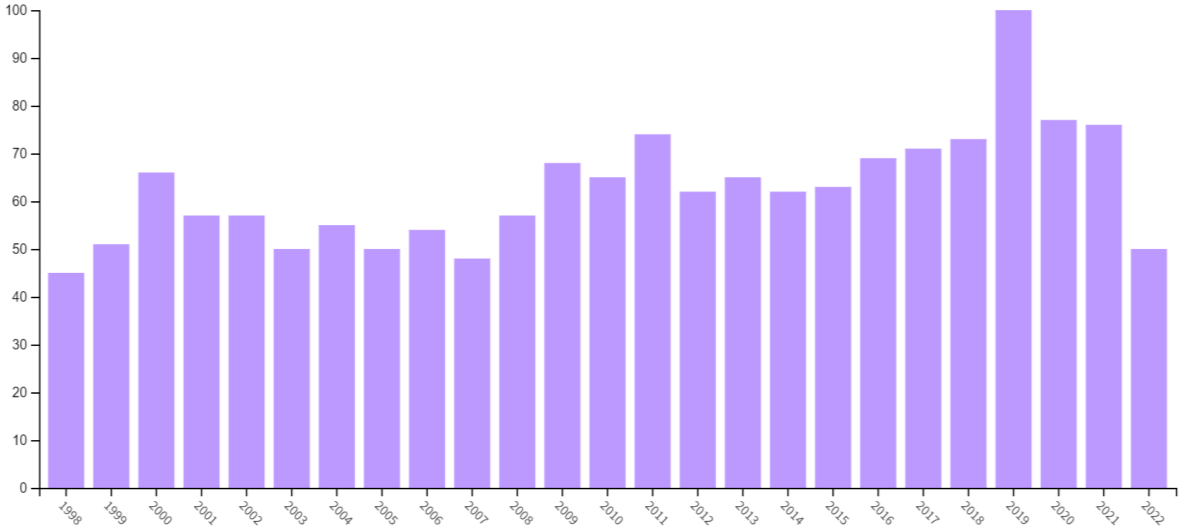


Figure 2 presents the distribution of papers published by countries. As seen, USA has been the most productive country dealing with issues touched upon by these journals. Almost half of the papers belong to USA. The other finding points to contributions from Canada and some countries in Europe and Asia. Still, it is not possible to talk about a global contribution.

Figure 2. *Distribution of publications by countries.*

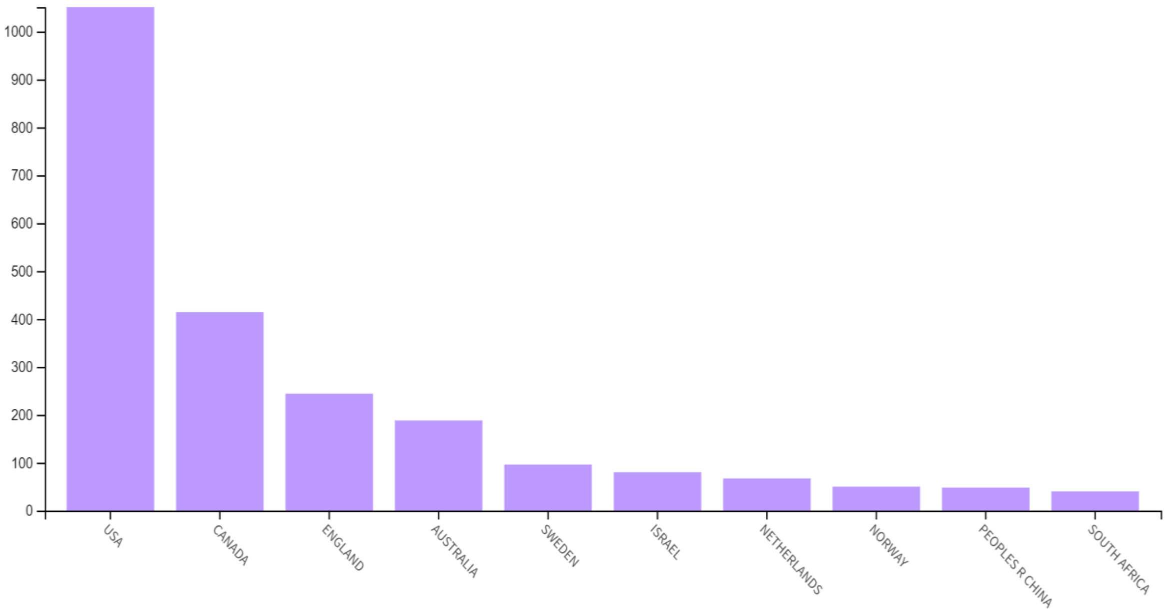


Figure 3. *The most productive authors.*

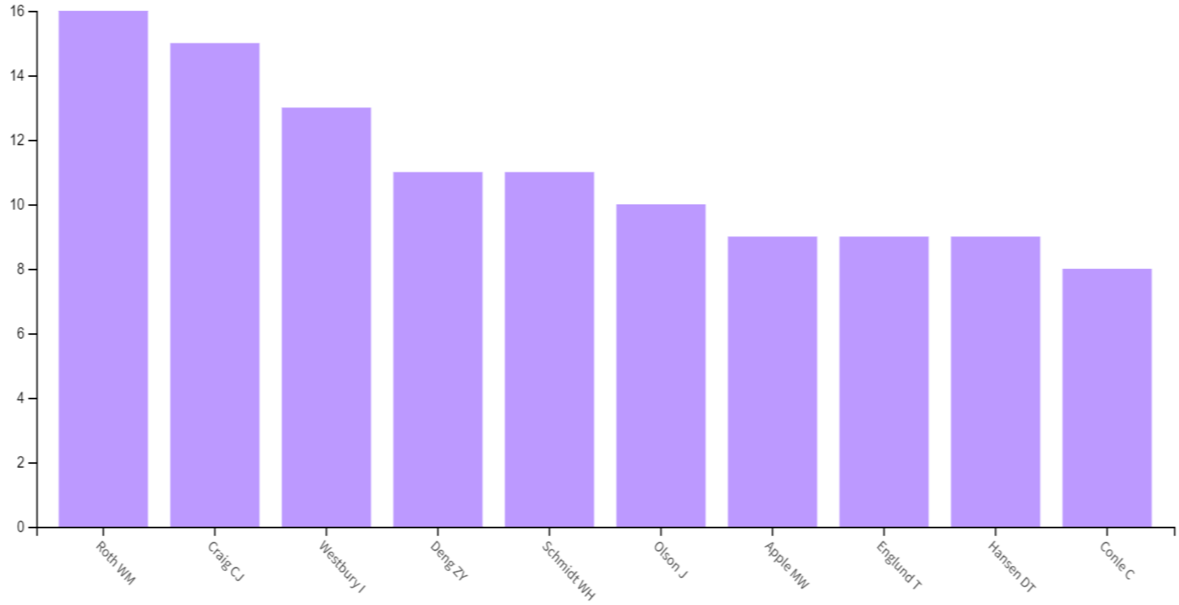
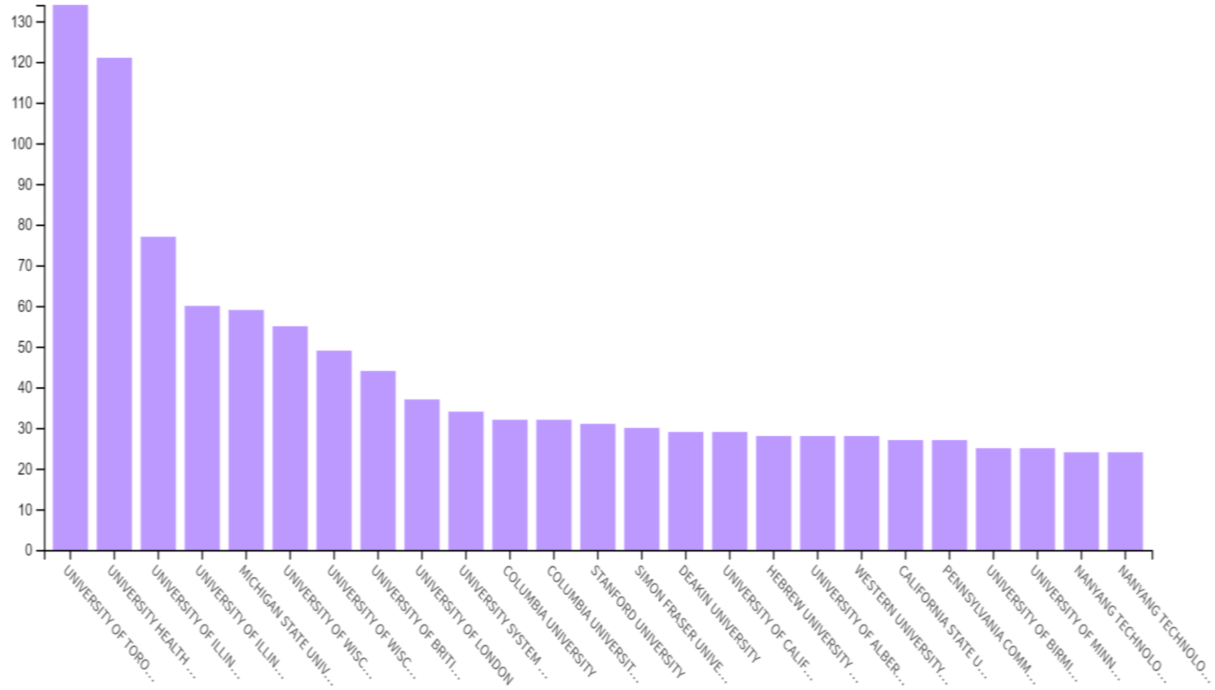


Figure 3 and 4 present findings on the most productive authors and institutions contributing to these journals. As seen in Figure 3, the most productive author was V.M. Roth, while the most productive institution was University of Toronto. It is possible to talk about contributions from various institutions, most located in USA.

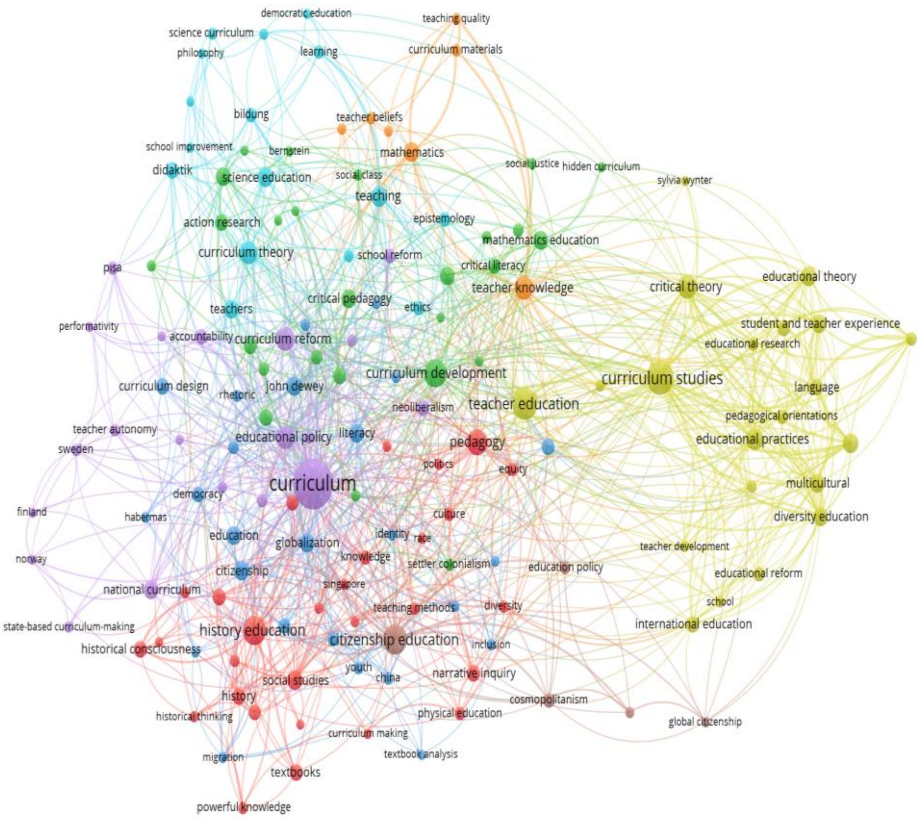
Figure 4. *The most productive institutions.*



3.2. Research Trends and Current Topics in Curriculum Field

Figure 5 presents the keywords used by the papers published in these journals. The minimum occurrence of the words was set to 5. The most noticeable finding as seen in the figure is that the most frequently used keywords look bigger than the less frequently used ones. The figure shows 9 clusters (red, blue, orange, brown, yellow, green, purple, turquoise and red). These clusters mean that these words are interrelated. The occurrence of these related words and concepts in these clusters is presented in Table 2.

Figure 5. *Co-occurrence of keywords.*



As seen in Figure 5 and Table 2, the terms most prevalently appearing across the examined papers include "curriculum" with a frequency of 107 occurrences, followed by "curriculum studies" (*f*=49), "teacher education" (*f*=42), "citizenship education" (*f*=36), "history education" (*f*=34), "curriculum development" (*f*=33), and "pedagogy" (*f*=27), among others. These findings suggest a semantic alignment with core issues in the field of curriculum studies.

Table 2. *Clusters of the words in publications.*

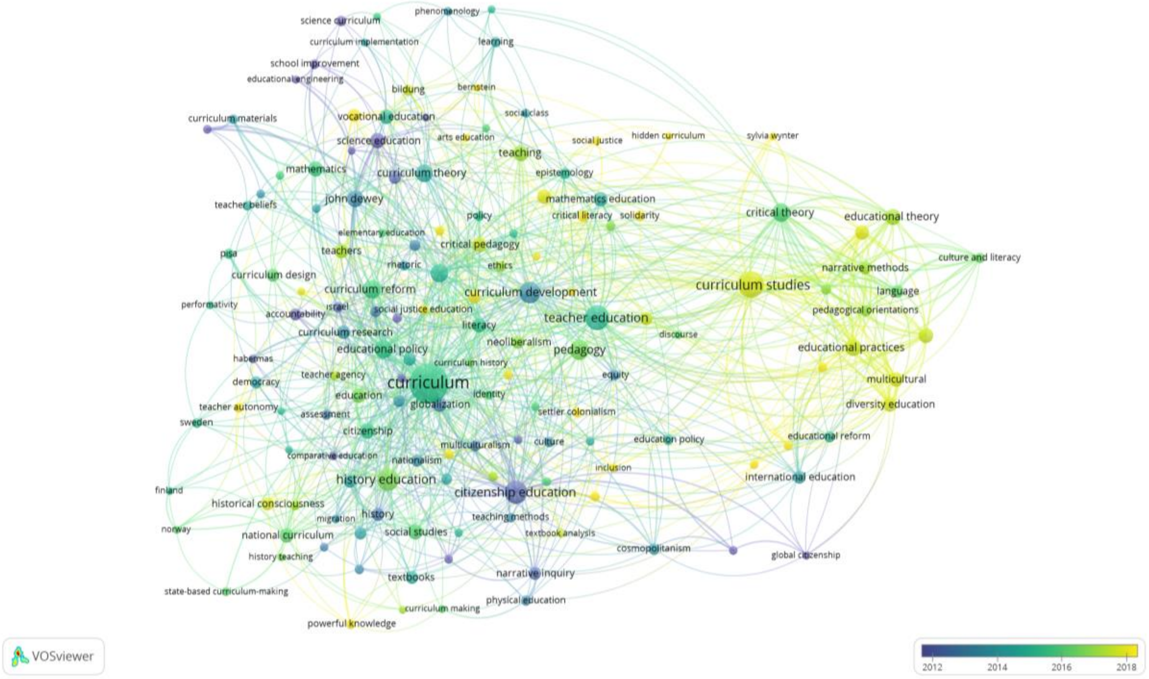
Clusters	Words (occurrence [*f*])
1st Cluster (Green)	Action research (11), agency (7), black feminism (5), critical literacy (10), critical pedagogy (14), curriculum change (12), curriculum development (33), curriculum research (13), ethnography (5), environmental education (9), hermeneutics (6), hidden curriculum (5), higher education (14), mathematics education (15), secondary education (12), settler colonialism (7), social justice (6), social justice education (8), social studies education (6), solidarity (8), student participation (5), teacher education (42), teaching methods (7), vocational education (14).
2nd Cluster (Purple)	Assessment (8), curriculum (107), accountability (10), Canada (6), comparative education (6), critical discourse analysis (10), curriculum reform (24), educational policy (23), Finland (5), national curriculum (15), history of education (6), neoliberalism (11), Norway (5), performativity (5), PISA (8), school reform (9), state-based-curriculum making (6), Sweden (7), teacher agency (7), teacher autonomy (7), teacher education curriculum (6), teacher professionalism (5).
3rd Cluster (Red)	Curriculum design (11), citizenship (16), conflict (5), democracy (10), discourse analysis (6), globalization (14), historical consciousness (12), historical thinking (7), history curriculum (10), history (11), history education (34), history teaching (5), history instruction (12), migration (6), powerful knowledge (7), secondary school curriculum (7), social studies (5), south Africa (6), textbooks (11), youth (7)

Table 2. *Continues.*

4th Cluster (Blue)	Actor-network theory (6), arts education (5), Bernstein (5), China (7), civic education (7), early childhood education (11), education (14), elementary education (6), funds of knowledge (6), identity (7), inclusion (6), Israel (8), literacy (15), moral education (6), multiculturalism (9), nationalism (10), recontextualism (5), rhetoric (7), social class (6), textbook analysis (5).
5th Cluster (Turquoise)	Bildung (9), curriculum implementation (5), curriculum theory (22), democratic education (5), didactic (5), educational change (5), educational engineering (5), epistemology (9), ethics (9), John Dewey (20), learning (8), phenomenology (6), philosophy (5), policy (7), school improvement (7), science education (17), science curriculum (8), teachers (15), teaching (18).
6th Cluster (Yellow)	Critical theory (24), culture and literacy (8), curriculum studies (49), diversity education (14), educational practices (19), educational reform (9), educational research (8), educational theory (18), gender issues in education (7), international education (11), language (11), multicultural (14), narrative methods (12), pedagogical orientations (9), school (6), socio-political conditions (17), student and teacher experiences (15), Sylvia Wynter (5).
7th Cluster (Red)	Culture (8), curriculum history (6), curriculum making (5), discourse (6), diversity (7), equity (7), knowledge (9), multicultural education (6), narrative inquiry (12), pedagogy (27), physical education (8), politics (6), race (6), Singapore (6), teacher development (6).
8th Cluster (Orange)	Curriculum materials (7), mathematics (16), mathematics curriculum (6), professional development (6), teacher beliefs (7), teacher knowledge (24), teaching quality (6)
9th Cluster (Brown)	Citizenship education (36), cosmopolitanism (9), education policy (8), European citizenship (6), global citizenship (5).

Figure 6 offers a temporal visualization of shifting research foci. Circa 2012, scholarly output predominantly centered on the theme of “citizenship education,” incorporating sub-topics such as “global citizenship,” “European citizenship,” and “globalization.” Subsequent focus transitioned towards “curriculum development” around 2014. The ensuing period, circa 2016, witnessed an emergent interest in themes including “teacher education,” “curriculum theory,” and specific analyses of “national curricula in Nordic countries.” Most recently, the prevailing research trends around 2018 have emphasized issues like “critical pedagogy,” “diversity education,” and “multicultural education,” collectively underscoring the imperative for democratic inclusivity within the curriculum.

Figure 6. *Co-occurrence of keywords between 2012-2018.*



4. DISCUSSION and CONCLUSION

Nothing is stable in the world, and everything is prone to change. In this respect, knowledge about any field, including curriculum, will always be open to change and challenge. Based on the hot topics discussed in the world of education, the scope of the curriculum field is expected to be upgraded to include these issues. As a matter of fact, curriculum as a field can be defined by dynamism in terms of its scope and focus which tend or are expected to change in line with specific changes brought about by time conditions. The findings of this bibliometric study managed to depict and visualize these changes over time, which can be called the journey of the curriculum field. This journey is discussed after a discussion of some descriptive findings below.

The descriptive findings indicated a decrease in the number of publications in these journals after 2019. This decrease in number might be attributed to the COVID-19 pandemic. As the COVID-19 pandemic, which "started in China in late 2019 and spread to all around the world" (Kaya, 2021, p. 302) shut the door on face-to-face education (Kaya, 2023), "most of the educational institutions were obliged to continue their education through online learning" (Kaya, 2021, p. 302). As a result, online learning has become the main research topic worldwide, which might be a reason for this decrease.

As promised, this study aimed to visualize the journey of the curriculum field over time. Time to discuss these findings now. As the research included in these journals highlighted as well, the focus of curricular studies at the beginning of the 21st century was on curricula of some nations. Especially, Nordic countries in Northern Europe such as Finland, Norway, and Sweden, and their curricula became the focus of curricular research due to their success in PISA (The Programme for International Student Assessment. The first success of the Finnish in PISA was in 2000, which was "greeted with surprise and disbelief" (Malinen et al., 2012) and identified as a "miracle" (Simola, 2005). After repetition of success in the subsequent exams; however, this success drew attention from many countries, resulting in a more detailed look at the Finnish education system, especially the Finnish Core Curriculum (Kaya, 2022). Research dealing with this issue has been included in the selected journals as well, because one of their aims was to publish contemporary issues concerning education and curriculum.

The evolving scholarly landscape has evidenced a marked pivot towards socio-political imperatives in the domains of education and curriculum studies. Notably, the thematic nucleus has coalesced around issues of inequality, encompassing multifaceted topics such as multiculturalism, feminism, black feminism, gender considerations, and diversity education. This thematic focus aligns conspicuously with the tenets of critical pedagogy, which advocates for dismantling oppressive societal structures through democratic pedagogical practices (Darder et al., 2003). Concomitant with increased global mobility and cross-border exchanges, nations have become increasingly heterogeneous, thereby necessitating curricular adaptations to cultivate national unity across ethnic, linguistic, and religious diversities. In this context, multicultural education emerges as a pragmatic instrument to achieve myriad objectives—from promoting diversity and equality to fostering mutual respect and facilitating optimal academic outcomes for all demographic groups (Levinson, 2007). Moreover, the extant literature reveals the subliminal existence of a 'hidden curriculum,' which tacitly indoctrinates students into conforming to pre-established hierarchies and power structures, including gender and economic hegemonies. Further converging with themes pertinent to critical pedagogy and critical theory, discussions related to the oppressive facets of colonialism and the instruction of history have also been underscored (McLaren, 2001). These thematic preoccupations elucidate the increasing adoption of discourse analysis as a methodological approach in these studies, possibly aiming to explicate societal mechanisms underpinning inequality. Moreover, multiple references to the pedagogical theories propounded by English sociologist Basil Bernstein—

centering on social struggle, symbolic control, and forms of power—further crystallize the thematic focus of the extant research corpus.

These concepts are also in line with the concept of Bildung by Wilhelm von Humboldt, which suggests the development of freedom and humanity in humans regardless of their status or class belonging through the teaching of content and the learning process. Humboldt defines the state within the limits that will not prevent and, on the contrary, protect the freedom that the individual needs in the process of shaping himself, because the original shaping of the individual and, therefore, the society depends on the absence of any external guiding intervention (Hotam 2019). In this sense, selection of content is of great importance. Rather than imposing one reality or one aspect of a specific content or knowledge, the individual should be allowed to create his/her own meaning out of various aspects of knowledge/content.

An additional salient observation warranting discussion pertains to the geographical distribution of contributions across countries, institutions, and authors within these journals. The data suggests a localized rather than global contribution. It is well-documented that migration trends have been accelerating, particularly toward economically developed nations such as the United States, Canada, and the United Kingdom, thereby leading to increasingly diverse and multicultural societies. These demographic shifts often intensify extant societal tensions, as evidenced by enduring racial dichotomies in these countries. Given that academic research aims to address pressing societal issues, the predominance of contributions from these nations in the journals under study could be interpreted as a response to such challenges. Another plausible explanation for this geographical concentration may reside in the location of the journals' publishers. Given these observations, it is incumbent upon journal editors to broaden their solicitation for contributions. Actively encouraging submissions from diverse geographical locations could enrich the global dataset pertaining to curriculum studies, thereby facilitating a more nuanced understanding through comparative analyses.

In summary, the thematic coherence among the studies published in these journals is indicative of an overarching consensus calling for comprehensive curricular reforms. The field of curriculum studies cannot afford to be indifferent to pressing educational challenges; rather, it bears the responsibility to acknowledge, interrogate, and articulate solutions to these issues. The exigencies of the present context compel the field to both engage proactively and respond critically. These challenges inherently fall under the purview of educational concerns and necessitate timely curricular adaptations to ameliorate them. Put succinctly, the extant research accentuates the emancipatory potential of education, achievable predominantly through curricular innovations. This emancipatory ethos echoes the democratic principles advanced by John Dewey and signals a call for democratic curriculum reform. Furthermore, it is worth noting that curriculum studies, a field rooted primarily in the 20th century, is undergoing an expansive metamorphosis. The field appears to be extending its disciplinary boundaries to encompass increasingly humanistic topics, thereby challenging its own traditional confines and aspiring toward a more inclusive, borderless scholarly landscape.

These findings are limited to data gathered from two journals, so further research can be conducted to include journals with similar aims and scopes in order to compare and contrast these findings and ultimately further define the scope of the curriculum field. In addition, most of these concepts and issues call for independent meta-studies to highlight the specifics inherent in them.

Acknowledgments

The author would like to thank the Editor and blind reviewers for their invaluable contributions.

Declaration of Conflicting Interests and Ethics

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

Orcid

Suat KAYA https://orcid.org/0000-0001-6593-3205

REFERENCES

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

Chang, Y.H., Chang, C.Y., & Tseng, Y.H. (2010). Trends of science education research: an automatic content analysis. *Journal of Science Education and Technology*, *19*(4), 315–331. https://doi.org/10.1007/s10956-009-9202-2

CI (Curriculum Inquiry). (2023). Available online: https://www.tandfonline.com/toc/rcui20/current

Darder, A., Baltodano, M. & Torres, R.D. (2003). *The Critical Pedagogy Reader.* Routledge.

Dewey. J. (2004). *Democracy and education: An Introduction to the philosophy of education.* Indian Edition. Aakar Books.

Fahimnia, B., Sarkis, J., & Davarzani, H. (2015). Green supply chain management: A review and bibliometric analysis. *International Journal of Production Economics*, *162*, 101-114. https://doi.org/10.1016/j.ijpe.2015.01.003

Hallinger, P., & Suriyankietkaew, S. (2018). Science mapping of the knowledge base on sustainable leadership, 1990-2018. *Sustainability*, *10*(12), 1-22. https://doi.org/10.3390/su10124846

Hinton, S. (2011). Ethnic diversity, national unity and multicultural education in China. *US-China Education Review B, 1*(5), 726-739.

Hotam, Y. (2019). Bildung: Liberal education and its devout origins. *Journal of Philosophy of Education*, *53* (4), 619-632. https://doi.org/10.1111/1467-9752.12386

JCS (Journal of Curriculum Studies). (2023). Available online: https://www.tandfonline.com/toc/tcus20/current.

Kelly, A.V. (2004). The *Curriculum theory and practice* (5th. Ed). Sage Publications.

Kaya, S. (2018). *Evaluation of the middle school English language curriculum developed in 2012 utilizing Stake's countenance evaluation model* [Unpublished doctoral dissertation]. Middle East Technical University.

Kaya, S. (2021). The factors predicting students' participation in online English courses. *Eurasian Journal of Educational Research 91, 301-320.* https://doi.org/10.14689/ejer.2021.91.14

Kaya, S. (2022). Analysis of Finnish core curriculum in relation to curriculum theories (p.49-67). (Ed. Ş. Durukan). In *Education & Science 2022-IV*. Efe Academy.

Kaya, S. (2023). Why do pre-service teachers prefer face to face, online or hybrid education?. *International Journal of Psychology and Educational Studies*, *10*(2), 379-392. https://dx.doi.org/10.52380/ijpes.2023.10.2.1024

Lee, M.H., Wu, Y.T., & Tsai, C.C. (2009). Research trends in science education from 2003 to 2007: A content analysis of publications in selected journals. *International Journal of Science Education, 31* (15), 1999-2020. https://doi.org/10.1080/09500690802314876

Levinson, M. (2007). Common schools and multicultural education. *Journal of Philosophy of Education*, *41*(4), 625–642. https://doi.org/10.1111/j.1467-9752.2007.00587.x

Malinen, O., Vaisanen, P., & Savolainen, H. (2012). Teacher education in Finland: A review of a national effort for preparing teachers for the future. *The Curriculum Journal*, *23*(4), 567– 584. https://doi.org/10.1080/09585176.2012.731011

McLaren, P. (2001). Che Guevara, Paulo Freire, and the politics of hope: Reclaiming critical pedagogy. *Critical Methodologies*, *1*(1), 108-131. https://doi.org/10.1177/15327086010010011

Oliva, P.F. (1997). *Developing the curriculum*. (4th Ed.). Longman.

Ornstein, A.C., & Hunkins, F.P. (2004). *Curriculum: Foundations, principles and issues*. Prentice Hall.

Romanelli, J.P., Fujimoto, J.T., Ferreira, M.D., & Milanez, D.H. (2018). Assessing ecological restoration as a research topic using bibliometric indicators. *Ecological Engineering*, *120*, 311-320. https://doi.org/10.1016/j.ecoleng.2018.06.015

Simola, H. (2005). The Finnish miracle of PISA: historical and sociological remarks on teaching and teacher education. *Comparative Education*, *41*(4), 455-470. https://doi.org/10.1080/03050060500317810

Staton-Spicer, A.Q, & Wulff, D.H. (1984). Research in communication and instruction: Categorization and synthesis. *Communicative Education, 33*(4), 377-391. https://doi.org/10.1080/03634528409384767

Zupic, I., & Cater, T. (2015). Bibliometric methods in management and organization. *Organizational Research Methods*, *18*(3), 429-472. https://doi.org/10.1177/1094428114562629

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 3, 507–531

https://doi.org/10.21449/ijate.1298691

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

Effects of gender and marital status on the perception of paternalistic leadership: A meta-analysis study

Mehmet Sabir Çevik [1,*]

[1] Yunus Emre Primary School, Siirt, Türkiye

ARTICLE HISTORY

Received: May 17, 2023
Revised: Aug. 18, 2023
Accepted: Aug. 23, 2023

Keywords:
Paternalistic leadership,
Gender,
Marital status,
Meta-analysis.

Abstract: This research aims to determine the overall effect size of gender and marital status on the perception of paternalistic leadership. In line with the research objective, studies on the perception of paternalistic leadership carried out between 2005 and 2022 in Türkiye were analysed with the meta-analysis method. Meta-analysis covered 22 studies on gender (n_{Gender}=9569) and 10 studies on marital status ($n_{\text{Marital Status}}$=6397) on the perception of paternalistic leadership. In this meta-analysis study utilising the random effects model, the Hedges' g value determining the standardised mean difference between the groups was used to calculate the effect sizes, and the origin of the heterogeneity in the research was tried to be determined by the moderator (sub-group) analyses. Research results revealed that the overall effect size of gender on the perception of paternalistic leadership was at a low level, with a value of 0.170, while the effect size of marital status was at a mean level, with a value of -0.523. However, it was determined in the moderator (sub-group) analyses that the effect size led to a statistically significant difference just in terms of sample groups in both gender and marital status variables.

1. INTRODUCTION

Discussions on leadership and effective leadership have gradually increased in recent years. Some of these discussions pertain to classical leadership approaches, and some to approaches emphasising contemporary and cultural contexts (Drost & Von Glinow, 1998; House et al., 2004). Moreover, there are also leadership approaches highlighting the leader's characteristics and advocating that these characteristics direct the behaviours of the employees in an organisation (Stahl, 2007). Yet, the common point of the discussions and explanations on leadership, in general, is viewed as the leaders' influence on and motivation of the organisation's employees (Anwar, 2013). Leaders can influence and motivate the organisation's members by displaying different leadership styles in various cultures (Türesin et al., 2015). Thus, different leadership styles or approaches have a formative effect on the acts and behaviours of the organisation's members (Mumford et al., 2002). In this context, paternalistic leadership is stated as one of the leadership styles emerging according to the cultural characteristics of the societies and influencing the acts and behaviours of the organisation's members (Cerit, 2013).

Paternalistic leadership originates from the sociocultural differences between Western and Eastern societies (Aycan, 2006). In other words, as a leadership style appearing in hierarchical

*CONTACT: Mehmet Sabir Çevik ✉ sahici1980@gmail.com Yunus Emre Primary School, Siirt, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

and traditional societies, paternalism is considered as a leadership approach prevailing more in Eastern than Western societies (Gürlek et al., 2020). Accordingly, it is known that societies in which paternalistic leadership is intensely observed display collectivist characteristics and high-power distances (Gelfand et al., 2007). Paternalism gained popularity in management and leadership because it is closely related to social characteristics, and organisations are structures affected by social characteristics (Martinez, 2003). The popularity of the paternalistic leadership style in the management and leadership fields is explained by its determinative role in organisational behaviours and organisational outputs (Bedi, 2020). In this context, there is a consensus among the researchers that paternalistic leadership increases positive organisational outputs (Demirer, 2012; Erben & Güneşer, 2008; Lee et al., 2018; Lin et al., 2015; Mussolino & Calabrò, 2014; Yeh et al., 2008); and that it hinders negative and undesirable outputs in organisations (Cheng et al., 2013; Dedahanov et al., 2019; Mulla & Krishnan, 2012; Wang & Cheng, 2010). Moreover, the literature includes significant research on the antecedents and consequences of paternalistic leadership. For instance, national and international literature covers various research carried out in several organisations on the relation of paternalistic leadership with organisational variables such as organisational citizenship (Göncü et al., 2014; Chu & Hung, 2009; Mete & Serin, 2015), organisational identification (Cheng et al., 2004; Korkmaz et al., 2018), organisational commitment (Pellegrini et al., 2010; Shi et al., 2020), organisational justice (Köksal, 2011), job satisfaction (Chamundeswari, 2013; Ekmen & Okçu, 2021; Sun & Wang, 2009), mobbing (Durmaz, 2019; Soylu, 2011), organisational creativity and organisational dissent (Ağladay & Dağlı, 2021), organisational happiness (Özgenel & Canulansı, 2021), job performance (Liang et al., 2007; Mert & Özgenel, 2020; Nigama et al., 2018), emotional labour (Zheng et al., 2020) and participation in decision making (Cansoy et al., 2020). Therefore, it appears that several variables can be associated with paternalistic leadership.

Antecedents of paternalistic leadership might include organisational variables as well as demographic (personal) variables such as gender and marital status (Erben & Güneşer, 2008; Kurt, 2013; Mete & Serin, 2015; Saylık, 2017; Taşdemir & Atalmış, 2021; Wu et al., 2011; Zhang et al., 2015). Thus, research examining the perception of paternalistic leadership in Türkiye according to demographic variables such as gender and marital status is remarkable. Some of the research revealed that gender causes a significant difference on the perception of paternalistic leadership (Cerit et al., 2011; Delice, 2020; Dursun, 2019; Kara et al., 2020; Karşu Cesur, 2015; Kılınç, 2019; Mert & Özgenel, 2020; Özgenel & Dursun, 2020; Saylık, 2017), while some advocated that it does not cause a significant difference (Ağalday, 2017; Arslan, 2016; Aydınoğlu, 2020; Bilici, 2017; Burgazlıoğlu, 2022; Dağlı & Ağalday, 2018; Hatipoğlu et al., 2019; İncegöz & Uslu, 2022; Koç, 2019; Korkmaz, 2018; Nal, 2018; Özgenel & Canulansı, 2021; Sarı, 2021). Concerning the marital status variable, some research pointed to a significant difference in the perception of paternalistic leadership (Abacı, 2020; Taşdemir & Atalmış, 2021), while some advocated that there is no significant difference (Ağalday, 2017; Aydınoğlu, 2020; Burgazlıoğlu, 2022; Korkmaz, 2018; Sarı, 2021; Saylık, 2017; Dağlı & Ağalday, 2018; Delice, 2020). All these indicate that the literature in Türkiye provides different and inconsistent results regarding the effect of gender and marital status variables on the perception of paternalistic leadership. Moreover, no research was found in the literature examining the effects of gender and marital status on the perception of paternalistic leadership with the meta-analysis method. Therefore, this research is considered to eliminate the uncertainty regarding the effect of gender and marital status variables on the perception of paternalistic leadership and to enable the synthesis of the research results. Besides, this research also examines the effects of gender and marital status on the perception of paternalistic leadership considering the variables, providing more accurate and precise results. The research results are considered to guide the researchers willing to investigate the perception of paternalistic leadership and provide the policymakers with foresight about the effect of gender and marital status on the perception of paternalistic leadership.

1.1. Paternalistic Leadership

The word paternalism, derived from the Latin word “pater”, is mostly used in a father’s taking care of his family and children. Paternalism means acting and behaving like a father and in a protective manner towards others (Bing, 2004; Suber, 1999). However, meanings attributed to paternalism are very complex and various (Aycan, 2006). For instance, paternalism is not only used as a negative term because of its derogatory connotation but also as a positive term in the sense of parents watching over their family members (Agich, 2003). In the management and leadership literature, the concept of paternalism has appeared as paternalistic leadership or paternal leadership. In the literature, paternalistic leadership has various definitions, such as helping the employees of the organisation in all matters under moral obligations (Farh & Cheng, 2000), meeting every need of the employees of the organisation with a paternal sensitivity (Afsar & Rehman, 2015), being involved in the private lives of the subordinates and protecting them (Pellegrini & Scandura, 2006), expecting respect and obedience from the employees (Aycan, 2006), dealing with and solving problems that the employees encounter outside their working lives (Huse & Mussolino, 2008). In light of these definitions and explanations, it is realised that paternalistic leadership aims to ensure a family atmosphere in organisational life, considers the organisation’s employees as family members, and involves a leadership approach based on obedience and respect.

Leadership approaches might vary among societies or cultures. A valid and prevailing leadership style in Eastern societies might not apply in Western societies (Fikret-Paşa, 2000; Westwood, 1997). Although the paternalistic leadership style is based on the teachings of Aristotle and Confucius and is one of the most common leadership approaches worldwide, it does not attract adequate attention in Western literature (Aycan et al., 2013). However, it was stated that paternalistic leadership had recently become prevalent in countries that can be considered Western, such as North America (Aycan et al., 2000). On the other hand, due to its content, the paternalistic leadership approach is a leadership style more suitable for the cultural textures of Asian societies; and it is common in countries such as China, Türkiye, Pakistan and India (Jackson, 2016). In organisational life, the paternalistic leadership style is observed in countries with high power distances and collectivist characteristics (Salminen Karlsson, 2015). Yet, the leadership style prevailing in a society cannot be dissociated from the culture and values of that society (Hofstede, 2006; Yukl, 2008). In other words, it might be asserted that the paternalistic leadership approach is closely related to social characteristics, and thus, based on the cultural values of a society, it might be stated whether it will become a prevailing leadership style in that society or not.

In the literature, the paternalistic leadership approach is conceptualised under different dimensions. Farh and Cheng (2000) addressed paternalistic leadership under the dimensions of “moral (ethical) leadership, benevolent leadership, authoritarian leadership,” while Aycan (2001) addressed it under “interest-based leadership and benign leadership”. Moral leadership means a leader being virtuous by displaying a high level of personal integrity. In contrast, while benevolent leadership corresponds to meeting all kinds of familial and personal needs of the organisation's employees, authoritarian leadership corresponds to a leader expecting subordinates to obey them without questioning and with respect (Liao et al., 2017). Interest-based leadership is the leader displaying intended behaviours in line with their own interests. In self-interested paternalism, the generosity or goodwill of the leader revolves around concerns about the work to be completed in the organisation (Hayek et al., 2010). However, benign leadership aims to promote the welfare, happiness and well-being of employees in a neutral and objective manner. In other words, paternalistic leaders with goodwill strive to meet the needs and expectations of their employees (Aycan, 2006). Based on these explanations, it may be stated that the moral, benevolent, and benign dimensions of paternalistic leadership correspond to a favourable and positive leadership approach. In contrast, authoritarian leadership and interest-based leadership dimensions correspond to a leadership approach that is undesirable or not much preferred in organisations.

1.2. Purpose of the Study

The research primarily aims to identify the effect sizes of gender and marital status on the perception of paternalistic leadership. In line with this primary objective, answers to the following questions were sought:

RQ1. What is the effect size of gender on the perception of paternalistic leadership?

RQ2. On the perception of paternalistic leadership, does the effect size of gender display a significant difference according to moderator (subgroup) variables (publication type, publication year, region of research, sample size, sample group and scale used)?

RQ3. What is the effect size of marital status on the perception of paternalistic leadership?

RQ4. On the perception of paternalistic leadership, does the effect size of marital status display a significant difference according to moderator (subgroup) variables (publication type, publication year, region of research, sample size, sample group and scale used)?

2. METHOD

2.1. Research Model

This research that aims to determine the effect sizes of gender and marital status on the perception of paternalistic leadership was carried out with the meta-analysis. Meta-analysis is the collection, interpretation, or synthesis with statistical methods of the empirical results of several quantitative research in any field (Lipsey & Wilson, 2001; Violato, 2019). The meta-analysis method examines the outcomes of different quantitative research with larger sample groups and through sound analyses (Cumming, 2012). The meta-analysis method was applied in this research as the aim was to synthesise the results of quantitative studies on the effect of gender and marital status on the perception of paternalistic leadership with larger sample groups and more robust analyses.

2.2. Literature Review Process

To obtain the studies carried out in Türkiye on paternalistic leadership, literature was reviewed by searching the keywords: “paternalist liderlik”, “babacan liderlik”, “paternalistic leadership”, and “paternalist leadership” in Turkish and English in “the National Thesis Centre of the Council of Higher Education (YÖK), Web of Science, ERIC, Google Scholar (Academic), National Academic Network and Information Centre (ULAKBİM), EBSCOhost, Science Direct, Sage Journals and ASOS” databases. The literature review was completed on 31.12.2022, and 122 studies were obtained in total. 122 studies obtained as a result of the literature review were identified according to the following inclusion criteria:

1.The studies were carried out in Türkiye between 2005 and 2022.

2.The studies are master’s theses, doctoral theses or articles published in refereed academic journals in Turkish or English.

3.The theses have access permits.

4. In case there was both a thesis study and an article study produced from the thesis using the same data in the literature, the article study produced from the thesis was included in the research.

5. The perception of paternalistic leadership was examined according to the variables of gender or marital status.

6. The overall total score for the perception of paternalistic leadership was reported.

7. Statistical information such as sample size, arithmetic mean, standard deviation, *p*-value and *t*-value were included in the studies to calculate effect sizes.
8. Full texts of the studies are accessible.

As a result of the literature review and based on the inclusion criteria, it was decided that the meta-analysis would include 22 studies on gender variable and ten on the marital status variable. As seen in Figure 1, the flow diagram of this meta-analysis was determined according to the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) flow model, as Moher et al. (2009) suggested.

Figure 1. *PRISMA flow diagram of the studies.*

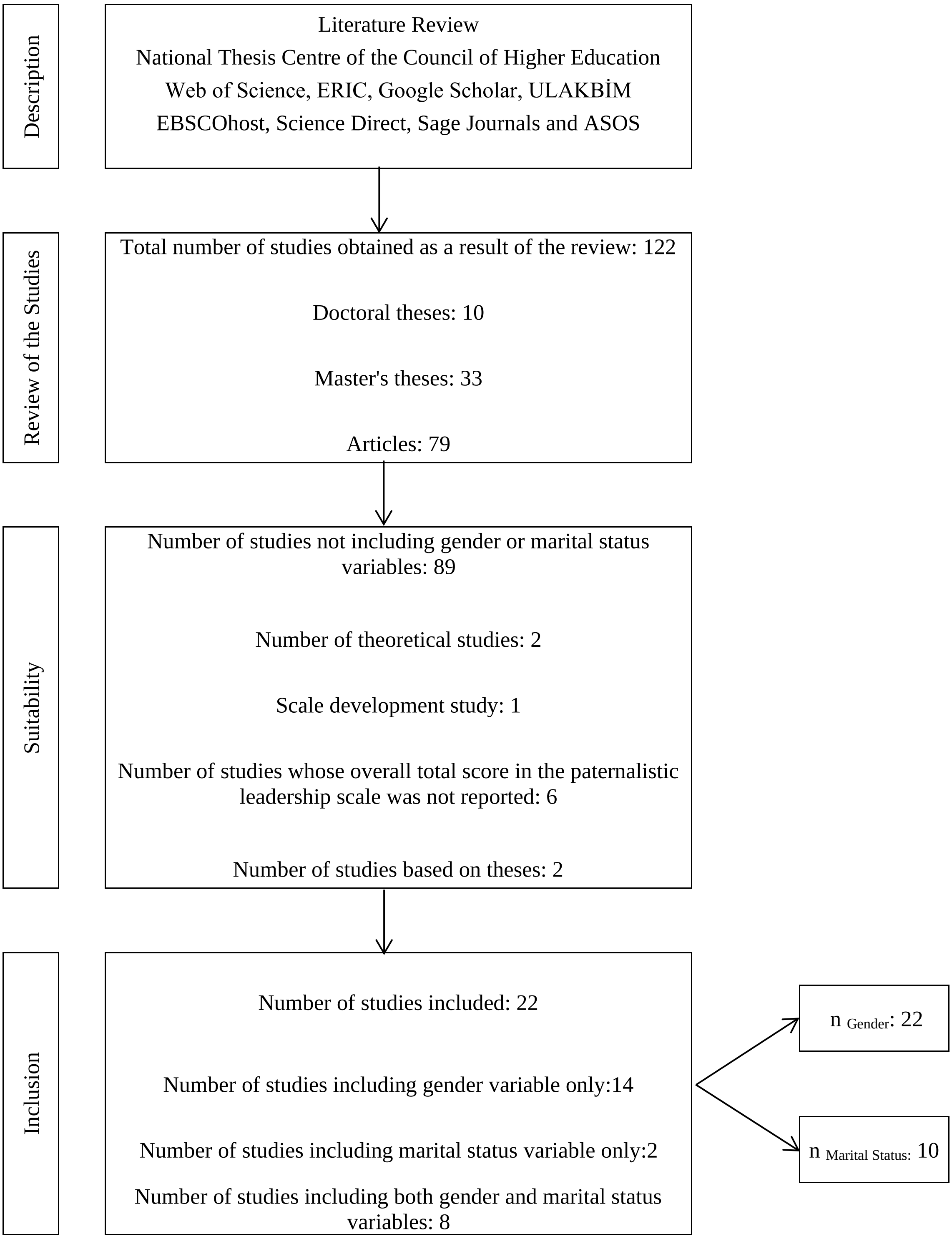

Description
Literature Review
National Thesis Centre of the Council of Higher Education
Web of Science, ERIC, Google Scholar, ULAKBİM
EBSCOhost, Science Direct, Sage Journals and ASOS
Review of the Studies
Total number of studies obtained as a result of the review: 122
Doctoral theses: 10
Master's theses: 33
Articles: 79
Suitability
Number of studies not including gender or marital status variables: 89
Number of theoretical studies: 2
Scale development study: 1
Number of studies whose overall total score in the paternalistic leadership scale was not reported: 6
Number of studies based on theses: 2
Inclusion
Number of studies included: 22
Number of studies including gender variable only:14
Number of studies including marital status variable only:2
Number of studies including both gender and marital status variables: 8
n Gender: 22
n Marital Status: 10

Table 1 displays descriptive information about the studies obtained regarding the gender and marital status variables as a result of the literature review.

Table 1. *Descriptive information about the studies included in the meta-analysis.*

Variables	Variable Information	Gender		Marital Status	
		f	%	f	%
Publication Type	Master's Thesis	9	40.91	4	40
	Doctoral Thesis	5	22.73	4	40
	Article	8	36.36	2	20
Publication Year	2011	1	4.55	-	-
	2015	1	4.55	-	-
	2016	1	4.55	-	-
	2017	3	13.64	2	20
	2018	3	13.64	2	20
	2019	4	18.18	-	-
	2020	5	22.73	3	30
	2021	2	9.09	2	20
	2022	2	9.09	1	10
Region of Research	Central Anatolia	4	18.18	3	30
	Marmara	7	31.82	1	10
	Southeastern Anatolia	2	9.09	2	20
	Mediterranean	2	9.09	2	20
	Black Sea	2	9.09	2	20
	Aegean	1	4.55	-	-
	Eastern Anatolia	1	4.55	-	-
	Other (mixed or not reported)	3	13.64	-	-
Sample Size	1-300	7	31.82	2	20
	301-600	10	45.45	3	30
	600 and above	5	22.73	5	50
Sample Group	Employees of Educational Organisations	14	63.64	7	70
	Other*	8	36.36	3	30
Scales Used	Cheng et al., 2004	3	13.64	2	20
	Pellegrini and Scandura, 2006	2	9.09	-	-
	Dağlı and Ağalday, 2017	7	31.82	3	30
	Aycan, 2006	6	27.27	-	-
	Other**	4	18.18	5	50

*Private sector, public employees, employees of enterprises and healthcare professionals,** Studies with the scales of Saylık (2017), Aycan et al. (2013), Saylık and Aydın (2020) and studies whose scales were not reported

As seen in Table 1, it was confirmed that there were 9 (40.91%) master's theses, 5 (22.73%) doctoral theses, and 8 (36.36%) articles examining the effect of gender on the perception of paternalistic leadership in Türkiye between 2005 and 2022, while there were 4 (40%) master's theses, 4 doctoral theses and 2 (20%) articles examining the effect of marital status. The number of studies published on the effect of gender on the perception of paternalistic leadership was highest in 2020 (n=5, 22.73%), and the number of studies published on the effect of marital status on the perception of paternalistic leadership was highest in 2020 (n=3, 30%). Research on paternalistic leadership, including gender variables, was mostly carried out in the Marmara region (n=7, 31.82%), while those including marital status variables were mostly carried out in the Central Anatolia region (n=3, 30%). With regard to sample size, it was determined that the studies on paternalistic leadership, including gender variables, were mostly carried out with varying numbers of participants between 301 and 600 (n=10, 45.45%), while the studies on paternalistic leadership, including marital status variable, were mostly carried out with 600 and more participants (n=5, 50%). The effect of gender and marital status on the perception of paternalistic leadership was mostly examined among the employees of educational organisations ($n_{\text{Sample Group-Gender}}$=14, 63.64%; $n_{\text{Sample Group-Marital Status}}$= 7, 70%). Lastly, it was found out that the most commonly used scale in the studies on paternalistic leadership, including gender variable, was the paternalistic leadership scale developed by Dağlı and Ağalday (2017) (n=7, 31.82%), while the most commonly used scale in the studies on paternalistic leadership including marital status variable was different and varied among the studies (n=5, 50%).

2.3. Data Coding

To ensure validity and reliability in the meta-analysis research, studies should be checked by coders (Açıkel, 2009; Stewart & Kamins, 2001). Accordingly, a coding form was drafted to determine whether the studies included in the meta-analysis by the researcher met the inclusion criteria. The coding form consists of the “publication type, publication year, region of research, sample size, sample group, the scale used, and statistical information about the studies”. The research code was written by two expert researchers who studied meta-analysis. Coding by these two expert researchers was calculated according to the reliability formula proposed by Miles and Huberman (2002) (Reliability=Agreement/Agreement+Disagreement), and the intercoder reliability was determined as 96%. The intercoder agreement is specified to be at least 80% (Patton, 2002). Therefore, the coding reliability of the research might be considered sufficient. Moreover, non-overlapping codes were also re-evaluated and corrected by the researchers.

2.4. Publication Bias

Publication bias is deliberately not publishing studies that do not provide expected significant statistics from research carried out on any subject (Makowski et al., 2019). In other words, researchers or academic journals tend not to publish statistically insignificant studies. This leads to publication bias among the studies applying the meta-analysis method (Borenstein et al., 2013). Presence of publication bias results in deviations in terms of the accuracy of the studies' average effect sizes (Field & Gillett, 2010). Accordingly, the presence of publication bias in this meta-analysis study was checked. Publication bias of the research was determined separately for both gender and marital status based on the Funnel plot (scatter plot), Begg and Mazumdar's rank correlation test, Rosenthal's Fail-Safe N value, Egger's regression test and Duval and Tweedie's trim and fill test results.

2.5. Heterogeneity

In meta-analysis studies, heterogeneity refers to the range of effect sizes of the studies included (Şen & Yıldırım, 2020). In meta-analysis studies, heterogeneity is examined with the *Q* test and I^2 value. Heterogeneity can be mentioned when the *Q* value calculated according to the degrees of freedom is higher than the chi-square value (x^2) or when the I^2 value is higher than 75% (Card, 2011; Cooper et al., 2009). On the condition that a meta-analysis study is heterogeneous, moderator (subgroup) analyses are needed. In other words, moderator analysis determines the causes of heterogeneity (Deeks et al., 2008). Accordingly, the effect size of gender and marital status on the perception of paternalistic leadership was also examined according to moderator variables (publication type, publication year, region of research, sample size, sample group and the scale used).

2.6. Selection of the Model

Meta-analysis studies are analysed according to fixed effects or random effects models. In the fixed effects model, all studies share the same effect size, and weightings are based on the number of observations. In contrast, in the random effects model, the effect sizes vary according to different characteristics (Cooper et al., 2009). In social sciences, the random effects model is advised to be used more in meta-analysis studies (Pigott & Polanin, 2020). Moreover, the model used in meta-analysis studies might be decided based on the heterogeneity test results (Q test and I2) (Dinçer, 2014). Accordingly, in determining the model to be used in this research, both the theoretical explanations and the heterogeneity test results (Q test and I2) were considered.

2.7. Calculation of the Effect Sizes

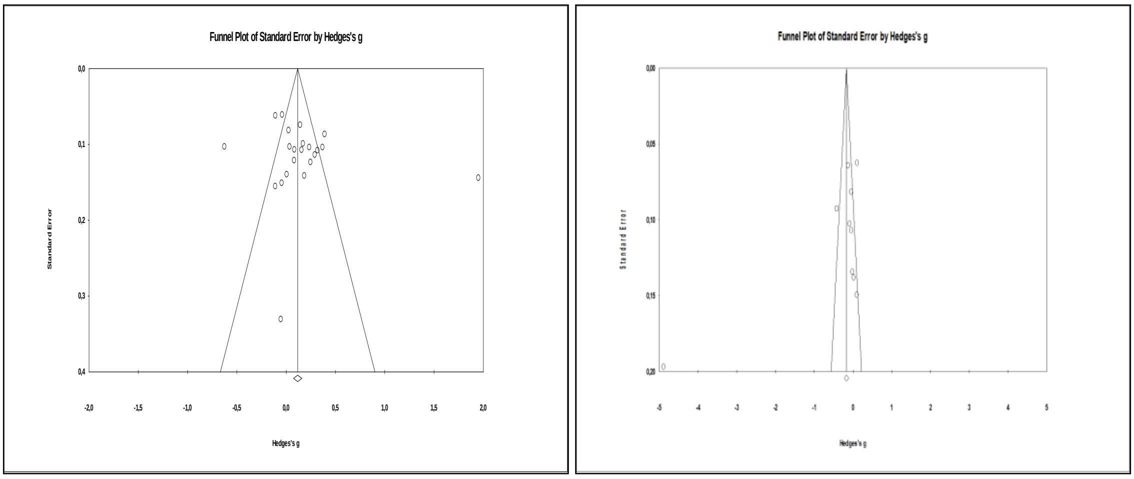
This meta-analysis study calculated effect sizes with the *Hedges' g* value, identifying the standardised mean difference between the groups. In this context, the data were interpreted according to a .05 significance level with the Comprehensive Meta-Analysis (CMA) statistical package program. Effect sizes were evaluated according to the criteria determined by Cohen (1992) as "≤ 0.2, low effect size; 0.50, medium effect size and ≥ 0.80, large effect size". A positive effect size on gender indicates that males have a higher perception of paternalistic leadership, while a positive effect size on marital status suggests that singles have a higher perception of paternalistic leadership. Besides, whether the effect size of gender and marital status on the perception of paternalistic leadership differs significantly in terms of "publication type, publication year, region of research, sample size, sample group and the scale used" was examined with moderator (subgroup) analyses, $Q_{Between}$, χ^2 and *p*-value.

3. FINDINGS

3.1. Findings Regarding the Publication Bias

Before the analyses on the effect sizes, the publication bias results of the research were checked. In this context, the publication bias of the research was determined separately for both gender and marital status by the Funnel plot (scatter plot), Begg and Mazumdar's rank correlation test, Rosenthal's Fail-Safe N value, Egger's regression test and Duval and Tweedie's trim and fill test results. Figure 2 displays the Funnel plot (scatter plot) graphics of the studies regarding a) gender and b) marital status, respectively.

Figure 2. *Funnel plot (scatter plot) graphics according to a) gender and b) marital status on the perception of paternalistic leadership.*



a) Gender

b) Marital Status

As seen in Figure 2, examining the research's Funnel plot (scatter plot) graphics on gender and marital status, it was determined that the effect sizes generally concentrated symmetrically around the standard error. In meta-analysis studies, the symmetric distribution of effect sizes around the standard error indicates the absence of publication bias (Borenstein et al., 2013). However, it is not correct to decide on the presence of publication bias based on just the Funnel plot (scatter plot) (Lipsey & Wilson, 2001; Petticrew & Roberts, 2006). Therefore, publication bias of the research on gender and marital status variables was determined by Begg and Mazumdar's rank correlation test, Rosenthal's Fail-Safe N value, Egger's regression test and Duval and Tweedie's trim and fill test results. Table 2 displays Begg and Mazumdar's rank correlation test, Rosenthal's Fail-Safe N value, and Egger's regression test results.

Table 2. *Begg and Mazumdar's rank correlation test, Rosenthal's Fail-Safe N value, Egger's regression test results.*

Reliability Test	Reliability Test Values			
	Gender		Marital Status	
Begg and Mazumdar's Rank Correlation Test	Tau	0.09957	Tau	-0.06667
	Z value for Tau	0.64855	Z value for Tau	0.26833
	p value (two sides)	0.51663	*p* value (two sides)	0.78845
Rosenthal's Fail-Safe N Value	Z value	6.52360	Z value	-9.726657
	p value	0.00000	*p* value	0.00000
	Alpha	0.05000	Alpha	0.05000
	Side	2.00000	Side	2.00000
	Z value for Alpha	1.95996	Z value for Alpha	1.95996
	Fail-Safe N Value	222	Fail-Safe N Value	237
Egger's Regression Test	Standard error	2.45805	Standard error	7.12092
	95% lower threshold value	-1.95296	95% lower threshold value	-28.93770
	95% upper threshold value	8.30183	95% upper threshold value	3.90405
	t-value	1.29145	*t*-value	1.75775
	df	20	*df*	8
	p value (two sides)	0.21128	*p* value (two sides)	0.11685

Table 2 confirms the absence of publication bias as the *p* values for gender and marital status were 0.51663 (p>0.05) and 0.78845 (p>0.05), respectively, according to the results of Begg and Mazumdar's rank correlation test. Moreover, Rosenthal's Fail-Safe N value was identified as 222 for gender and 237 for marital status. 222 for gender and 237 for marital status refer to the number of studies that should be included to refrain from mentioning a significant effect. It is not possible to reach 222 and 237 in practice, and the N/(5k+10) value is higher than 1 for gender [222/(5x22+10)=1.850>1] and for marital status [237/(5x10+10)=3.95>1], and thus, these indicate that there is no publication bias (Mullen et al., 2001). Besides, statistically insignificant *p* values in the Egger test (p_{Gender}=0.21128>0.05; $p_{\text{Marital status}}$=0.11685>0.05) (Rothstein et al., 2005) confirm the absence of publication bias in the research. Table 3 displays the results of Duval and Tweedie's trim and fill method, another indicator of the availability or absence of publication bias.

Table 3. *Results of Duval and Tweedie's Trim and fill method on gender and marital status.*

			Confidence Interval (95%)		
Gender	Difference	Point Estimate	Lower Threshold	Upper Threshold	Q
Observed Value		0.17005	0.01682	0.32328	262.69384
Adjusted Value	0	0.17005	0.01682	0.32328	262.69384
Marital Status					
Observed Value		-0.52303	-1.01954	-0.02652	611.96025
Adjusted Value	0	-0.52303	-1.01954	-0.02652	611.96025

As seen in Table 3, the number of trimmed studies on both gender (Observed Value $_{\text{Point Estimate}}$=0.17005; Adjusted Value P$_{\text{oint Estimate}}$=0.17005) and marital status (Observed Value $_{\text{Point Estimate}}$=-0.52303; Adjusted Value $_{\text{Point Estimate}}$) = -0.52303) was determined as 0, and this might be interpreted as the absence of publication bias. Accordingly, depending on the results of the Funnel plot (scatter plot), Begg and Mazumdar's rank correlation test, Rosenthal's Fail-Safe N value, Egger's regression test and Duval and Tweedie's trim and fill method, it might be asserted that there is no publication bias in this meta-analysis study as a whole.

3.2. Findings Regarding the Heterogeneity Tests

In order to decide on the effect size model for the research, heterogeneity tests were carried out on gender and marital status variables. Accordingly, Table 4 displays the heterogeneity test results for the model to be used in calculating the effect sizes according to gender and marital status on the perception of paternalistic leadership.

Table 4. *Heterogeneity test results of the research on gender and marital status.*

			95% Confidence Interval			Heterogeneity test		
Gender	k	Point Estimate	Lower Threshold	Upper Threshold	Q value	*df*	*p*	I^2
Fixed Effects	22	0.117	0.074	0.159	262.694	21	0.000	92.006
Random Effects	22	0.170	0.017	0.323				
Marital Status								
Fixed Effects	10	-0.171	-0.230	-0.113	611.960	9	0.000	98.529
Random Effects	10	-0.523	-1.020	-0.027				

k: Number of studies

As seen in Table 4, the Q value for gender was determined as 262.694, while the Q value for marital status was determined as 611.960. Concerning gender, the Q value (Q_{Gender}=262.694) corresponds to 32.671 at 21 degrees of freedom and 0.05 significance level in the chi-square table (x^2), while according to marital status, the Q value ($Q_{\text{Marital Status}}$=611.960) corresponds to 16.919 at 9 degrees of freedom and 0.05 significance level in the chi-square table (x^2). Besides, the Higgins I^2 value of the research on gender was determined as 92.006, while the Higgins I^2 value on marital status was determined as 98.529. Q values of the research are beyond the chi-square (x^2) table values and are significant at the p=0.05 level, and the Higgins I^2 values are higher than 75%, and these mean that the data are heterogeneous in terms of gender and marital status (Card, 2011; Cooper et al., 2009; Higgins & Thompson, 2002). Moreover, the availability of intervening variables in the research, such as the publication type, publication year, region of research, sample size, sample group, and the scale used, points out the possibility of change in effect sizes in the research (Üstün & Eryılmaz, 2014). Consequently, based on all these analyses and grounds, the research was identified as heterogeneous, and it was decided to use the random effects model in the research.

3.3. Findings Regarding the Effect Size

This part addresses the effect sizes of the studies examining the perception of paternalistic leadership according to gender and marital status in the random effects model. Table 5 displays the effect sizes of the perception of paternalistic leadership according to gender.

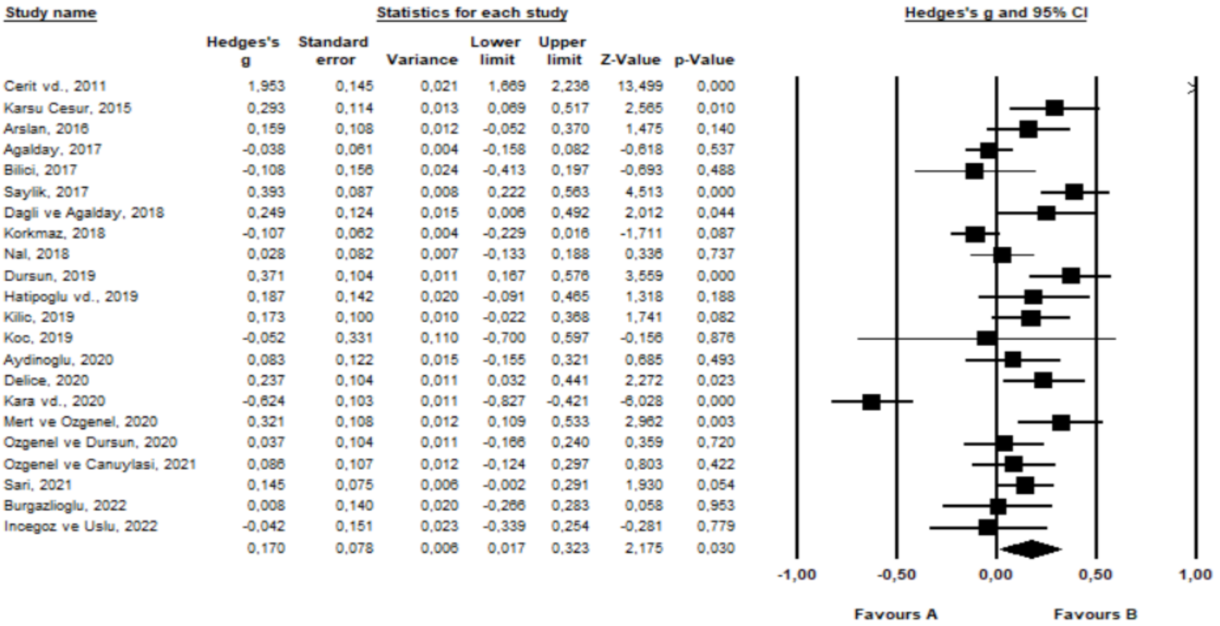
Table 5. *Effect sizes of the perception of paternalistic leadership on gender.*

Research Title	Effect Size	Confidence Interval (95%)		Z	*p*	n
	(Hedges's g)	Lower Threshold	Upper Threshold			
Cerit et al., 2011	1.953	1.669	2.236	2.236	0.000*	284
Karşu Cesur, 2015	0.293	0.069	0.517	0.517	0.010*	346
Arslan, 2016	0.159	-0.052	0.370	0.370	0.140	349
Ağalday, 2017	-0.038	-0.158	0.082	0.082	0.537	1059
Bilici, 2017	-0.108	-0.413	0.197	0.197	0.488	171
Saylık, 2017	0.393	0.222	0.563	0.563	0.000*	700
Dağlı and Ağalday, 2018	0.249	0.006	0.492	0.492	0.044	261
Korkmaz, 2018	-0.107	-0.229	0.016	0.016	0.087	1032
Nal, 2018	0.028	-0.133	0.188	0.188	0.737	683
Dursun, 2019	0.371	0.167	0.576	0.576	0.000*	420
Hatipoğlu et al., 2019	0.187	-0.091	0.465	0.465	0.188	200
Kılıç, 2019	0.173	-0.022	0.368	0.368	0.082	405
Koç, 2019	-0.052	-0.700	0.597	0.597	0.876	57
Aydınoğlu, 2020	0.083	-0.155	0.321	0.321	0.493	413
Delice, 2020	0.237	0.032	0.441	0.441	0.023*	370
Kara et al., 2020	-0.624	-0.827	-0.421	-0.421	0.000*	400
Mert and Özgenel, 2020	0.321	0.109	0.533	0.533	0.003*	431
Özgenel and Dursun, 2020	0.037	-0.166	0.240	0.240	0.720	420
Özgenel and Canuylası, 2021	0.086	-0.124	0.297	0.297	0.422	449
Sarı, 2021	0.145	-0.002	0.291	0.291	0.054	717
Burgazlıoğlu, 2022	0.008	-0.266	0.283	0.283	0.953	210
İncegöz and Uslu, 2022	-0.042	-0.339	0.254	0.254	0.779	192
Random Effects Model	0.170	0.017	0.323	2.175	0.030*	9569

* *p*< 0.05

As seen in Table 5, it was determined that the effect sizes of the studies on gender carried out with a total of 9569 participants vary between -0.624 and 1.953; and the study with the highest effect size (1.953) was carried out by Cerit et al. (2011), while the study with the lowest effect size (0.008) by Burgazlıoğlu (2022). Besides, according to the random effects model, the overall effect size of paternalistic leadership perception on gender is 0.170 [Confidence Interval (95%): 0.017; 0.323; *p*=0.030<0.05], and it was determined that male participants had significantly higher perceptions of paternalistic leadership than female participants. The overall effect size calculated according to gender (Effect Size$_{\text{Gender}}$ = 0.170) corresponds to a "low effect size" according to Cohen's (1992) effect size classification. This result indicates that the perception of paternalistic leadership significantly differs according to gender. Figure 3 displays the forest plot of the perception of paternalistic leadership regarding gender.

Figure 3. *Forest plot of the perception of paternalistic leadership on gender.*



Study name	Statistics for each study						
	Hedges's g	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
Cerit vd., 2011	1,953	0,145	0,021	1,669	2,236	13,499	0,000
Karsu Cesur, 2015	0,293	0,114	0,013	0,069	0,517	2,565	0,010
Arslan, 2016	0,159	0,108	0,012	-0,052	0,370	1,475	0,140
Agalday, 2017	-0,038	0,061	0,004	-0,158	0,082	-0,618	0,537
Bilici, 2017	-0,108	0,156	0,024	-0,413	0,197	-0,693	0,488
Saylik, 2017	0,393	0,087	0,008	0,222	0,563	4,513	0,000
Dagli ve Agalday, 2018	0,249	0,124	0,015	0,006	0,492	2,012	0,044
Korkmaz, 2018	-0,107	0,062	0,004	-0,229	0,016	-1,711	0,087
Nal, 2018	0,028	0,082	0,007	-0,133	0,188	0,336	0,737
Dursun, 2019	0,371	0,104	0,011	0,167	0,576	3,559	0,000
Hatipoglu vd., 2019	0,187	0,142	0,020	-0,091	0,465	1,318	0,188
Kilic, 2019	0,173	0,100	0,010	-0,022	0,368	1,741	0,082
Koc, 2019	-0,052	0,331	0,110	-0,700	0,597	-0,156	0,876
Aydinoglu, 2020	0,083	0,122	0,015	-0,155	0,321	0,685	0,493
Delice, 2020	0,237	0,104	0,011	0,032	0,441	2,272	0,023
Kara vd., 2020	-0,624	0,103	0,011	-0,827	-0,421	-6,028	0,000
Mert ve Ozgenel, 2020	0,321	0,108	0,012	0,109	0,533	2,962	0,003
Ozgenel ve Dursun, 2020	0,037	0,104	0,011	-0,166	0,240	0,359	0,720
Ozgenel ve Canuylasi, 2021	0,086	0,107	0,012	-0,124	0,297	0,803	0,422
Sari, 2021	0,145	0,075	0,006	-0,002	0,291	1,930	0,054
Burgazlioglu, 2022	0,008	0,140	0,020	-0,266	0,283	0,058	0,953
Incegoz ve Uslu, 2022	-0,042	0,151	0,023	-0,339	0,254	-0,281	0,779
	0,170	0,078	0,006	0,017	0,323	2,175	0,030

As seen in Figure 3, the squares represent the effect sizes of the research, while the diamond shape in the form of a rhombus at the bottom of the figure represents the overall effect size. Lines on both sides of the squares display the distribution of each study's lower and upper thresholds according to a 95% confidence interval. According to Figure 3, 6 of the 22 studies included in this meta-analysis study have negative effect sizes, while 16 have positive ones.

Table 6 displays the effect sizes of the perception of paternalistic leadership according to the marital status variable.

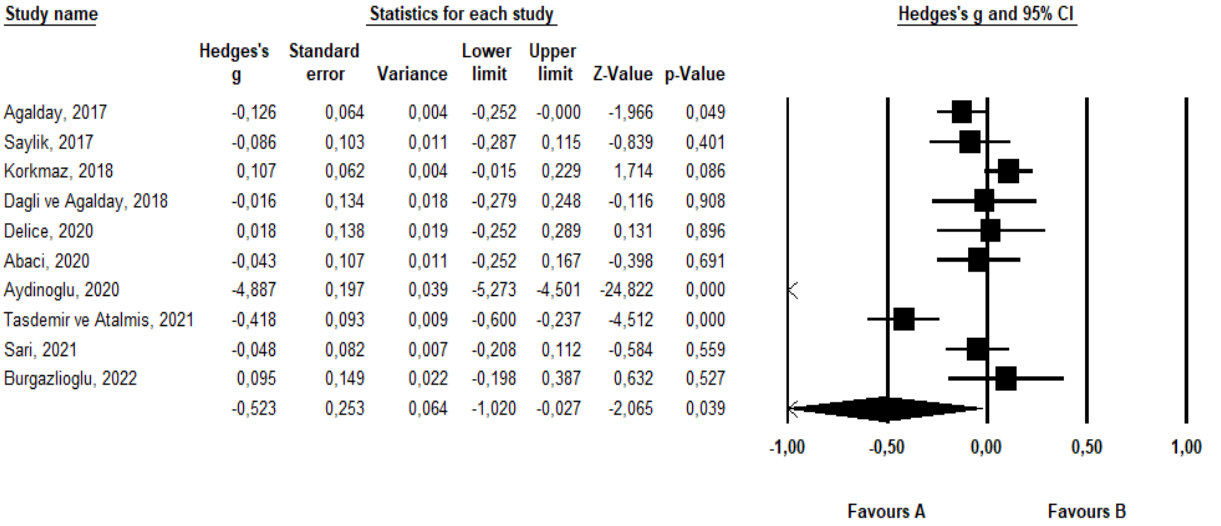
Table 6. *Effect sizes of the perception of paternalistic leadership according to marital status variable.*

		Confidence Interval (95%)				
Research Title	Effect Size (Hedges's g)	Lower Threshold	Upper Threshold	Z	*p*	n
Ağalday, 2017	-0.126	-0.252	0.000	-1.966	0.049	1632
Saylık, 2017	-0.086	-0.287	0.115	-0.839	0.401	700
Korkmaz, 2018	0.107	-0.015	0.229	1.714	0.086	1032
Dağlı and Ağalday, 2018	-0.016	-0.279	0.248	-0.116	0.908	261
Delice, 2020	0.018	-0.252	0.289	0.131	0.896	370
Abacı, 2020	-0.043	-0.252	0.167	-0.398	0.691	422
Aydınoğlu, 2020	-4.887	-5.273	-4.501	-24.822	0.000	413
Taşdemir and Atalmış, 2021	-0.418	-0.600	-0.237	-4.512	0.000	640
Sarı, 2021	-0.048	-0.208	0.112	-0.584	0.559	717
Burgazlıoğlu, 2022	0.095	-0.198	0.387	0.632	0.527	210
Random Effects Model	-0.523	-1.020	-0.027	-2.065	0.039	6397

As seen in Table 6, it was established that the effect sizes of the studies on marital status, carried out with a total of 6397 participants, vary between -4.887 and 0.107, and the study with the highest effect size (-4.887) was carried out by Aydınoğlu (2020), while the study with the lowest effect size (0.018) by Delice (2020). Besides, according to the random effects model, the overall effect size of paternalistic leadership perception according to marital status is -0.523

[Confidence Interval (95%): -1.020; -0.027; *p*=0.039<0.05], and it was determined that married participants had significantly higher perceptions of paternalistic leadership than single participants. The overall effect size calculated according to marital status (Effect SizeMarital Status =-0.523) corresponds to a "medium effect size" according to Cohen's (1992) effect size classification. Thus, this result indicates that the perception of paternalistic leadership differs significantly according to marital status. Figure 4 displays the forest plot of the perception of paternalistic leadership regarding marital status.

Figure 4. *Forest plot of the perception of paternalistic leadership according to marital status.*



Study name	Statistics for each study							Hedges's g and 95% CI
	Hedges's g	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value	
Agalday, 2017	-0,126	0,064	0,004	-0,252	-0,000	-1,966	0,049	
Saylik, 2017	-0,086	0,103	0,011	-0,287	0,115	-0,839	0,401	
Korkmaz, 2018	0,107	0,062	0,004	-0,015	0,229	1,714	0,086	
Dagli ve Agalday, 2018	-0,016	0,134	0,018	-0,279	0,248	-0,116	0,908	
Delice, 2020	0,018	0,138	0,019	-0,252	0,289	0,131	0,896	
Abaci, 2020	-0,043	0,107	0,011	-0,252	0,167	-0,398	0,691	
Aydinoglu, 2020	-4,887	0,197	0,039	-5,273	-4,501	-24,822	0,000	
Tasdemir ve Atalmis, 2021	-0,418	0,093	0,009	-0,600	-0,237	-4,512	0,000	
Sari, 2021	-0,048	0,082	0,007	-0,208	0,112	-0,584	0,559	
Burgazlioglu, 2022	0,095	0,149	0,022	-0,198	0,387	0,632	0,527	
	-0,523	0,253	0,064	-1,020	-0,027	-2,065	0,039	

As seen in Figure 4, the squares represent the effect sizes of the studies, while the diamond shape in the form of a rhombus at the bottom of the figure represents the overall effect size. Lines on both sides of the squares display the distribution of each study's lower and upper thresholds according to a 95% confidence interval. Based on Figure 3, it was determined that 7 of the 10 studies included in this meta-analysis study had negative effect sizes while 3 had positive effect sizes.

3.4. Findings Regarding the Moderator (Subgroup) Effect Analyses

Tables 7 and 8 display the analysis results on the effect sizes of gender and marital status on the perception of paternalistic leadership regarding moderator variables (publication type, publication year, region of research, sample size, sample group and the scale used). Table 7 displays the analysis results according to the effect size of gender on the perception of paternalistic leadership on moderator variables.

Table 7. *Analysis results on the effect size of gender in the perception of paternalistic leadership according to moderator variables.*

Moderator	k	Effect Size (Hedges's g)	Confidence Interval (95%) Lower Threshold	Confidence Interval (95%) Upper Threshold	Q_b	*df*	*p*
Publication Type							
Master's Thesis	9	0.177	0.092	0.262	1.592	2	0.451
Doctoral Thesis	5	0.065	-0.103	0.232			
Article	8	0.267	-0.195	0.730			
Publication Year							
Between 2011 and 2018	9	0.305	0.004	0.605	1.817	2	0.403
Between 2019 and 2020	9	0.088	-0.129	0.305			
Between 2021 and 2022	4	0.088	-0.015	0.192			
Region of Research							
Central Anatolia	4	0.085	-0.177	0.347	3.094	5	0.686
Marmara	7	0.185	0.064	0.307			
Southeastern Anatolia	2	0.086	-0.193	0.364			
Mediterranean	2	-0.194	-1.037	0.650			
Black Sea	2	1.044	-0.728	2.816			
Other**	5	0.094	-0.001	0.189			
Sample Size							
Between 1-300	7	0.333	-0.182	0.848	0.865	2	0.649
Between 301-600	10	0.108	-0.083	0.298			
601 and above	5	0.078	-0.083	0.238			
Sample Group							
Employees of Edu.Organis.	14	0.309	0.118	0.500	9.322	1	0.002*
Other***	8	-0.099	-0.279	0.080			
Scales Used							
Cheng et al., 2004	3	0.026	-0.152	0.203	2.926	4	0.570
Pellegrini and Scandura, 2006	2	0.986	-0.900	2.873			
Dağlı and Ağalday, 2017	7	0.155	0.035	0.275			
Aycan, 2006	6	-0.010	-0.340	0.319			
Other****	4	0.158	-0.060	0.376			

*$p < 0.05$, **Studies with several regions or whose region is not reported *** Private sector, public employees, employees of enterprises and healthcare professionals; ****Studies with the scales of Saylık (2017), Aycan et al. (2013), Saylık and Aydın (2020) and studies whose scales were not reported, k= Number of studies; Q_b=Intergroup Q value.

As seen in Table 7, it was determined that the effect size of gender on the perception of paternalistic leadership did not display a statistically significant difference according to publication type (Q_b=1.592; *df*=2; p>0.05), publication year (Q_b=1.817; *df*=2; p>0.05), the region of research (Q_b=3.094; *df*=5; p>0.05), sample size (Q_b=0.865; *df*=2; p>0.05) and the scale used (Q_b=2.926; *df*=4; p>0.05), but there was a significant difference only according to the sample group (Q_b=9.322; *df*=1; p<0.05). In other words, it was ascertained that only the sample group is a determining variable on the effect size of gender on the perception of paternalistic leadership.

Table 8 displays the analysis results on the effect size of marital status on the perception of paternalistic leadership according to moderator variables.

Table 8. *Analysis results on the effect size of marital status on the perception of paternalistic leadership according to moderator variables.*

Moderator	k	Effect Size (Hedges's g)	Confidence Interval (95%) Lower Threshold	Confidence Interval (95%) Upper Threshold	Q_b	*df*	*p*
Publication Type							
Master's Thesis	4	-0.017	-0.124	0.090	4.875	2	0.087
Doctoral Thesis	4	-1.228	-2.420	-0.036			
Article	2	-0.229	-0.623	0.165			
Publication Year							
Between 2011 and 2018	4	-0.026	-0.156	0.103	1.898	2	0.387
Between 2019 and 2020	3	-1.632	-4.218	0.953			
Between 2021 and 2022	3	-0.136	-0.431	0.158			
Region of Research							
Central Anatolia	3	-1.612	-3.678	0.455	2.060	1	0.151
Other**	7	-0.096	-0.218	0.025			
Sample Size							
Between 1-300	2	0.034	-0.162	0.229	2.618	2	0.270
Between 301-600	3	-1.632	-4.218	0.953			
601 and above	5	-0.108	-0.273	0.056			
Sample Group							
Employees of Edu.Organis.	7	-0.779	-1.514	-0.044	5.058	1	0.025*
Other***	3	0.072	-0.027	0.171			
Scales Used							
Cheng et al., 2004	2	-2.387	-7.281	2.508	0.870	2	0.647
Dağlı and Ağalday, 2017	3	-0.086	-0.179	0.006			
Other****	5	-0.103	-0.291	0.085			

* *p*< 0.05, **Studies with several regions or whose region is not reported *** Private sector, public employees, employees of enterprises and healthcare professionals; ****Studies with the scales of Saylık (2017), Aycan et al. (2013), Saylık and Aydın (2020) and studies whose scales were not reported, k= Number of studies; Q_b=Intergroup Q value.

As in Table 8, the effect size of marital status on the perception of paternalistic leadership did not display a statistically significant difference according to the publication type (Q_b=4.875; *df*=2; *p*>0.05), publication year (Q_b=1.898; df=2; *p*>0.05), region of the research (Q_b=2.060; *df*=1; *p*>0.05), sample size (Q_b=2.618; *df*=2; *p*>0.05) and the scale used (Q_b=0.870; *df*=2; *p*>0.05). Based on Table 8, it was determined that there was a significant difference regarding only the sample size (Q_b=5.058; *df*=1; *p*< 0.05). In other words, it was ascertained that only the sample group is a determining variable on the effect size of marital status on the perception of paternalistic leadership.

4. DISCUSSION and CONCLUSION

This research aims to determine the effect of gender and marital status variables on the perception of paternalistic leadership through the meta-analysis method. Moreover, it was also aimed in the research to figure out whether the effect sizes differ according to the publication type, publication year, region of the research, sample size, sample group and the scale used. Research results revealed that gender had a low effect size, and marital status had a medium effect size on the perception of paternalistic leadership. Besides, it was also found that the effect sizes of both gender and marital status displayed a significant difference only in terms of the sample group.

One of the most important results of the research is that the effect size of gender on the perception of paternalistic leadership was at a low level. Concerning the effect size of gender, it was found that the paternalistic leadership perception of the male participants was higher than that of female participants. Accordingly, it might be asserted that gender is an effective but not a determining variable in the perception of paternalistic leadership. In other words, the gender variable might be regarded as a variable with a low effect on the perception of paternalistic leadership. Practices in the organisation or organisational behaviours might vary according to gender (Britton, 2000). Certain leadership behaviours, such as establishing good relations with the employees, helping and supporting them, were considered feminine by Oplatka (2004). Similarly, Saylık (2017) explains the higher paternalistic leadership perceptions of male participants compared to female participants because most managers are men, and paternalistic leadership behaviours show more male-oriented characteristics. Naturally, feminine characteristics of some leadership behaviours might result in the males' expecting leaders of an organisation to be more paternalistic (Cerit et al., 2011). Literature covers different conclusions concerning the perception of paternalistic leadership according to gender. Gender was claimed to cause a significant difference in the perception of paternalistic leadership in some studies (Cerit et al., 2011; Delice, 2020; Dursun, 2019; Kara et al., 2020; Karşu Cesur, 2015; Kılıç, 2019; Mert & Özgenel, 2020; Özgenel & Dursun, 2020; Saylık, 2017), while it was claimed not to cause a significant difference in some other studies (Ağalday, 2017; Arslan, 2016; Aydınoğlu, 2020; Bilici, 2017; Burgazlıoğlu, 2022; Dağlı & Ağalday, 2018; Hatipoğlu et al., 2019; İncegöz & Uslu, 2022; Koç, 2019; Korkmaz, 2018; Nal, 2018; Özgenel & Canuylası, 2021; Sarı, 2021). However, while it was revealed in only one research that the paternalistic leadership perception of female participants was higher than that of male participants (Kara et al., 2020), other studies asserted that the paternalistic leadership perception of male participants was higher than that of female participants in general (Cerit et al., 2011; Delice, 2020; Dursun, 2019; Karşu Cesur, 2015; Kılınç, 2019; Mert & Özgenel, 2020; Özgenel & Dursun, 2020; Saylık, 2017). In almost all of the research carried out with samples from Türkiye, gender does not have a significant effect on the perception of paternalistic leadership, and men have higher perceptions of paternalistic leadership than that of women, and these might be related to Türkiye's male-dominated social dynamics and cultural values with collectivist characteristics. Thus, Salminen Karlsson's (2015) and Jackson's (2016) statement that paternalistic leadership style is typical in countries with high levels of collectivist characteristics and Hofstede's (2006) and Yukl's (2008) assertion that the dominant leadership style in a country is not independent of the culture of the concerned society support the research results as a whole.

Another notable result revealed by the research is that the effect size of the marital status variable on the perception of paternalistic leadership is at the medium level. Moreover, the research also established that the married have higher levels of paternalistic leadership perception than the singles. Based on the research results, marital status is a determining variable in the perception of paternalistic leadership among the participants. Literature covers research pointing that marital status causes a significant difference in the perception of paternalistic leadership (Abacı, 2020; Taşdemir & Atalmış, 2021), as well as research advocating the absence of any substantial difference (Ağalday, 2017; Aydınoğlu, 2020; Burgazlıoğlu, 2022; Dağlı & Ağalday, 2018; Delice, 2020; Korkmaz, 2018; Sarı, 2021; Saylık, 2017). Moreover, out of the research, two of them (Abacı, 2020; Taşdemir & Atalmış, 2021) pointing to significant differences established that the married participants had higher perceptions of paternalistic leadership than the singles, as also claimed in this research. Married participants have essential family responsibilities and have to care for their families more often, and these might have increased the awareness of the leaders of the organisation on paternalistic leadership behaviours. Besides, married participants' struggle to earn a living and fear of job loss due to financial concerns might have led to more positive perceptions of paternalistic leadership among them in comparison to that of singles. Ağalday (2017), in a study examining the paternalistic leadership behaviours of primary school principals, explains why married participants find school principals more paternalistic because school principals empathise with married teachers and act more benevolently because they are generally married. Though the literature sets forth different reasons for the higher perceptions of paternalistic leadership among the married participants compared to the singles, it is remarkable that this meta-analysis study identified marital status as an effective variable on the perception of paternalistic leadership.

Moderator analyses under the research revealed that the effect of both gender and marital status on the perception of paternalistic leadership differs only according to the sample group. In other words, it might be asserted that the research's effect sizes vary according to whether participants are employees of educational organisations or not. Accordingly, it was observed that the effect sizes of the research with participants composed of the employees of educational organisations are significantly higher than that of research with participants other than those of educational organisations. Aycan (2006) claimed that paternalistic leadership ensures a family atmosphere in the working environment and enables the employees to establish close relations with each other. In organisational life, the relations of employees with each other in the business environment are regarded as one of the main determinants of attitudes and behaviours towards the leader and the organisation (Nahrgang et al., 2009). In terms of educational organisations, it was asserted that the constant interaction of school administrators with the teachers shapes teachers' ideas and attitudes about the school (Alev, 2020). Therefore, the effectiveness and quality of organisations such as schools might be ensured through positive relations and interactions to be established among the employees (Korkmaz, 2005). Accordingly, higher effect sizes among the employees of educational organisations than other sample groups might be explained by the intensity of paternalistic behaviours such as interaction, communication, support and helpfulness in educational organisations. In organisations with great and extensive human resources, individuals might need each other and interact more. Therefore, differences in the effect sizes of the research according to gender and marital status according to the sample group might be considered an expected result.

The results of this meta-analysis should be addressed by considering certain limitations. The most important limitation of this research is that it only covers the previous research carried out in Türkiye. Therefore, the research results might rather be generalised for Türkiye. Another limitation is that the analyses in the research were made over the overall scores of the scales instead of the dimensions of the scales. In other words, studies not reporting the overall scores of the paternalistic leadership scale were not included in this meta-analysis study. In the research, carrying out the moderator analyses only with categorical variables might be considered another limitation. Against these limitations, several suggestions might be made to the practitioners and researchers. It may be useful for organisation leaders to help and support their female employees more in their work, to display ethical behaviours that will embrace everyone and create a family atmosphere without discriminating between married or single employees in the organisation, and to demonstrate leadership behaviours that are far from oppressive authoritarian behaviours. Researchers might be suggested to examine the effects of variables other than gender and marital status on the perception of paternalistic leadership, to analyse the effect size of the paternalistic leadership scale according to dimensions, to include studies in the international literature, and to include continuous variables in moderator analyses.

Declaration of Conflicting Interests and Ethics

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

Orcid

Mehmet Sabir Çevik https://orcid.org/0000-0002-8814-4747

REFERENCES

Sources marked with an asterisk (*) indicate studies included in the meta-analysis.

*Abacı, Ş. (2020). *The effects of paternalist leadership behaviors of managers on employees business perceptions: A study in the textile industry* [Unpublished master's thesis]. İstanbul Gelişim University.

Açıkel, C. (2009). Meta analiz ve kanıta dayalı tıptaki yeri [Meta-analysis and its place in evidence-based medicine]. *Klinik Psikofarmoloji Bülteni, 19*(2), 164-172. https://search.trdizin.gov.tr/tr/yayin/detay/100090/

Afsar, B., & Rehman, M. (2015). The relationship between workplace spirituality and innovative work behavior: The mediating role of perceived person–organization fit. *Journal of Management, Spirituality & Religion, 12*(4), 329-353. https://doi.org/10.1080/14766086.2015.1060515

Agich, G. (2003). *Dependence and autonomy in old age: An ethical framework for long-term care*. Cambridge University Press.

*Ağalday, B. (2017). *The relationship between primary school principals' paternalistic leadership behaviours and teachers' organizational creativity and organizational dissent levels* [Unpublished doctoral dissertation]. Dicle University.

Ağalday, B., & Dağlı, A. (2021). The investigation of the relations between paternalistic leadership, organizational creativity and organizational dissent. *Research in Educational Administration & Leadership, 6*(4), 748-794. https://doi.org/10.30828/real/2021.4.1

Alev, S. (2020). Okullarda örgütsel sinizmin yordayıcısı olarak lider-üye etkileşimi [Leader-member interaction as a predictor of organizational cynicism in schools]. *Trakya Eğitim Dergisi, 10*(2), 347-360. https://doi.org/10.24315/tred.618955

Anwar, H. (2013). Impact of paternalistic leadership on employees outcome a study on the banking sector of Pakistan. *IOSR Journal of Business and Management, 7*(6), 109-115. https://doi.org/10.9790/487X-076109115

*Arslan, Ö. (2016). *The correlation between school directors' paternalist leadership level and teachers' organisational cynism level* [Unpublished master's thesis]. Uşak University.

Aycan, Z. (2001). Paternalizm: Yönetim ve liderlik anlayışına ilişkin üç görgül çalışma [Paternalism: Three empirical studies on management and leadership]. *Yönetim Araştırmaları Dergisi, 1*(1), 11-31. http://yad.baskent.edu.tr/files/2001_cilt_1_1.pdf

Aycan, Z. (2006). Paternalism: Towards conceptual refinement and operationalization. In K.S. Yang, K.K. Hwang & U. Kim (Eds.), *Scientific advances in indigenous psychologies: Empirical, philosophical, and cultural contributions* (pp. 445-466). Springer.

Aycan, Z., Kanungo, R., Mendonca, M., Yu, K., Deller, J., Stahl, G., & Kurshid, A. (2000). Impact of culture on human resource management practices: A 10-country comparison. *Applied Psychology, 49*(1), 192-221. https://doi.org/10.1111/1464-0597.00010

Aycan, Z., Schyns, B., Sun, J.M., Felfe, J., & Saher, N. (2013). Convergence and divergence of paternalistic leadership: A cross-cultural investigation of prototypes. *Journal of International Business Studies*, 44, 962-969. https://doi.org/10.1057/jibs.2013.48

*Aydınoğlu, N. (2020). *Investigation of the effects of authentic and paternalist leadership behavior of administrators on teachers' motivation, job satisfaction and organizational* *commitment (Ankara province private schools example)* [Unpublished doctoral dissertation]. İstanbul Gelişim University.

Bedi, A. (2020). A meta-analytic review of paternalistic leadership. Applied Psychology, *69*(3), 960-1008. https://doi.org/10.1111/apps.12186

*Bilici, H.F. (2017). *Burnout, work engagement, turnover, paternal leadership and a research* [Unpublished master's thesis]. İstanbul Arel University.

Bing, S. (2004). *Sun Tzu was a sissy: Conquer your enemies, promote your friends, and wage the real art of war*. Harper Collins.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., & Rothstein, H.R. (2013). *Introduction to meta-analysis*. John Wiley & Sons, Ltd.

Britton, D.M. (2000). The epistemology of the gendered organization. *Gender and Society*, *14*(3), 418-434. https://doi.org/10.1177/089124300014003004

*Burgazlıoğlu, F. (2022). *The effect of y generation employees' paternalist leadership perception on organizational commitment msc thesis* [Unpublished master's thesis]. İstanbul Arel University.

Cansoy, R., Polatcan, M., & Parlar, H. (2020). Paternalistic school principal behaviours and teachers' participation in decision making: The intermediary role of teachers' trust in principals. *Research in Educational Administration & Leadership*, *5*(2), 553-584. https://doi.org/10.30828/real/2020.2.8

Card, N.A. (2011). *Applied meta-analysis for social science research: Methodology in the social sciences.* Guilford.

Cerit, Y. (2013). Paternalist liderlik ile öğretmenlere yönelik yıldırma davranışları arasındaki ilişki [The relationship between paternalistic leadership and mobbing behaviors towards teachers]. *Kuram ve Uygulamada Eğitim Bilimleri Dergisi*, *13*(2), 839-851. https://search.trdizin.gov.tr/tr/yayin/detay/145539/

*Cerit, Y., Özdemir, T., & Akgün, N. (2011). Sınıf öğretmenlerinin okul müdürlerinin paternalist liderlik davranışları sergilemelerini istemeye yönelik görüşlerinin bazı demografik değişkenler açısından incelenmesi [Examining the views of classroom teachers on asking school principals to exhibit paternalistic leadership behaviors in terms of some demographic variables]. *AİBÜ Eğitim Fakültesi Dergisi, 11*(1), 87-99. https://dergipark.org.tr/tr/download/article-file/16836

Chamundeswari, S. (2013). Job satisfaction and performance of schoolteachers. *International Journal of Academic Research in Business and Social Sciences*, *3*(5), 420-428. https://hrmars.com/papers_submitted/9599/job-satisfaction-and-performance-of-school-teachers.pdf

Cheng, B.S., Boer, D., Chou, L.F., Huang, M.P., Yoneyama, S., Shim, D., Sun, J.M., Lin, T. T., Chou, W.J., & Tsai, C.Y. (2013). Paternalistic leadership in four east asian societies generalizability and cultural differences of the triad model. *Journal of Cross-Cultural Psychology, 45*(1), 82-90. https://doi.org/10.1177/0022022113490070

Cheng, B.S., Chou, L.F., Wu, T.Y., Huang, M.P., & Farh, J.L. (2004). Paternalistic leadership and subordinate responses: Establishing a leadership model in Chinese organizations. *Asian Journal of Social Psychology*, *7*(1), 89-117. https://doi.org/10.1111/j.1467-839X.2004.00137.x

Chu, P.C., & Hung, C.C. (2009). The relationship of paternalistic leadership and organizational citizenship behavior: The mediating effect of upward communication. *Journal of Human Resource and Adult Learning, 5(2),* 66-73.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155-159. https://doi.org/10.1037/0033-2909.112.1.155

Cooper, H., Hedges, L.V., & Valentine, J.C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2. bs.). Sage.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and metaanalysis*. Routledge.

Dağlı, A., & Ağalday, B. (2017). Developing A headmasters' paternalistic leadership behaviours scale in Turkey. *Journal of Education and Practice*, *8*(30), 190-200. https://www.researchgate.net/publication/324063247_Developing_a_Headmasters'_Paternalistic_Leadership_Behaviours_Scale_in_Turkey

*Dağlı, A., & Ağalday, B. (2018). Okul müdürlerinin paternalist liderlik davranışlarının incelenmesi [Examination of school principals' paternalistic leadership behaviors]. *Elektronik Sosyal Bilimler Dergisi, 17*(66), 518-534. https://doi.org/10.17755/esosder.341663

Dedahanov, A.T., Bozorov, F., & Sung, S. (2019). Paternalistic leadership and innovative behavior: Psychological empowerment as a mediator. *Sustainability, 11*(6), 1-14. https://doi.org/10.3390/su11061770

Deeks, J.J., Higgins, J.P.T. & Altman, D.G. (2008). Analysing data and undertaking meta-analyses. J.P.T. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions* içinde (s. 243-296). John Wiley & Sons.

*Delice, A. (2020). *The relationship between the administrators' paternalistic leadership characteristics and the effectiveness of schools (Kahramanmaraş sample)* [Unpublished master's thesis]. Sütçü İmam University.

Demirer, P. (2012). *Is paternalistic leadership empowering: a contingency framework* [Unpublished master's thesis]. Koç University.

Dinçer, S. (2014). *Applied meta-analysis in educational sciences. Pegem Academy Publishing.*

Drost, E.A., & Von Glinow, M.A. (1998). Leadership behavior in Mexico: Etic philosophies/emic practices. *Research in International Business and International Relations, 7,* 3-28.

Durmaz, C. (2019). *The moderator effect of individualism-collectivism and the mediating effect of mobbing on the relationship between paternalistic leadership and organizational cynicism* [Unpublished doctoral dissertation]. Hacettepe University.

*Dursun, İ.E. (2019). *The effect of paternalistic leadership behaviors of school principals on creating school culture* [Unpublished master's thesis]. Sabahattin Zaim University.

Ekmen, F., & Okçu, V. (2021). The relation between paternalistic leadership behaviors of school administrators and pre-school teachers job satisfaction. *European Journal of Education Studies, 8*(6), 142-164. https://doi.org/10.46827/ejes.v8i6.3776

Erben, G.S., & Güneşer, A.B. (2008). The relationship between paternalistic leadership and organizational commitment: Investigating the role of climate regarding ethics. *Journal of Business Ethics, 82*(4), 955-968. https://doi.org/10.1007/s10551-007-9605-z

Farh, J.L., & Cheng, B.S. (2000). Paternalistic leadership in Chinese organizations: A cultural analysis. *Indigenous Psychological Research in Chinese Societies*, *13*, 127-80. https://doi.org/10.1057/9780230511590_5

Field, A.P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology, 63*(3), 665-694. https://doi.org/10.1348/000711010X502733

Fikret-Paşa, S. (2000). Leadership characteristics in the Turkish environment. In Z.Aycan (Ed.), *Management, leadership and human resources practices in Turkey* (pp. 225-241), Türk Psikologları Derneği.

Gelfand, M.J., Erez, M., & Aycan, Z. (2007). Cross-cultural organizational behavior. *Annual Reviews of Psychology*, 58, 479-514. https://doi.org/10.1146/annurev.psych.58.110405.085559

Göncü, A., Aycan, Z., & Johnson, R.E. (2014). Effects of paternalistic and transformational leadership on follower outcomes. *International Journal of Management and Business, 5*(1), 36-58. https://doi.org/10.1002/9781118785317.weom060156

Gürlek, M., Yeşiltaş, M., Tuna, M., Kanten, P., & Çeken, H., (2020). Paternalistic leadership and organizational identification: The mediating role of forgiveness climate. *International Journal of Hospitality and Tourism Administration*, *1*(1), 1-29. http://acikerisim.mu.edu.tr/xmlui/bitstream/handle/20.500.12809/6271/%C3%87eken.pdf?sequence=1&isAllowed=y

*Hatipoğlu, Z., Akduman, G., & Demir, B. (2019). Babacan liderlik tarzının çalışan görev performansı ve duygusal bağlılık üzerindeki etkisi [The effect of paternalistic leadership style on employee task performance and emotional commitment]. *İşletme Araştırmaları Dergisi*, *11*(1), 279-292. https://doi.org/10.20491/isarder.2019.599

Hayek, M., Novicevic, M.M., Humphreys, J.H., & Jones, N. (2010). Ending the denial of slavery in management history: Paternalist leadership of Joseph Emory Davis. *Journal of Management History*, *16*(3), 367-379. https://doi.org/10.1108/17511341011051252

Higgins, J.P.T., & Thompson, S.G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*, 1539-1558. https://doi.org/10.1002/sim.1186

Hofstede, G.H. (2006). What did globe really measure? Researchers' minds versus respondents' minds. *Journal of International Business Studies, 37*(6), 882-896. https://doi.org/10.1057/palgrave.jibs.8400233

House, R.J., Hanges, P.J., Javidan, M., Dorfman, P.W., & Gupta, V. (2004). *Culture, leadership, & organizations: The GLOBE study of 62 societies.* Sage Publications.

Huse, M., & Mussolino, D. (2008). Paternalism and governance in family firms. *ICSB World Conference*, June 22-25, Halifax, Nova Scotia, Canada.

*İnceöz, S., & Uslu, T. (2022). Paternalist, açık ve ilişki odaklı liderlik tarzlarının, çalışanların kurumsal yönetişim algıları ile ilişkilerinin incelenmesi [Examination of the relationship between paternalistic, open and relationship-oriented leadership styles and employees' perceptions of corporate governance]. *Sosyal, Beşeri ve İdari Bilimler Dergisi, 5*(12), 1690-1713. https://www.sobibder.org/index.php/sobibder/article/view/352

Jackson, T. (2016). Paternalistic leadership: Themissing link in cross-cultural leadership studies? *International Journal of Cross-Cultural Management, 16*(1), 3-7. https://doi.org/10.1177/1470595816637701

*Kara, E., Kaya, A., Başboğa, M.İ., Güvel, Ş., Çelik, C., & Koçak, B. (2020). Paternalist liderlik ve işten ayrılma niyeti üzerine bir araştırma [A research on paternalistic leadership and turnover intention]. *BMIJ, 8*(4): 118-138. https://doi.org/10.15295/bmij.v8i4.1710

*Karşu Cesur, D. (2015). *The relationship between paternalistic leadership and organizational culture: The case of Sakarya University* [Unpublished master's thesis]. Sakarya University.

*Kılıç, E. (2019). *With the paternalist leadership levels of school managers the relationship between teachers' perceptions of organizational support* [Unpublished master's thesis]. Uşak University.

*Koç, E. (2019). *Investigation of the relationship between job satisfaction of the employees in provincial directorate of youth and sports and paternalist leadership* [Unpublished master's thesis]. Marmara University.

*Korkmaz, F. (2018). *The mediating role of employee's work engagement in the effect on organizational identification of paternalistic leadership behaviour a comperative analysis between public and private sector* [Unpublished doctoral dissertation]. Kırıkkale University.

Korkmaz, F., Gökdeniz, İ., & Zorlu, K. (2018). Paternalist liderlik davranışının örgütsel özdeşleşme üzerindeki etkisinde çalışanların işe tutkunluk düzeylerinin aracılık rolü [The mediating role of employees' work engagement in the effect of paternalistic leadership behavior on organizational identification]. *İşletme Araştırmaları Dergisi, 10*(3), 950-973. https://doi.org/10.20491/isarder.2018.508

Korkmaz, M. (2005). Duyguların ve liderlik stillerinin öğretmenlerin performansı üzerinde etkisi [The effect of emotions and leadership styles on teachers' performance]. *Kuram ve Uygulamada Eğitim Yönetimi, 43,* 401-422. https://dergipark.org.tr/tr/pub/kuey/issue/10354/126786

Köksal, O. (2011). Paternalizm ile algılanan örgütsel adalet arasındaki ilişkinin tespitine yönelik bir araştırma [A research on the determination of the relationship between paternalism and perceived organizational justice] *Cumhuriyet Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, *12*(2), 159-170. https://search.trdizin.gov.tr/tr/yayin/detay/130930/

Kurt, İ. (2013). Babacan liderlik ile çalışanların işlerine yaratıcı katılım algıları arasındaki ilişkiyi araştırmaya yönelik bir çalışma [A study to investigate the relationship between paternalistic leadership and employees' perceptions of creative participation in their work]. *Sosyal ve Beşerî Bilimler Dergisi*, *5*(1), 321-330. https://dergipark.org.tr/tr/download/article-file/117364

Lee, J.Y., Jang, S.H., & Lee, S.Y. (2018). Paternalistic leadership and knowledge sharing with outsiders in emerging economies: Based on social exchange relations within the China context. *Personnel Review, 47*(5), 1094-1115. https://doi.org/10.1108/PR-03-2017-0068

Liang, S.K., Ling, H.C., & Hsieh, S.Y. (2007). The mediating effects of leader-member exchange quality to influence the relationship between paternalistic leadership and organizational citizenship behaviors. *Journal of American Academy of Business, 10*(2), 127-137.

Liao, S., Widowati, R., Hu, D., & Tasman, L. (2017). The mediating effect of psychological contract in the relationships between paternalistic leadership and turnover intention for foreign workers in Taiwan. *Asia Pacific Management Review, 22*(2), 80-87. https://doi.org/10.1016/j.apmrv.2016.08.003

Lin, C.P., Lin, M.Z., & Li, Y.B. (2015). An empirical study on the effect of paternalistic leadership on employees' voice behaviors–the intermediary role of psychological empowerment. *Journal of Interdisciplinary Mathematics, 18*(6), 789-810. https://doi.org/10.1080/09720502.2015.1108089

Lipsey, M.W., & Wilson, D.B. (2001). *Practical meta analysis.* SAGE Publications.

Makowski, D., Piraux, F., & Brun, F. (2019). *From experimental network to meta-analysis: Methods and applications with R for agronomic and environmental sciences.* Springer.

Martinez, P.G. (2003). Paternalism as a positive form of leader subordinate exchange. *Management Research*, *1*(3), 227-242. https://doi.org/10.1108/15365430380000529

*Mert, P., & Özgenel, M. (2020). A relational research on paternalist leadership behaviors perceived by teachers and teachers' performance. *Educational Policy Analysis and Strategic Research*, *15*(2), 41-60. https://doi.org/10.29329/epasr.2020.251.3

Mete, Y.A., & Serin, H. (2015). Okul yöneticilerinin babacan liderlik davranışı ile öğretmenlerin örgütsel vatandaşlık ve örgütsel sinizm davranışları arasındaki ilişki [The relationship between school administrators' fatherly leadership behavior and teachers' organizational citizenship and organizational cynicism behaviors]. *Hasan Âli Yücel Eğitim Fakültesi Dergisi, 12*(2), 147-159. https://dergipark.org.tr/tr/pub/iuhayefd/issue/8803/110083

Miles, M.B., & Huberman, A.M. (2002). *The qualitative researcher's companion*. Sage Publications.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), 1-6. https://doi.org/10.1371/journal.pmed.1000097

Mulla, Z.R., & Krishnan, V. (2012). Effects of beliefs in Indian philosophy: Paternalism and citizenship behaviors. *Great Lakes Herald*, *6*(2), 26-35. https://www.researchgate.net/pu blication/234065950_Effects_of_Beliefs_in_Indian_Philosophy_Paternalism_and_Citizenship_Behaviors

Mullen, B., Muellerleile, P., & Bryant, B. (2001). Cumulative meta-analysis: A consideration of indicators of sufficiency and stability. *Personality and Social Psychology Bulletin, 27*(11), 1450-1462. https://doi.org/10.1177/01461672012711006

Mumford, M.D., Ginamaire, M.S., Blaine, G., & Jill, M.S. (2002). Leading creative people: Orchestrating expertise and relationships. *Leadership Quarterly. 13*(6), 705-750. https://doi.org/10.1016/S1048-9843(02)00158-3

Mussolino, D., & Calabro, A. (2014). Paternalistic leadership in family firms: Types and implications for intergenerational succession. *Journal of Family Business Strategy, 5*(2), 197-210. https://doi.org/10.1016/j.jfbs.2013.09.003

Nahrgang, J., Morgeson, F., & Ilies, R. (2009). The development of leader-member exchanges: Exploring how personality and performance influence leader and member relationships over time. *Organizational Behavior and Human Decision Processes*, *108*, 256-266. https://doi.org/10.1016/j.obhdp.2008.09.002

*Nal, M. (2018). *An analysis of the relationship between health administrators' paternalistic leadership behavior, employee job satisfaction and perceptions of organizational justice* [Unpublished doctoral dissertation]. Marmara University.

Nigama, K., Selvabaskar, S., Surulivel, S.T., Alamelu, R., & Joice, D.U. (2018). Job satisfaction among school teachers. *International Journal of Pure and Applied Mathematics*, *119*(7), 2645-255. https://acadpubl.eu/jsi/2018-119-7/articles/7c/80.pdf

*Özgenel, M., & Canuylasu, R. (2021). Okul müdürlerinin paternalist liderlik davranışlarının örgütsel mutluluğa etkisi [The effect of school principals' paternalistic leadership behaviors on organizational happiness]. *Eğitim ve Teknoloji*, *3*(1), 14-31. https://doi.org/10.26677/TR1010.2020.361

*Özgenel, M., & Dursun, İ.E. (2020). Okul müdürlerinin paternalist liderlik davranışlarının okul kültürüne etkisi [The effect of school principals' paternalistic leadership behaviors on school culture]. *Sosyal, Beşeri ve İdari Bilimler Dergisi*, *3*(4), 284-302. https://doi.org/10.26677/TR1010.2020.361

Patton, M.Q. (2002). *Qualitative research and evaluation methods*. Sage Publications.

Pellegrini, E.K., & Scandura, T.A. (2006). Leader-member exchange (LMX), paternalism, and delegation in the Turkish business culture: An empirical investigation. *Journal of International Business Studies*, 37, 264-79. https://doi.org/10.1057/palgrave.jibs.8400185

Pellegrini, E.K., Scandura, T.A., & Jayaraman, V. (2010). Cross-cultural generalizability of paternalistic leadership: An expansion of leader-member exchange theory. *Group & Organization Management*, *35*(4), 391-420. https://doi.org/10.1177/1059601110378456

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences.* Blackwell Publishing.

Pigott, T.D., & Polanin, J.R. (2020). Methodological guidance paper: High-quality meta-analysis in a systematic review. *Review of Educational Research, 90*(1), 24-46. https://doi.org/10.3102/0034654319877153

Rothstein, H.R., Sutton, A.J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments.* John Wiley & Sons, Ltd.

Salminen-Karlsson, M. (2015). Expatriate paternalistic leadership and gender relations in small European software firms in India. *Culture and Organization*, *21*(5), 409-426. https://doi.org/10.1080/14759551.2015.1068776

*Sarı, T. (2021). *The relationship between school administrators 'paternalist leadership behavior and teachers' job satisfaction* [Unpublished master's thesis]. Ondokuz Mayıs University.

*Saylık, A. (2017). *The relationship between paternalistic leadership behaviours of school principals and culture dimensions of Hofstede* [Unpublished doctoral dissertation]. Ankara University.

Saylık, A., & Aydın, İ. (2020). Okul müdürlerinin paternalist liderlik davranışları ölçeğinin geliştirilmesi: Geçerlik ve güvenirlik çalışması [Development of The Paternalist Leadership Behavior Scale of School Principals: Validity and Reliability Study]. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, 13*(1), 273-300. https://doi.org/10.30964/auebfd.631892

Shi, X., Yu, Z., & Zheng, X. (2020) Exploring the relationship between paternalistic leadership, teacher commitment, and job satisfaction in Chinese schools. *Frontiers in Psychology, 11*, 1-12. https://doi.org/10.3389/fpsyg.2020.01481

Soylu, S. (2011). Creating a family or loyaltybased framework: The effects of paternalistic leadership on workplace bullying. *Journal of Business Ethics, 99*(2), 217-231. https://doi.org/10.1007/s10551-010-0651-6

Stahl, M.J. (2007). The influential leader. *Leader to Leader*, 46, 49-54. https://doi.org/10.1002/ltl.257

Stewart, D.W., & Kamins, M.A. (2001). Developing a coding scheme and coding study reports. M.W. Lipsey, & D.B. Wilson (Eds.), *Practical meta-analysis* (pp. 73-90). Sage.

Suber, P. (1999). Paternalism. In Christopher Berry Gray (Ed.), *Philosophy of law: An encyclopedia* (pp. 632-635). Garland Pub. II.

Sun, J.M., & Wang, B. (2009). Servant leadership in China: Conceptualization and measurement. In W.H. Mobley, and Y.W. Ming Li (Eds.), *Advances in global leadership* (pp. 321-344). Emerald Group Publishing Limited.

Şen, S., & Yıldırım, İ. (2020). *CMA ile meta analiz uygulamaları.* Anı Yayıncılık.

*Taşdemir, İ., & Atalmış, E.H. (2021). Okul müdürlerinin paternalist liderlik davranışları ile yaratıcı liderlik özellikleri arasındaki ilişkinin incelenmesi [Examining the relationship between school principals' paternalistic leadership behaviors and creative leadership characteristics]. *Sakarya Üniversitesi Eğitim Fakültesi Dergisi, 21*(1), 84-103. https://dergipark.org.tr/tr/pub/sakaefd/issue/62657/944598

Türesin Tetik, H., & Köse, S. (2015). Örgüt çalışanlarının paternalistik liderlik algıları ve öğrenilmiş güçlülük düzeyleri arasındaki ilişkilerin incelenmesi [Investıgatıon of the relatıonshıp between perceptıons of paternalıstıc leadershıp and learned resourcefulness levels of the employees]. *Uluslararası Yönetim İktisat ve İşletme Dergisi, 11*(26), 29-56. https://doi.org/10.17130/ijmeb.2015.11.26.352

Üstün, U., & Eryılmaz, A. (2014). Etkili araştırma sentezleri yapabilmek için bir araştırma yöntemi: Meta-analiz [A research methodology to conduct effective research syntheses: Meta-Analysis]. *Eğitim ve Bilim, 39*(174), 1-32. https://doi.org/10.15390/EB.2014.3379

Violato, C. (2019). *Assessing competence in medicine and other health professions.* CRC Press.

Wang, A.C., & Cheng, B.S. (2010). When does benevolent leadership lead to creativity? The moderating role of creative role identity and job autonomy. *Journal of Organizational Behavior, 31*(1), 106-121. https://doi.org/10.1002/job.634

Westwood, R.I. (1997). Harmony and patriarchy: The cultural basis for paternalistic headship among the overseas Chinese. *Organization Studies, 18*, 445-480. https://doi.org/10.1177/017084069701800305

Wu, M., Huang, X., Li, C., & Liu, W. (2011). Perceived interactional justice and trust-in supervisor as mediators for paternalistic leadership. *Management and Organization Review, 8*(1), 97-121. https://doi.org/10.1111/j.1740-8784.2011.00283.x

Yeh, H.R., Chi, H.K., Chiou, C.Y. (2008). The influences of paternalistic leadership, job stress, and organizational commitment on organizational performance: An empirical of policeman in Taiwan. *The Journal of International Management Studies, 3*(2), 85-91.

http://nhuir.nhu.edu.tw/bitstream/987654321/27172/1/The+Influences+of+Paternalistic.pdf

Yukl, G. (2008). *Leadership in organizations*. Prentice Hall.

Zhang, Y., Huai, M.Y., & Xie, Y.H. (2015). Paternalistic leadership and employee voice in China: A dual process model. *The Leadership Quarterly*, 26, 25-36. https://doi.org/10.1016/j.leaqua.2014.01.002

Zheng, X., Shi, X., & Liu, Y. (2020). Leading teachers' emotions like parents: Relationships between paternalistic leadership, emotional labor and teacher commitment in China. *Frontiers in Psychology*, *11*, 1-9. https://doi.org/10.3389/fpsyg.2020.00519

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 3, 532–543

https://doi.org/10.21449/ijate.1295077

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

The validity and reliability of the Juvenile Arthritis Functional Assessment Report (JAFAR) in children/adolescents with Juvenile Idiopathic Arthritis: The Turkish version study

Merve Bali[1], Elif Gür Kabul[2,*], Bilge Başakcı Çalık[1], Gülçin Otar Yener[3], Zahide Ekici Tekin[3], Selçuk Yüksel[3]

[1]Pamukkale University, Faculty of Physiotherapy and Rehabilitation, Denizli, Türkiye
[2]Uşak University, Faculty of Health Sciences, Physiotherapy and Rehabilitation, Uşak, Türkiye
[3]Pamukkale University, Department of Pediatric Rheumatology, Faculty of Medicine, Denizli, Türkiye

ARTICLE HISTORY

Received: May 10, 2023
Revised: July 10, 2023
Accepted: July 27, 2023

Keywords:
Juvenile Idiopathic Arthritis,
Children,
Adolescence.

Abstract: The aim of the study was to describe the validity and reliability of the Turkish version of Juvenile Arthritis Functional Assessment Report (JAFAR) in children/adolescents with Juvenile Idiopathic Arthritis (JIA). Sixty-nine children/adolescents with JIA were included in the study. JAFAR(TR)-Child and Parent forms were applied to the patients with JIA and to their parents for test retest at one-week intervals, the patients did not receive additional treatment and his/her pharmacological treatment did not change for that week. Test-retest reliability was evaluated by intraclass correlation coefficient (ICC), and internal consistency reliability of multi-item subscales was evaluated by calculating Cronbach's alpha coefficient. Correlations between JAFAR(TR)-Child and Parent with the Pediatric Quality of Life Inventory 3.0. Module Arthritis (PedsQL), the Childhood Health Assessment Questionnaire (CHAQ), and the Juvenile Arthritis Disease Activity Score (JADAS) were evaluated to determine construct validity. The ICC value for the test/retest reliability of JAFAR(TR)-Child was 0.963 and of JAFAR(TR)-Parent was 0.576. JAFAR(TR)-Child total score had low to moderate correlations with PedsQL Child (r=-0.34; *p*=0.004), CHAQ (r=0.40; *p*=0.001), and JADAS total score (r=0.42; *p*=0.000). JAFAR(TR)-Parent total score had moderate to high correlations with PedsQL Parent (r=-0.55; *p*=0.000), CHAQ (r=0.72; *p*=0.000) and JADAS total score (r=0.53; *p*= 0.000). The Turkish version of JAFAR was found to be clinically valid and reliable in JIA.

1. INTRODUCTION

Juvenile Idiopathic Arthritis (JIA) is a chronic autoimmune disease, in which arthritis occurs in one or more joints below 16 years old at least 6 weeks (Petty et al., 2004). Joint pain, muscle atrophy, weakness, contracture, joint swelling, and movement-related abnormalities are seen in the symptoms of individuals diagnosed with JIA (Hansmann et al., 2015).

Knowing to what extent rheumatic diseases, which begin to show their effects in childhood, affect the child's functionality in daily activities is very important information in the management of these children's diseases. JIA is the most common cause of functional disability

*CONTACT: Elif Gür Kabul ✉ elifgur1988@hotmail.com Uşak University, Faculty of Health Sciences, Physiotherapy and Rehabilitation, Uşak, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

in childhood. Therefore, early evaluation is important. Studies show that children/adolescents with JIA have less functional ability, physical activity participation, and fitness compared to those of the healthy peers. This inadequacy also causes physical disability in JIA children (Henderson et al., 1995; Takken et al., 2002; Klepper, 2008; 2003; Lelieveld et al., 2008; Tarakci et al., 2011).

Questionnaires are widely used to assess functional status in children/adolescents with JIA. The questionnaires such as The Juvenile Arthritis Functional Assessment Scale (JAFAS), Juvenile Arthritis Self-Report Index, The Juvenile Arthritis Multidimensional Assessment Report (JAMAR), and Childhood Health Assessment Questionnaire (CHAQ) are generally used for evaluation. CHAQ demonstrates high internal reliability and test-retest reliability (Cronbach's alpha≥0.98; r=0.8) (Singh et al., 1994) and consists of 8 subsections (dressing, reach, eating, arising, walking, grip, hygiene, and activities) and 30 questions. However, some problems have been reported such as difficulty in understanding the questions during the evaluation and the same answers are always given by the patients after a while (Kisaarslan & Sözeri, 2016). The Juvenile Arthritis Functional Assessment Scale (JAFAS) is a scale focused on musculoskeletal function. The assessor looks at how long it takes to do the activities. The Juvenile Arthritis Self-Report Index is a two-part questionnaire consisting of 100 questions focused on physical activity. The fact that the clinical application of these two questionnaires takes a long time and that their Turkish validity and reliability has not been done creates a disadvantage. JAMAR, which has Turkish validity, consists of 15 questions. Physical function is evaluated in the first question and in the other questions, and a general evaluation is made by examining pain, intensity, presence of painful or swollen joints, morning stiffness and duration, disease activity level, treatment content, and school problems; however, it has limitations in questioning functionality.

These questionnaires evaluating functionality are few in number, and only CHAQ and JAMAR have been validated and found reliable in Turkish (Tarakci et al., 2013; Demirkaya et al., 2018). However, the CHAQ and JAMAR alone are insufficient to assess actual functionality. Juvenile Arthritis Functional Assessment Report (JAFAR) is a functional questionnaire that covers assessment of physical function, aids/devices, help from others, and pain. JAFAR evaluates the ability to perform 23 physical functions (based on daily functional movements) without any help for the past week. Each item has a three-point Likert answer system (“0” all the time; “1” sometimes; “2” almost never). It also measures the severity of the child pain for child and her/his parent with a 10 cm line (10=Very Bad Pain, 0=No pain). JAFAR is a simple and convenient questionnaire for clinical studies that can be easily filled by the patient and comprehensively evaluates physical function (Howe et al., 1991). We aimed to examine the validity and reliability of the Turkish version of the Juvenile Arthritis Functional Assessment Report.

2. METHOD

2.1. Patients

Sixty-nine children/adolescents with JIA between the ages of 7-18 (29 boys, 40 girls; mean age=13.36 ± 2.97 years) followed by Pamukkale University Pediatric Rheumatology Clinic and Pediatric Rheumatology Physiotherapy and Rehabilitation Unit were included in the study. In the related literature, the sample size should be 3-10 times the number of scale items in scale studies (Cattell, 1978; Comrey & Lee, 1992; Tavşancıl, 2002, s. 5–6; Hair et al., 2009).

Inclusion criteria were diagnosed with JIA according to the criteria of International League of Associations for Rheumatology to be between the ages of 6-18 in order to be included in the study.

Exclusion criteria were (a) having another autoimmune disease, (b) having neurological disease, (c) presence of any orthopedic, cardiopulmonary problem that can affect functionality and daily living activities, (e) having a psychiatric illness that affects cooperation, and (f) having a history of orthopedic surgery in the last one year.

Approval that there was no ethical problem for the study was obtained from Pamukkale University at the meeting number of 18 dated 24.10.2019. All participants were informed verbally before participating in the study and consent forms were signed by the participants.

2.2. Procedures

A cross sectional study design was planned.

2.3. Clinical Data

All participants were evaluated by the investigator in approximately 40-45 minutes and a session. After the demographic information of the patients was recorded, the quality of life was evaluated with Pediatric Quality of Life Inventory 3.0. Module Arthritis (PedsQL), disability levels with Childhood Health Assessment Questionnaire (CHAQ), and disease activities with the Juvenile Arthritis Disease Activity Score (JADAS). JAFAR questionnaire was applied to children/adolescents with JIA and their parents for test retest at one-week intervals, the patients did not receive additional treatment and his/her pharmacological treatment did not change for that week.

2.3.1. *Translation and cultural adaptation of Juvenile Arthritis Functional Assessment Report (JAFAR)*

During the JAFAR cross-cultural adaptation process, previously recommended procedures were followed in five stages (Beaton et al., 2000; Wild et al., 2005). The JAFAR (TR) is presented in Appendix.

2.3.2. *Juvenile Arthritis Functional Assessment Report (JAFAR)*

Juvenile Arthritis Functional Assessment Report is a functional assessment criterion developed based on the children with JIA and parents of the children. JAFAR is valid for JIA. JAFAR consists of two forms, Juvenile Arthritis Functional Assessment Report for Children (JAFAR-C) and Juvenile Arthritis Functional Assessment Report for Parents (JAFAR-P). JAFAR evaluates the ability to perform physical functions without assistance with 23 items for the past week. Each item has a three-point Likert answer system (“0” all the time; “1” sometimes; “2” almost never). A lower score means better physical functionality. JAFAR also assesses whether aids/devices are used, whether help from others is needed, and child pain with a 10 cm line (10=Very Bad Pain, 0=No pain) (Howe et al., 1991).

2.3.3. *Pediatric Quality of Life Inventory 3.0. Module Arthritis (PedsQL)*

PedsQL 3.0 Arthritis Module was developed to evaluate the quality of life in children with rheumatic disease. PedsQL 3.0 Arthritis Module has “Pain and hurt”, “Daily activities”, “Treatment”, “Worry” and “Communication” subsections and consists of 22 items in total. Each item is evaluated from 0 to 4 (Never-0, Always-4). PedsQL 3.0 Arthritis Module has separate forms for children of different age groups and their parents (2-4 years old, 5-7 years old, 8-12 years old and 12-18 years old). In our study, 8-12 age and 12-18 age child and parent forms were used. As the score decreases, the quality of life decreases (Tarakci et al., 2013).

2.3.4. *Childhood Health Assessment Questionnaire (CHAQ)*

CHAQ evaluates functional abilities in children. The scale can be applied to all children between the ages of 6 months and 18 years. The CHAQ is composed of disability and discomfort indexes. CHAQ Disability Index consists of 30 questions and 8 subsections, including dressing, eating, reach, arising, walking, grip, hygiene, and activities. Calculation of

CHAQ Disability Index is based on all scores from 8 sections summed and divided by 8. CHAQ Discomfort Index assessed pain and global evaluation measured by two 0-100 mm visual analog scales. Higher score means more severe functional disability (Ozdogan et al., 2001).

2.3.5. *Juvenile Arthritis Disease Activity Score (JADAS)*

Juvenile Arthritis Disease Activity Score (JADAS) was evaluated for the disease activity for children (Consolaro et al., 2009). JADAS consists of four parts:

1. D-GAS (Doctor- Visual Analogue Scale),
2. H-GAS (Patient- Visual Analogue Scale),
3. Number of active joints (71, 27,10 joints): Active joint is defined as the presence of swollen joint and/or tender joint.
4. Evaluation of sedimentation between 0-10: SEDIM: ESR (mm/hour)-20/10.

If the sedimentation rate is 120 or higher, the score is considered 10.

JADAS is calculated by the arithmetic sum of four parts (Nordal et al., 2012). In our study, JADAS 27 was used. JADAS 27 includes cervical, elbows, wrists, 1-3 metacarpophalangeal joints, proximal interphalangeal joints, hip joints, knees, and ankles (Horneff & Becker, 2014).

2.4. Statistical Analysis

SPSS 25.0 software (IBM SPSS Statistics 25 software (Armonk, NY: IBM Corp) was used for the analyses. Categorical variables were shown as number and percent while continuous variables as mean ± Standard deviation (SD) and median (minimum – maximum values). The conformity of continuous numerical variables to the normal distribution was examined using the ShapiroWilk test. External construct validity was analyzed with Spearmanrho correlation coefficient. The internal consistency reliability was analyzed with the Cronbach's alpha coefficients. For intraclass correlation coefficient (ICC), two way mixed was used for test-retest reliability (ICC; <0.50=poor reliability, between 0.50 and 0.75: moderate reliability, between 0.75 and 0.90: good reliability, >0.90: excellent reliability). Statistical significance level was accepted as $p<0.05$.

3. RESULTS

The mean age of the patients in the study was 13.36 ± 2.97 years and 58% were girls and adolescents. The demographic data of the patients are summarized in Table 1. Descriptive information about the outcome measures is given in Table 2.

3.1. Construct validity

Descriptive data of JAFAR Parent and JAFAR Child total scores are given in Table 2.

Table 1. *Demographic data of patients with JIA***.**

	n	%	Med (IQR)	Min - Max
Age			13 (11 - 16)	8 – 18
Gender				
Girl	40	58.0		
Boy	29	42.0		
Diagnosis Age (year)			11 (7 - 14)	1 – 17
BMI			20.2 (17.61 - 23.55)	12.62 - 32.95

Table 2. *Descriptive data of the outcome measures.*

PedsQL Child	Mean± S.D.	Med (IQR)	Min - Max
Pain and hurt	70.65 ± 26.8	81.25 (50 - 93.73)	0 - 100
Daily activities	93.38 ± 16.46	100 (92.5 - 100)	10 - 100
Treatment	79.75 ± 20.34	85.7 (69.64 - 96.41)	28.5 - 100
Worry	76.58 ± 26.69	83.33 (62.47 - 100)	0 - 116.66
Communication	82.11 ± 21.46	91.6 (66.66 - 100)	33.3 - 100
Total	80.5 ± 15.44	85.87 (71.33 - 92.65)	43.21 - 100
PedsQL Parent	Mean± S.D.	Med (IQR)	Min - Max
Pain and hurt	70.74 ± 29.19	81.25 (53.13 - 100)	0 - 100
Daily activities	92.68 ± 17.4	100 (97.5 - 100)	10 - 100
Treatment	76.91 ± 20.9	78.57 (60.71 - 94.63)	32.14 - 100
Worry	69.91 ± 29.89	75 (50 - 100)	0 - 100
Communication	84.41 ± 23.44	100 (70.83 - 100)	16.66 - 100
Total	78.93 ± 17.02	81.13 (70.49 - 90.83)	37.49 - 100
CHAQ			
Dressing	0.58 ± 0.95	0 (0 - 1)	0 - 3
Eating	0.26 ± 0.63	0 (0 - 0)	0 - 3
Reach	0.54 ± 0.96	0 (0 - 1)	0 - 3
Arising	0.49 ± 0.8	0 (0 - 1)	0 - 3
Walking	0.39 ± 0.75	0 (0 - 1)	0 - 3
Grip	0.39 ± 0.81	0 (0 - 0.5)	0 - 3
Hygiene	0.45 ± 0.78	0 (0 - 1)	0 - 3
Activities	0.67 ± 1.02	0 (0 - 1)	0 - 3
Disability Index Total	0.47 ± 0.59	0.25 (0 - 0.69)	0 - 3
Pain	35.17 ± 26.25	40 (10 - 55)	0 - 90
Global Evaluation	36.54 ± 26.04	40 (10 - 60)	0 - 90
JADAS			
Total	9.23 ± 6.66	9 (3.5 - 13)	0 - 28
JAFAR			
Child 1. Evaluation Total	1.88 ± 2.85	1 (0 - 3)	0 - 12
Child 2. Evaluation Total (retest)	1.68 ± 2.64	0 (0 - 2.5)	0 - 12
Parent 1. Evaluation Total	3.74 ± 5.32	1 (0 - 6.5)	0 - 23
Parent 2. Evaluation Total (retest)	4.29 ± 14.03	1 (0 - 3)	0 - 100

PedsQL: Pediatric Quality of Life Inventory 3.0. Module Arthritis, CHAQ= Childhood Health Assessment Questionnaire JADAS= Juvenile Arthritis Disease Activity Score, JAFAR= Juvenile Arthritis Functional Assessment Report

3.2. External Validation

For the validity of the child and parent forms of JAFAR, the relationship between the JAFAR Child and Parent total score and subsections and total score of the PedsQL Child and Parent forms, subsections and total score of CHAQ Disability Index, Pain and Global Evaluation of CHAQ and the JADAS total score was examined and is given in Table 3.

JAFAR-Child total score had a significant negative correlation with pain and hurt subsection (*r*=-0.521; *p*=0.000) and total score (*r*=-0.347; *p*=0.004)) of PedsQL Child. PedsQL score approaching 100 means better quality of life, while JAFAR total score approaching zero means better physical functionality. For this reason, the negative correlation indicates that JAFAR is also suitable in the evaluation.

JAFAR-Child total score had a significant positive correlation with arising (*r*=0.475; *p*=0.000), walking (*r*=0.320; *p*=0.000), hygiene (*r*=0.305; *p*=0.011), activities (*r*=0.255; *p*=0.035) subsections and total score (*r*=0.401; 0.001) of CHAQ Disability Index, Pain of CHAQ (*r*=0.375; *p*=0.001), Global Evaluation of CHAQ (*r*=0.445; *p*=0.001), and JADAS total score (*r*=0.422; *p*=0.000). The higher the score in CHAQ, the higher the disability level, and the higher the score in JADAS, the higher the disease activity. For this reason, the positive correlation indicates that JAFAR is also suitable in the evaluation.

JAFAR-Parent total score had a significant negative correlation with all subsections (except for Worry) (*r*=-0.679/-0.370; *p*<0.05) and total score (*r*=-0.553; *p*=0.000) of PedsQL Parent.

JAFAR-Parent total score had a significant positive correlation with all subsections and total score of CHAQ Disability Index, Pain of CHAQ, Global Evaluation of CHAQ (*r*=0.723/0.320; *p*<0.05) and JADAS total score (*r*=0.539; *p*=0.000) (*p*<0.05).

Table 3. *Correlation between JAFAR-Child and JAFAR-Parent with PedsQL, CHAQ and JADAS questionnaires.*

	JAFAR-Child rho; *p*	JAFAR-Parent rho; *p*
PedsQL Child		
Pain and hurt	-0.521; 0.001	-0.673; 0.001
Daily activities	-0.198; 0.103	-0.468; 0.001
Treatment	-0.085; 0.489	-0.330; 0.006
Worry	-0.178; 0.143	-0.313; 0.009
Communication	-0.136; 0.265	-0.472; 0.001
Total	-0.347; 0.004	-0.652; 0.001
PedsQL Parent		
Pain and hurt	-0.480; 0.001	-0.679; 0.001
Daily activities	-0.314; 0.009	-0.396; 0.001
Treatment	-0.267; 0.027	-0.370; 0.002
Worry	-0.228; 0.059	-0.187; 0.125
Communication	-0.125; 0.306	-0.442; 0.001
Total	-0.372; 0.002	-0.553; 0.001
CHAQ		
Dressing	0.082; 0.504	0.462; 0.001
Eating	0.151;0.216	0.320; 0.007
Reach	0.135; 0.267	0.595; 0.001
Arising	0.475; 0.001	0.671; 0.001
Walking	0.320; 0.001	0.572; 0.001
Grip	0.160; 0.190	0.357; 0.003
Hygiene	0.305; 0.011	0.541; 0.001
Activities	0.255; 0.035	0.400; 0.001
Disability Index Total	0.401; 0.001	0.723; 0.001
Pain	0.375; 0.001	0.455; 0.001
Global Evaluation	0.445; 0.001	0.527; 0.001
JADAS		
Total	0.422; 0.001	0.539; 0.001

PedsQL: Pediatric Quality of Life Inventory 3.0. Module Arthritis, CHAQ= Childhood Health Assessment Questionnaire JADAS= Juvenile Arthritis Disease Activity Score, JAFAR= Juvenile Arthritis Functional Assessment Report

3.3. Internal Consistency Reliability

The internal consistency coefficient for the JAFAR-Parent pain was 0.659, the internal consistency coefficient for the JAFAR-Parent total score was 0.576, the internal consistency coefficient for the JAFAR-Child pain was 0.879, the internal consistency coefficient for the JAFAR-Child total score was 0.963, and the scale was found reliable (Table 4).

Table 4. *Test-retest reliability of JAFAR-Child and JAFAR-Parent.*

	ICC	95% CI of ICC Lower-upper	Reliability
JAFAR-Child Pain	0.879	0.804 – 0.925	Good
JAFAR-Child Total	0.963	0.94 – 0.977	Good
JAFAR-Parent Pain	0.659	0.449 – 0.789	Moderate
JAFAR-Parent Total	0.576	0.315 – 0.737	Moderate

JAFAR: Juvenile Arthritis Functional Assessment Report, CI: Confidence Interval; ICC: Intraclass correlation coefficient two-way mixed model-absolute agreement; Intraclass correlation coefficient values less than 0.50 indicate poor reliability, values between 0.50 and 0.75 indicate moderate reliability, values between 0.75 and 0.90 indicate good reliability, and values greater than 0.90 indicate excellent reliability.

4. DISCUSSION and CONCLUSION

The Turkish version of the JAFAR was found to be clinically valid and reliable for use in clinical evaluations and rehabilitation interventions in patients with JIA.

Determination of daily functional abilities of children/adolescents with juvenile chronic arthritis is of primary importance (Murray & Passo, 1995). Functional abilities of children/adolescents with JIA decreased as they were less likely to participate in social activities and tended to lead a more sedentary life (Gare et al., 1993).

CHAQ and JAMAR are widely used for functional evaluation in clinics in Türkiye. A meta-analysis study, examining the functional evaluation of children/adolescents with JIA in the Turkish population, emphasized that the options for functional assessment are limited (Kuntze et al., 2018). Kisaarslan et al. (2016), in their review of outcome measures at JIA, reported that the CHAQ has some problems such as difficulty in understanding the questions during the evaluation and the same answers are always given by the patients after a while. We think that other problems such as the inability to apply for the children's parents and younger children may encounter problems in answering because of the difficulty in understanding some of the CHAQ questions also make the CHAQ less adequate.

The JAMAR, another valid and reliable questionnaire in Turkish, consists of 15 questions. Since only the first question has 15 sub-parameters, it limits the practical evaluation of functional problems in the clinic (Demirkaya et al., 2018). However, JAFAR evaluates functional assessment in detail with 23 questions applied to both parents and children/adolescents with JIA (Howe et al., 1991). The quick, practical and meaningful evaluation is very important in JIA, because the evaluation of functionality guides the management as well as clinical diagnosis and treatment.

The internal consistency of JAFAR was quite good and found to be reliable. As a result of external validity analysis, the relationship between JAFAR-Child, JAFAR-Parent and CHAQ, PedsQL, JADAS, and their subsections was found to be moderately significant.

The limitation of this study is the inability to reach all children and adolescents with JIA diagnosed and followed-up in the clinic, since the evaluation part coincided with the COVID-19 pandemic period.

When the literature is examined, we see that JAFAR is not valid and reliable in any language other than English. Since JAFAR is a questionnaire that can evaluate the functional level quickly and easily, we believe that it will be beneficial in terms of evaluating the perspectives of children and adolescents with JIA and their families, determining the functional disabilities of their children and taking measures for this situation. Therefore, we recommend examining the validity and reliability of this questionnaire in languages other than English.

In conclusion, The JAFAR-TR scale is a valid and reliable outcome measure assessing the physical function of children/adolescents with JIA.

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, 24.10.2019, 60116787-020/75835.

Authorship Contribution Statement

Merve Bali: Collected data, wrote the article**. Elif Gür Kabul:** Design, collected data, statistical analyses. **Bilge Başakcı Çalık:** Design, commented on statistics, critical revision of the manuscript. **Gülçin Otar Yener:** Diagnosed patients and checked eligibility for inclusion criteria**. Zahide Ekici Tekin:** Diagnosed patients and checked eligibility for inclusion criteria**.** **Selçuk Yüksel:** Commented on statistics, critical revision of the manuscript.

Orcid

Merve Bali https://orcid.org/0000-0002-6955-9596
Elif Gür Kabul https://orcid.org/0000-0003-3209-1499
Bilge Başakcı Çalık https://orcid.org/0000-0002-7267-7622
Gülçin Otar Yener https://orcid.org/0000-0003-2575-6309
Zahide Ekici Tekin https://orcid.org/0000-0002-5446-667X
Selçuk Yüksel https://orcid.org/0000-0001-9415-1640

REFERENCES

Beaton, D.E., Bombardier, C., Guillemin, F., & Ferraz, M.B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine (Phila. Pa. 1976), 25*(24), 3186–3191. https://doi:10.1097/00007632-200012150-00014

Cattell, R.B. (1978). *The scientific use of factor analysis in behavioral and life sciences* (2nd edition). Plenum.

Comrey, A.L., & Lee, H.B. (1992). *A first course in factor analysis* (2nd edition). Lawrence Erlbaum.

Consolaro, A., Ruperto, N., Bazso, A., Pistorio, A., Magni-Manzoni, S., Filocamo, G., Malattia, C., Viola, S., Martini, A., Ravelli, A., & Paediatric Rheumatology International Trials Organisation. (2009). Development and validation of a composite disease activity score for juvenile idiopathic arthritis. *Arthritis and Rheumatism, 61*(5), 658-666. https://doi:10.1002/art.24516

Demirkaya, E., Ozen, S., Sozeri, B., Ayaz, N.A., Kasapcopur, O., Unsal, E., Makay, B.B., Barut, K., Fidanci, B.E., Simsek, D., Cakan, M., Consolaro, A., Bovis, F., Ruperto, N., & Paediatric Rheumatology International Trials Organisation (PRINTO). (2018). The Turkish version of the juvenile arthritis multidimensional assessment report (JAMAR). *Rheumatology International*, *38*(1), 395–402. https://doi:10.1007/s00296-018-3982-8

Gare, A.B., Fasth, A., & Wiklund, I. (1993). Measurement of functional status in juvenile chronic arthritis; evaluation of a Swedish version of the Child Health Assessment Questionnaire. *Clinical and Experimental Rheumatology*, *11*(5), 569-576.

Hair, J.F., Black, W., Babin, B., & Anderson, R. (2009). *Multivariate data analysis* (7th Edition). Upper Saddle River.

Hansmann, S., Benseler, S.M., & Kuemmerle-Deschner, J.B. (2015). Dynamic knee joint function in children with juvenile idiopathic arthritis (JIA). *Pediatric Rheumatology Online Journal*, *13*(1), 1–11. https://doi:10.1186/s12969-015-0004-1

Henderson, C.J., Lovell, D.J., Specker, B.L., & Campaigne, B.N. (1995). Physical activity in children with juvenile rheumatoid arthritis: Quantification and evaluation. *Arthritis Care and Research*, *8*(2), 114–119. https://doi:10.1002/art.1790080210

Horneff, G., & Becker, I. (2014). Definition of improvement in juvenile idiopathic arthritis using the juvenile arthritis disease activity score. *Rheumatology (Oxford)*, *53*(7), 1229–1234. https://doi:10.1093/rheumatology/ket470

Howe, S., Levinson, J., Shear, E., Hartner, S., McGirr, G., Schulte, M., Lovell D. (1991). Development of a disability measurement tool for juvenile rheumatoid arthritis. The juvenile arthritis functional assessment report for children and their parents. *Arthritis and Rheumatism*, *34*(7), 873–880. https://doi:10.1002/art.1780340713

Kisaarslan, A.P., & Sözeri, B. (2016). The outcome measures in Juvenile idiopathic arthritis: Review. *Türkiye Klinikleri Pediatri*, *25*(2), 101–109. https://doi:10.1002/art.11055

Klepper, S.E. (2003). Exercise and fitness in children with arthritis: Evidence of benefits for exercise and physical activity. *Arthritis and Rheumatism*, *49*(3), 435–443.

Klepper, S.E. (2008). Exercise in pediatric rheumatic diseases. *Current Opinion in Rheumatology*, *20*(5), 619–624. https://doi:10.1097/BOR.0b013e32830634ee

Kuntze, G., Nesbitt, C., Whittaker, J.L., Nettel-Aguirre, A., Toomey, C., Esau, S., Doyle-Baker, P.K., Shank, J., Brooks, J., Benseler, S., Emery, C.A. (2018). Exercise therapy in Juvenile Idiopathic Arthritis: A systematic review and meta-analysis. *Archives of Physical Medicine and Rehabilitation*, *99*(1), 178-193. https://doi:10.1016/j.apmr.2017.05.030

Lelieveld, O.T.H.M., Armbrust, W., van Leeuwen, M.A., Duppen, N., Geertzen, J.H.B., Sauer, P.J.J., van Weert, E. (2008). Physical activity in adolescents with juvenile idiopathic arthritis. *Arthritis and Rheumatism*, *59*(10), 1379–1384. https://doi:10.1002/art.24102

Murray, K.J., & Passo, M.H. (1995). Functional measures in children with rheumatic diseases. *Pediatric Clinics of North America*, *42*(5), 1127–1154. https://doi:10.1016/s0031-3955(16)40056-8

Nordal, E.B., Zak, M., Aalto, K., Berntson, L., Fasth, A., Herlin, T., Lahdenne, P., Nielsen, S., Peltoniemi, S., Straume, B., & Rygg, M. (2012). Validity and predictive ability of the juvenile arthritis disease activity score based on CRP versus ESR in a Nordic population-based setting. *Annals of the Rheumatic Diseases*, *71*(7), 1122-1127. https://doi:10.1136/annrheumdis-2011-200237

Ozdogan, H., Ruperto, N., Kasapçopur, O., Bakkaloglu, A., Arisoy, N., Ozen, S., Ugurlu, U., Unsal, E., Melikoglu, M., & Paediatric Rheumatology International Trials Organisation. (2001). The Turkish version of the Childhood Health Assessment Questionnaire (CHAQ) and the Child Health Questionnaire (CHQ). *Clinical and Experimental Rheumatology*, *19*(4), S158-162.

Petty, R.E., Southwood, T.R., Manners, P., Baum, J., Glass, D.N., Goldenberg, J., He, X., Maldonado-Cocco, J., Orozco-Alcala, J., Prieur, A.M., Suarez-Almazor, M.E., Woo, P., & International League of Associations for Rheumatology. (2004). International League of Associations for Rheumatology classification of juvenile idiopathic arthritis: Second revision, Edmonton, 2001. *The Journal of Rheumatology*, *31*(2), 390–392.

Singh, G., Athreya, B.H., Fries, J. F., & Goldsmith, D.P. (1994). Measurement of health status in children with juvenile rheumatoid arthritis. *Arthritis and Rheumatism*, *37*(12), 1761–1769. https://doi:10.1002/art.1780371209

Takken, T., Hemel, A., Van Der Net, J., & Helders, P.J.M. (2002). Aerobic fitness in children with juvenile idiopathic arthritis: A systematic review. *The Journal of Rheumatology, 29*(12), 2643–2647.

Tarakci, E., Baydogan, S.N., Kasapcopur, O., & Dirican, A. (2013). Cross-cultural adaptation, reliability, and validity of the Turkish version of PedsQL 3.0 Arthritis Module: A quality-of-life measure for patients with juvenile idiopathic arthritis in Turkey. *Quality of Life Research, 22*(3), 531–536. https://doi:10.1007/s11136-012-0180-0

Tarakci, E., Yeldan, I., Mutlu, E.K., Baydogan, S.N., & Kasapcopur, O. (2011). The relationship between physical activity level, anxiety, depression, and functional ability in children and adolescents with juvenile idiopathic arthritis. *Clinical Rheumatology, 30*(11), 1415–1420. https://doi:10.1007/s10067-011-1832-0

Tavşancıl, E. (2002). *Tutumların ölçülmesi ve SPSS ile veri analizi (1st edition).* Nobel publishing house.

Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., Erikson, P., & ISPOR Task Force for Translation and Cultural Adaptation. (2005). Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: Report of the ISPOR task force for translation and cultural adaptation. *Value in Health, 8*(2), 94-104. https://doi:10.1111/j.1524-4733.2005.04054.x

APPENDIX

JAFAR - Juvenile Arthritis Functional Assessment Report - Turkish version

JAFAR - Jüvenil Romatoid Artrit Fonksiyonel Değerlendirme Formu

7 Yaş ve Üstü Jüvenil Romatoid Artritli Çocuklar İçin

Hasta Adı-Soyadı:
Hasta Doğum Tarihi: Değerlendirme Tarihi:

1.Bölüm: Yetenek Ölçeği

Bu anket, çocuğunuzun hastalığının onun günlük yaşamdaki fonksiyonlarını nasıl etkilediğini öğrenmek amacıyla oluşturulmuştur.

Lütfen bu sayfanın arkasına herhangi bir yorum eklemekten çekinmeyin.

Son bir haftayı düşünerek çocuğunuzun yapabildiği aktivitelere göre uygun cevabı işaretleyin.

Lütfen bütün soruları cevaplayınız.

Geçtiğimiz hafta,	Her zaman	Bazen	Neredeyse hiç
1. Gömleğini askıdan almak	___	___	___
2. Gömleğini iliklemek	___	___	___
3. Kazağını başının üzerinden giymek	___	___	___
4. Musluk açmak	___	___	___
5. Yere oturup sonrasında kalkmak	___	___	___
6. Havlu ile sırtını kurulamak	___	___	___
7. Yüzünü yıkamak	___	___	___
8. Ayakkabı bağcığını bağlamak	___	___	___
9. Çorap giymek	___	___	___
10. Diş fırçalamak	___	___	___
11. Kollardan destek almadan sandalyeden kalmak	___	___	___
12. Yatağa yatmak	___	___	___
13. Çatal ve bıçak kullanarak yiyecekleri kesmek	___	___	___
14. Boş bardağı ağıza götürmek	___	___	___
15. Daha önceden açılmış kavanozu açmak	___	___	___
16. Yardımsız 50 adım yürümek	___	___	___
17. Beş basamak çıkmak	___	___	___
18. Ayak parmaklarının ucunda yükselmek	___	___	___
19. Başın üzerine uzanmak	___	___	___
20. Yataktan kalmak	___	___	___
21. Ayakta dururken yerden bir şey almak	___	___	___
22. Kapı tokmağını çevirerek açmak	___	___	___
23. Başını döndürüp omzunun üzerinden bakmak	___	___	___

2. Bölüm: Yardımcı Araç ve Cihazlar

Çoğunuzun herhangi bir aktivite sırasında kullandığı araç veya cihazlar varsa işaretleyin.

	Kulanıyor	Kullanmıyor
Baston	___	___
Walker/Yürüteç	___	___
Koltuk Değneği	___	___
Tekerlekli Sandalye	___	___
Kalınlaştırılmış kalem	___	___
Düğme kancası	___	___
Fermuar çekeceği	___	___
Ayakkabı çekeceği	___	___
Özel mutfak gereçleri	___	___
Özel sandalye	___	___
Özelleştirilmiş klozet	___	___
Küvet sandalyesi	___	___
Kavanoz açacağı	___	___
Küvet tutunma barları	___	___
Uzanma çubukları	___	___

Çocuğunuz yukarıdakilerden başka bir yardımcı araç, gereç, alet veya cihaz kullanıyor mu?

Eğer evet ise, tanımlayınız. ____

3. Bölüm: Başkalarından Yardım

Çocuğunuz aşağıdaki aktiviteler sırasında herhangi birine ihtiyaç duyuyorsa işaretleyin.

	Yardıma ihtiyacı yok	Yardıma ihtiyacı var
Sabahları giyinirken	___	___
Sabahları yıkanırken	___	___
Yatağa girip çıkarken	___	___
Yemek yerken	___	___
Evin etrafında dolaşırken	___	___
Sandalyeye oturup kalkarken	___	___
Nesnelere uzanıp alırken	___	___

4. Bölüm: Ağrı

Ayrıca çocuğunuzun hastalığından dolayı ağrıdan etkilenip etkilenmediğini öğrenmek istiyoruz.

GEÇTİĞİMİZ HAFTA çocuğunuzun hastalığından dolayı ne kadar ağrısı olduğunu düşünüyorsunuz? Ağrı şiddetini aşağıda verilen çizgi üzerinde işaretleyiniz.

0 ____ 100

Ağrı yok   Çok şiddetli ağrı

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 3, 544–562

https://doi.org/10.21449/ijate.1167705

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

A comparative study of ensemble methods in the field of education: Bagging and Boosting algorithms

Hikmet Şevgin [1,*]

[1]Van Yüzüncü Yıl University, Faculty of Education, Department of Educational Sciences, Van, Türkiye

ARTICLE HISTORY

Received: Aug. 27, 2022
Revised: July 10, 2023
Accepted: Aug. 26, 2023

Keywords:
Educational Data Mining,
Ensemble Learning,
Bagging,
Boosting.

Abstract: This study aims to conduct a comparative study of Bagging and Boosting algorithms among ensemble methods and to compare the classification performance of TreeNet and Random Forest methods using these algorithms on the data extracted from ABİDE application in education. The main factor in choosing them for analyses is that they are Ensemble methods combining decision trees via Bagging and Boosting algorithms and creating a single outcome by combining the outputs obtained from each of them. The data set consists of mathematics scores of ABİDE (Academic Skills Monitoring and Evaluation) 2016 implementation and various demographic variables regarding students. The study group involves 5000 students randomly recruited. On the deletion of loss data and assignment procedures, this number decreased to 4568. The analyses showed that the TreeNet method performed more successfully in terms of classification accuracy, sensitivity, F1-score and AUC value based on sample size, and the Random Forest method on specificity and accuracy. It can be alleged that the TreeNet method is more successful in all numerical estimation error rates for each sample size by producing lower values compared to the Random Forest method. When comparing both analysis methods based on ABİDE data, considering all the conditions, including sample size, cross validity and performance criteria following the analyses, TreeNet can be said to exhibit higher classification performance than Random Forest. Unlike a single classifier or predictive method, the classification or prediction of multiple methods by using Boosting and Bagging algorithms is considered important for the results obtained in education.

1. INTRODUCTION

The retrieval of information that needs to be obtained in order to make speculations concerning an event or situation from a community instead of a single person definitely provides the opportunity to make stronger inferences with poorer error rate. In the daily life as well, the attempt to obtain a greater deal of information that can be gained regarding an event or situation, and the overall evaluation of the collected data, is ultimately the result of attempting to reach a more precise conclusion. However, during a decision phase yielding important results, the opinions of experts who are thought to help make decisions are consulted. For example, the opinions of several specialists are asked before a life- threatening operation. In addition, ensemble- based decision- making processes are also administered to elect a manager or to

*CONTACT: Hikmet Şevgin ✉ hikmetsevgin@gmail.com Van Yüzüncü Yıl University, Faculty of Education, Department of Educational Sciences, Van, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

decide on a new law (Polikar, 2012). Likewise, ensemble methods performs analysis methods and, in this respect, it has received increasing attention in recent years with its use with various multiple classification systems, data mining methods and machine learning algorithms (Do-Nascimento et al., 2019; Lee et al., 2010; Zhang & Ma, 2012). The methods that were initially used to reduce the variance of classification and predictive analyses and to increase the accuracy of classification were then successfully utilized for various purposes such as feature selection and the determination of confidence interval (Abeel et al., 2010; Kumari, 2012; Saeys et al., 2008; Zhang & Ma, 2012).

Technological advancement and novel statistical algorithms have allowed for a better understanding of data mining and improved its use. The emergence and development of ensemble learning in the last quarter can be regarded as a reflection of this process. On account of the combination of basic statistical methods to generate ensemble learning methods, the results with high classification success and precise prediction as well as low error variance have been obtained (Bauer & Kohavi, 1999; Hansen & Salamon, 1990; Onan, 2015; Opitz & Shavlik, 1996; Polikar, 2006; Sagi & Rokach, 2018) and, in this respect, its use has recently increased in various areas such as health, economy, banking, agriculture, engineering, business and education (Akman, 2010; Şevgin & Önen 2022).

There have been several studies employing ensemble methods encountered in the literature (Abidi et al., 2020; Baskin et al., 2017a; Baskin et al., 2017b; Dietterich, 2000; Dietterich, 2002; Freund & Schapire, 1996; Friedman, 2001; Kapucu & Cubukcu, 2021; Kausar et al., 2020; Li et al., 2022; Mousavi & Eftekhari, 2015; Pong-Inwong & Kaewmak, 2016; Steinki & Mohammad, 2015; Wang et al., 2018). It is worth noting that the researchers who conduct studies on data mining and machine learning have fallen behind in discovering the success of Ensemble-based learning methods in terms of classification and prediction-based decision-making (Polikar, 2012). Nevertheless, with the studies carried out in recent years, it has been seen that a great deal of knowledge and literature have been obtained especially in the field of education (Abdar et al., 2018; Abellán & Castellano, 2017; Aggarwal et al., 2021; Almasri et al., 2019; Ashraf et al., 2021; Ashraf et al., 2020; Arun et al., 2021; Guo et al., 2021; Karalar et al., 2021; Keser & Aghalarova, 2022; Kotsiantis et al., 2010; Injadat et al., 2020a; Injadat et al., 2020b; Premalatha & Sujatha, 2021). This comparative study focusing on Bagging and Boosting (Akman, 2010; Zhou, 2012) algorithms that are the most well-known Ensemble methods may contribute to the literature and, particularly the field of educational data mining, in order to list and utilize the concept of Ensemble Learning and its methods among advanced statistical methods in the field of education.

In the field of education, both in the phase of various and big data processing that poses opportunities for the construction of education within the Ministry and in the analysis process of multidimensional, complicated and noisy data obtained from students and teachers through large- scale tests, it is of importance to achieve strong and non-deviating outputs. Indeed, the use of ensemble methods can be considered as flexibility (Strobl et al., 2009) for the data analysis in the noisy data by its nature that we often call traditional which do not provide various assumptions required for the parametric methods. Thus, the achievement of the output with lower error variances in the field of education can be contributed. Considering the situations where decisions regarding students such as fail- pass or successful- unsuccessful are made or variables that affect student achievement are examined, the realization of analyses with high classification and prediction success and poor error rate may ensure the results in terms of high classification/ decision validity. It is clear that the use of ensemble methods in education serves to obtain results with high classification and prediction success and to gain results with high classification/ decision validity. Therefore, it is considered important to utilize ensemble methods to obtain evidence concerning classification/ precision validity in the procedures to be performed for classification and prediction.

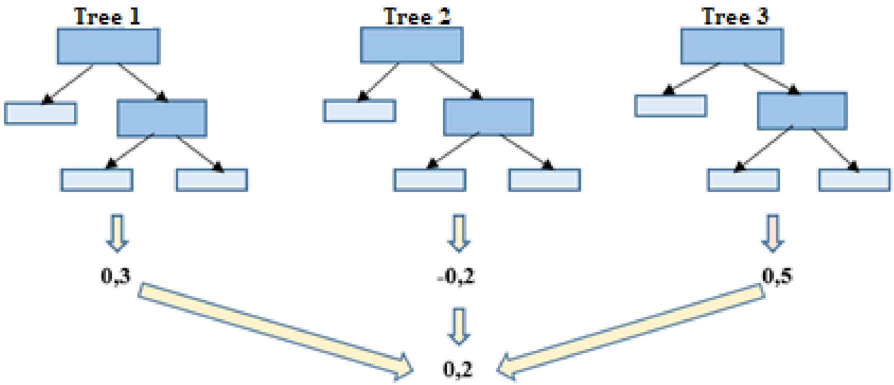
1.1. Ensemble Learning

Recently, in the process of statistically synthesizing the data obtained through scientific research, the idea of combining multiple methods to produce a new model based on classification or prediction has been emphasized by the researchers and been the subject of publications in recent years. Tukey is the first researcher who has introduced the concept of ensemble learning (1977) where he had used linear regression model to fit the original data as first step and then again linear regression model to fit the residual as a second step (Sagi & Rokach, 2018). Later, in the 1990s, Hansen and Salamon shared the outputs of neural network ensembles. In addition, in 1996, Breiman first proposed ideas for the Bagging algorithm and in the same year, Freund and Schapire came up with the first boosting algorithm. Subsequently, the AdaBoost algorithm was introduced by Freund and Schapire (1996) as a result of combining multiple weak classifiers to build one strong classifier. Moreover, certain studies on the development of ensemble methods using boosting algorithms such as Gradien Boosting presented by Friedman et al. (2000) and Multiple Additive Regression Trees (MART) proposed by Friedman and Meulman (2003) have been encountered. In the meantime, numerous ensemble methods which perform ensemble learning by using Bagging and Boosting algorithms have been developed (Kumari, 2012; Polikar, 2006; Schapire, 2003; Zhou, 2012).

The fact that the information is obtained from the narration of more than one person who witnessed the same event rather than of a person, in other words, the information gathered from ensembles provide more reliable results with high accuracy. Learning in this way is called ensemble learning (Polikar, 2012). Similarly, the combination of the predictions of several base estimators is generally better than the prediction of one best predictor. A group of predictive methods is gathered under the title of Ensemble and the process of making predictions from the ensemble is called Ensemble Learning (Geron, 2019). To sum, Ensemble methods can be considered as the combination of multiple methods to produce outputs with higher success (Quinlan, 1996), that is outputs with higher levels of reliability (Akman, 2010; Maclin & Opitz, 1997) in contrast to the outputs based on classification and prediction obtained from single methods. These methods, combined together to give an ensemble, can be a decision tree (C&RT, C5, CHIAD, ID3, QUEST) as well as such methods as MARS, YSA, SVA (Chen & Guestrin, 2016; Clarke et al., 2009; Freund & Schapire, 1996; Friedman, 2001; Friedman & Meulman, 2003; Quinlan, 1996; Sutton, 2005; Zhou, 2012). The algorithms that combine these methods and give an ensemble are Boosting, Bagging, Stacking, Max Voting, Averaging, Weighted Averaging and Blending algorithms (Baskin et al. 2017a; Zhang & Ma, 2012; Zhou, 2012). Of these algorithms, Bagging and Boosting are the most elaborated and known ensemble learning algorithms (Akman, 2010; Zhou, 2012). Within the scope of this study, Bagging and Boosting algorithms are included.

As stated above, although Bagging and Boosting algorithms can be applied to several methods, it has been seen that they are mostly used together with decision trees in the literature. In certain sources, however, ensemble methods are referred as Tree- based Ensemble Methods (Akman, 2010). The TreeNet method, which creates ensembles using classification and regression trees (C&RT) with the boosting algorithm, and the Random Forest (Breiman, 2001), which creates ensembles using C&RT with the bagging algorithm, are included in the present study. In certain sources, although Random Forest is considered as an Ensemble method independently due to the fact that it creates random subspaces to do a random selection of a subset of features to use to grow each tree (Geron, 2019; Han et al., 2012), it is also included in the Bagging title since it utilizes Bagging algorithm in the formation of ensemble (Clarke et al., 2009; Nisbet et al., 2009). Hastie et al. (2009) stated that Random Forest method was a modification of the Bagging algorithm. The main factor choosing TreeNet and Random Forest methods for the current study is that both methods are Ensemble methods that combine single decision trees (classification and regression trees - C&RT) with Bagging and Boosting algorithms and combine the outputs obtained from each of them into a single output. An example representing the working principle of ensemble methods is presented in Figure 1 below:

Figure 1. *The illustration of the working principle of Ensemble Model*



In Figure 1, the value of each tree is combined to produce the final value of the ensemble. The combination process differs since Bagging and Boosting algorithms use their own techniques. During the consolidation process, boosting algorithm iteratively constructs a series of decision trees being trained whereas Bagging algorithm consists of simple random sampling with replacement. These algorithms and the analyses that use them are respectively elaborated below.

1.1.1. *Boosting*

In Boosting algorithm, each model is constructed on the incorrectly predicted data of the previous model (Friedman, 2001). In other words, each model learns from the errors of the previous model. This is realized by weighting the data points and the whole process continues sequentially (Friedman & Meulman, 2003). Then, the weak learners are eliminated one by one and the strong learner is reached (Polikar, 2012). The last model is yielded from the weighted average of all models (Zhou, 2012).

Boosting algorithm [Rokach (2019)].

```
Input: $I$ (a weak inducer), $S$. (a training set) and $k$ (the sample size for the first classifier)
Output: $M_1$, $M_2$, $M_3$
1: $S_1 \leftarrow$ Randomly selected $k < m$ instances from $S$ without replacement;
2: $M_1 \leftarrow I(S_1)$
3: $S_2 \leftarrow$ Randomly selected instances (without replacement) from $S - S_1$ such that half of them
   are correctly classified by $M_1$.
4: $M_2 \leftarrow I(S_2)$
5: $S_3 \leftarrow$ any instances in $S - S_1 - S_2$ that are classified differently by $M_1$ and $M_2$.
```

As shown above, boosting algorithm has an iterative characteristic. The algorithm generates three classifiers. The sample S_1, which is used to train the first classifier M_1, is randomly selected from the original data set. The second classifier, M_2, is trained on a sample M_2, half of which consists of instances that are incorrectly classified by M_1, and the other half is composed of instances that are correctly classified by M_2. The last classifier, M_3, is trained with instances that the two previous classifiers disagree on (Rokach, 2019).

The error rate of the M_i model is calculated using the given the formula below:

$$error(M_i) = \sum_{j=1}^{d} wj \times error(X_j) \quad (1)$$

In this equation, *error* (X_j) is the classification error of Xj. If the group is incorrectly classified, error (Xj) = 1, otherwise it is 0 (zero) (Han et al., 2012). If the performance of the classifier, M_i, is poor, the classification error exceeds 0.5, in which case M_i is abandoned. Instead, the operation is retried by generating a new Si training data (Han et al., 2012). The error rate of M_i affects the updating of the weights of the training set. If the observations are correctly classified, the weighting of observations is multiplied by the value obtained from the equation below:

$$\frac{error(M_i)}{(1-error(M_i)} \tag{2}$$

When the weights of all correctly classified observations are updated, the weights of all observations (including those that are incorrectly classified) are normalized so that their sum remains the same as before. As a result, the weights of misclassified observations are increased and the weights of correctly classified observations are reduced. The lower the error rate is for a classifier, the higher the accuracy rate is (Han et al., 2012). The weight calculated for each M_i classifier is represented by the equation below:

$$log\,\frac{1-error(M_i)}{error(M_i)} \tag{3}$$

Based on boosting algorithm, various alternatives such as AdaBoost (Adaptive Boosting – Freund & Schapire, 1996), Gradient Boosting (Friedman, 2001), XGBoost (Chen & Guestrin, 2016) have been developed to determine the weights used in the training and classification phases of the boost iteration. However, AdaBoost and Gradient Bosting are commonly used algorithms (Sinharay, 2016).

1.1.2. *Bagging*

Bagging is an abbreviation for Bootstrap-Aggregating. It was first proposed by Leo Breiman in 1996. It is a simple, yet effective method for generating an ensemble of classifiers. The ensemble classifier that is created by this method consolidates the outputs of various learned classifiers into a single classification and this results in a classifier whose accuracy is greater than the accuracy of each individual classifier (Rokach, 2019). Bootstrap in the bagging algorithm is represented as resampling (Breiman, 1996). In this method, each classifier in the ensemble is trained on a sample of instances taken with replacement (allowing repetitions) from the training set. All classifiers are trained using the same learning algorithm. Therefore, some of the original instances may appear more than once in a training set, and some may not be included at all (Efron & Tibshirani, 1993).

Bagging Algorithm [Rokach (2019)].

Input: *I* (a base inducer), *T* (the number of iterations), *S* (the original training set), μ (the sample size).

```
1: $t \leftarrow 1$
2: Repeat
3: $S_t \leftarrow$ a sample of μ instances from $S$ with replacement.
4: Construct classifier $M_t$ using $I$, with $S_t$ as the training set.
5: $t \leftarrow t + 1$
6: until $t > T$
```

The Bagging algorithm works as shown above. The classifiers are all trained using the same learning algorithm. The algorithm receives an induction algorithm *'I'* which is used for training all members of the ensemble. The stopping criterion in line six terminates the training when the ensemble size reaches *'I'*. One of the main advantages of bagging is that it can be implemented easily in a parallel mode by training the various ensemble classifiers on different processors (Rokach, 2019).

The most important feature that distinguishes the Bagging algorithm from the Boosting algorithm is that sampling with replacement is used. That is, it is likely to use a sample more than once in the Bagging algorithm. However, in Boosting algorithm, the sample that has been used is not used again. The common feature of the Bagging and Boosting algorithms is that in both algorithms, they generate the last classifier through multiple voting for classification models, and the last estimator through the average of parameter estimates for regression models (Ferreira & Figueiredo, 2012). In this respect, it has been considered important in terms of using the data obtained in the field of education in the analysis of classification and prediction. Besides, unlike the results obtained by a single method, the use of results obtained through more than one method has also been regarded noteworthy in terms of the reliability and validity of the results obtained. Finally, it has been thought that it may contribute to the field in terms of using novel methods built on Bagging and Boosting algorithms in education. In fact, it has been seen that both the Boosting and Bagging algorithms are included in certain studies conducted in the field of education. However, this study is remarkable in terms of the fact that it elaborates the concept of *'Ensemble Learning'* entitled under data mining and machine learning and compares the methods based on the most known algorithms, Bagging and Boosting, on the data in the field of education. Therefore, "The purpose of the study is to conduct a comparative study of Bagging and Boosting algorithms among ensemble methods and to examine the classification performance of both methods on the data obtained in the field of education through TreeNet and Random Forest". To this end, answers to the following questions have been sought:

1) Do the performance measurements of TreeNet and Random Forest methods using Bagging and Boosting algorithms obtained according to each sample size based on 3,5 and 10-fold cross validity on the ABİDE data using Bagging and Boosting algorithms differ?

2) Is there a difference between TreeNet and Random Forest method using Bagging and Boosting algorithms on the ABİDE data based on the comparison of RMSE, MSE, MAD and MRAD values?

2. METHOD

The study was designed with quantitative research and a relational survey model was used with a descriptive approach. The relational model allows researchers to obtain information regarding a large group by examining a sample (Leedy & Ormrod, 2005).

2.1. Data Set

The data set of the study consists of mathematics scores of ABIDE (Academic Skills Monitoring and Evaluation) 2016 administered to 8th grade students. ABIDE implementation includes Turkish, Mathematics, Science and Social Studies achievement tests prepared for 8th grade students. However, the Mathematics achievement test was focused in the current research. For the data of 5000 students randomly recruited from the data set, data deletion was carried out for the demographic data and the values were assigned to the obtained from the scales through (MCAR) regression since it is below %5 for the loss data (Tabachnick & Fidell, 2015). As a result of the deletion of loss data and assignment procedures, this number decreased to 4568. The dependent variable (students' maths achievement), which is a continuous variable, was dual-categorized by considering the first quarter of %25 (low maths achievement) and the fourth quarter of %25 (high maths achievement). 2284 (1034 female and 1250 male) students, 1142 in the first quarter and 1142 in the fourth quarter, constitute the sample of the study. Those in the first quarter with maths scores between 343,10- 440,14 refer to the students with low maths achievement whereas those in the fourth quarter with maths scores between 556,62-776,02 refer to the students with high maths achievement.

2.1.1. *Measurement tools*

The current research consists of mathematics achievement test in ABİDE implementation, demographic information collected by student survey and the variables collected at the scale level that are the attitude towards the school, peer bullying, parental approach, liking of mathematics course, self-efficacy perception towards the mathematics course, the value given to the mathematics course and teacher's instructional activities.

Prior to the data analysis through ensemble methods, the reliability, validity and multiple connection problems of the scales used in the research were examined. With the purpose of determining the reliability coefficient, McDonald's (ω) reliability index was employed instead of Chronbach Alpha reliability index due to the fact that the factor loads of the items were not equal (Yurdugül, 2006). McDonald's (ω) reliability index of the scales ranged from 0.77 (parental approach) the lowest to 0.94 (teacher's instructional activities) the highest and these values can be said to be at acceptable levels. In order to prove the validity of the scale, exploratory factor analysis was performed and it was found that each scale had one dimensional and that factor loads of the items varied between 0.369 the lowest and 0.875 the highest. Since the factor loads related to the items are above the acceptable minimum value, 0.30 (Çokluk et al., 2012), it can be said that they are above the acceptable value. Moreover, Tolerance and VIF values were examined for multi connection problem, and it was revealed that Tolerance values ranged between 0.520 and 0.916 and VIF values varied between 1.091and 1.922. Since these values are higher than 0.100 for Tolerance and lower than 10 for VIF (Schroeder et al., 1990), it can be stated that there is no multi connection problem.

2.2. Data Analysis

In the research, the data set was divided into four data sets as 250, 500, 1000 and 2000 in terms of sample size through simple random sampling without replacement. The observations in each data set were assigned to the data set in a way that they were subjected to 3-fold, 5-fold and 10-fold cross validation.

In this study, in the context of ensemble methods, performance criteria based on sample size were compared for TreeNet analysis method using Boosting algorithm and Random Forest method using Bagging algorithm in the background. In data analysis, the educational version of the SPM 7.0 statistical package program and open source Phyton-based Orange package 3.34 version were utilized. In addition, the evaluation of performance criteria yielded by confusion matrix was made through the test data and the 2nd category (Successful) was considered as the focus group.

2.2.1. *TreeNet*

The TreeNet method is based on stochastic gradient boosting algorithm to determine the weights used in the training and classification phases of the incremental iteration (Padmaja et al., 2016). Stochastic gradient boosting, developed by Friedman (2002), is used to address a regression task by optimizing the mean squared error. It is a non- parametric method where each successive learner is trained following the pseudo - residual errors of the preceding learner, thus finding solutions to classification and regression problems (Friedman, 2002; Hastie et al., 2009). The TreeNet (TM Salford Systems, inc.) method has various titles due to commercial concerns such as Multiple Additive Regression Trees-MART (TM Jerill, inc.), Boosted Regression Trees-BRT (TM Stat Soft, inc.), Gradient Boosting Trees (GBT) and Gradient Boosting Model (GBM) (Elish & Elish, 2009; Hill & Lewicki, 2006). TreeNet is successfully applied in science fields where complex relationships of numerous variables are modelled by adding classification trees when the dependent variable is categorical and the regression trees are added when the variable is continuous (Şevgin & Önen, 2022).

2.2.2. *Random forest*

Random Forest method is a special modification of Bagging algorithm (Amrieh et al., 2016; Hastie et al., 2009). It was created as a result of the application of the Random Subspace technique proposed by Ho (1998) on the Bagging method (Biau, 2012). In the bagging method, decision trees are generated by selection from the data set independently of one another through bootstrap technique. However, the Random Subspace method does a random selection of a subset of features to use to grow each tree (Akman, 2010). In Random Forest method, each decision tree that generates the decision forest is created by bootstrap sampling randomly selected from the original data set with replacement. The Random Forest proposed by Breiman (2001) is a non-parametric method applied in science fields where complex relationships of numerous variables are modelled by adding classification trees to regression trees through bootstrap sampling method when the dependent variable is two- class or multi- class (Biau & Scornet, 2016; Geneur et al., 2017).

Recent studies have shown that ensemble learning methods outperform traditional regression methods (Elith et al., 2006). It can be said that TreeNet and Random Forest are among best performing ensemble methods (More detailed information for these two methods, see Breiman, 2001; Friedman, 2002).

2.2.3. *Confusion matrix*

A confusion matrix is a technique for summarizing the performance of a classification algorithm. A confusion matrix is a two-dimensional matrix (“actual” and “predicted”), indexed in one dimension by the true class of an object and in the other by the class that the classifier assigns (Ting, 2017) and it allows easily discovering whether the system mixes the two classes (Şevgin, 2020). Table 1 presents an example of confusion matrix for a two - class classification task.

Table 1. *Confusion matrix.*

		Predicted Class		
		Unsuccessful	Successful	Total
Actual Class	Unsuccessful	TN	FP	TN+FP
	Successful	FN	TP	FP+TP
	Total	TN+FN	FP+TP	TN+FN+FP+TP

(TP: True Positive TN: True Negative FP: False Positive Fn: False Negative)

Confusion matrices represent counts from predicted and actual values. It is applied to binary classification. In this regard, the confusion matrix represents true positive (TP) values, false positive (FP) values, true negative (TN) values and false negative (FN) values (Ting, 2017). The output for True Positive and True Negative shows the instances predicted accurately. However, False Positive and False Negative represent the instances predicted incorrectly. Accuracy is calculated as the sum of two accurate predictions (TP + TN) divided by the total number of data sets (P + N). The best accuracy is 1.0, and the worst is 0.00. Ideally, the sum of TP and TN should have an approximate value to the total of the pattern and the sum of FP and FN values should be close to zero (Han et al., 2012).

2.2.4. *Performance criteria for the categorical dependent variable*

In this research, accuracy- percentage- sensitivity- precision ratios, AUC value of ROC curve and F1 score were used as performance criteria. The formulas are given below:

Accurate classification rate indicates how well the method used in classification problems predicts the class distributions of the data and is often expressed as a percentage.

$$Accurate\ Classification\ Rate = \frac{(\mathrm{TP+TN})}{(\mathrm{TP+FP+TN+FN})} \quad (4)$$

Specificity refers to the probability of a negative test result, conditioned on the individual truly being negative and it takes a value between 0 and 1. This value is usually expressed as a percentage.

$$Specificity = \frac{(TN)}{(TN+FP)} \quad (5)$$

Sensitivity represents how well a test can identify true positives and it reveals a value between 0 and 1. This value is usually expressed as a percentage.

$$Sensitivity = \frac{(TP)}{(TP+FN)} \quad (6)$$

The numerical value of accuracy represents the proportion of true positive results in the selected population and yields a value between 0 and 1. This value is usually expressed as a percentage.

$$Precision = \frac{(TP)}{(TP+FP)} \quad (7)$$

The F- score (also known as the F1- score or F-measure) is defined as the harmonic mean of precision and recall scores of a model in order to ensure a balanced measure of overall classification performance.

$$F1 - Score = 2x\frac{sensitivity\ x\ precision}{sensitivity+precision} \quad (8)$$

2.2.5. *Performance criteria for the continuous dependent variable*

The RMSE, MSE, MAD, and MRAD values which give error values for numerical prediction, allow data mining and machine learning methods to be examined and compared to one another.

RMSE (Root Mean Square Error): RMSE measures the average difference between a statistical model's predicted values and the actual values. The RMSE value is the measurement of how close the predictions are to the actual values. A low RMSE value refers to a better model performance.

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}} \quad (9)$$

MSE (Mean Square Error): MSE is defined as mean or average of the square of the difference between actual and estimated values. Unlike RMSE, MSE is computed without taking the square root. The MSE value quantifies the size of prediction errors and a low MSE value means a better model performance.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \quad (10)$$

MAD (Mean Absolute Deviation): MAD is a measure of the average absolute distance between each data value and the mean of a dataset. The MAD value measures the size of prediction errors, yet, unlike RMSE and MSE, it can be more sensitive to larger extreme outliers since it does not take the square of the deviation.

$$MAD = \frac{\sum_{i=1}^{n}|\chi_i - \bar{x}|}{n} \quad (11)$$

MRAD (Mean Relative Absolute Deviation): MRAD is the average distance between each data point and the mean. MRAD provides an independent assessment of the scale of the measured values by calculating the prediction errors to the actual values. Besides, it is useful or comparing values measured in different times.

$$MRAD = \frac{\frac{(\sum_{i=1}^{n}|(x_i - \bar{x})|)}{n}}{\bar{x}*100} \tag{12}$$

2.2.6. *Cross validation*

Cross validation, also being referred to as rotation estimation, is a resampling technique used in statistical modelling and machine learning to evaluate the performance and generalization ability of two or more models. Cross validation involves dividing the existing dataset into k subsets, training the model on a subset of the data, and evaluating its performance on the remaining fold(s) (Olson & Delen, 2008). In K-fold cross-validation, the full data set is randomly divided into various subsets of k of approximately equal size. The classification model is trained and tested k times. In the present study, 3-fold, 5-fold and 10-fold cross-validity was applied to evaluate the performance of the methods. In other words, a cross-validity was performed in which one- third, one- fifth and one- tenth of the data set were considered as test data.

3. RESULTS

In this section, the TreeNet method using the boosting algorithm in the background and the Random Forest method using the Bagging algorithm are examined in different sample sizes, 3-fold, 5-fold and 10-fold cross validity rates. At each sample size and each cross-validity rate, the number of trees that is required by the TreeNet and Random Forest methods to generate the optimal model is presented in Table 2.

Table 2. *The number of trees where Treenet and random forest models are established.*

		250	500	1000	2000
TreeNet	3K	648	312	762	484
	5K	446	465	700	475
	10K	561	426	739	465
Random Forest	3K	526	258	589	461
	5K	433	436	547	417
	10K	489	438	628	423

Table 2 represents the number of trees needed to determine the optimal number of trees in the area under the ROC curve for TreeNet (Hastie et al., 2009). For Random Forest, the value with the lowest error rate in the decision forest refers to the number of trees needed for the most appropriate model to be established (Huffer and Park, 2020; Probst and Boulesteix, 2017).

3.1. Findings on the TreeNet and Random Forest Methods by Sample Size

The classification performances yielded by both analysis methods as a result of 3-fold cross validation for each level of the sample size taken from the study group are presented in Table 3 as a percentage. In Table 3, it is seen that for both analysis methods with 3-fold cross-validity, they received the same value in terms of accurate classification rate in 500 sample size although TreeNet method was higher than Random Forest method in 250, 1000 and 2000 sample sizes. In terms of specificity, TreeNet method was found to be higher in 250 smaple sizes whereas Random Forest was revealed to be higher in 500, 1000 and 2000 sample sizes. In terms of sensitivity, it is seen that TreeNet method is higher than Random Forest method in all sample sizes. In terms of accuracy, it is seen that the TreeNet method is higher in the sample size of 250 and 1000 and the Random Forest method is higher in the sample size of 500 and 2000. However, in terms of F1- score, it has been revealed that the TreeNet method is higher than the

Random Forest method in all sample sizes. In terms of AUC value, it has been found that the Random forest method is higher in the sample size of 250 and, however, that TreeNet method is higher in the sample sizes of 500, 1000 and 2000.

Table 3. *Percentages of classification performance by sample sizes for 3-Fold Cross validity.*

			250	500	1000	2000
3K	TreeNet	Accurate Classification Rate	%76.80	%71.40	%77.20	%77.20
		Specificity	%74.56	%70.59	%76.24	%78.35
		Sensitivity	%78.68	%72.24	%78.18	%76.00
		Accuracy	%78.68	%70.24	%76.33	%77.10
		F1- score	%78.68	%71.23	%77.25	%76.54
		AUC value	%83.98	%80.84	%85.77	%84.80
	Random Forest	Accurate Classification Rate	%72.80	%71.40	%74.10	%76.15
		Specificity	%67.54	%72.94	%77.03	%79.52
		Sensitivity	%77.21	%69.80	%71.11	%72.62
		Accuracy	%73.94	%71.25	%75.21	%77.28
		F1- score	%75.54	%70.52	%73.10	%74.88
		AUC value	%80.71	%81.35	%83.61	%84.79

The classification performances obtained by both analysis methods as a result of 5-fold cross validation for each level of the sample size taken from the study group are presented in Table 4 as a percentage.

Table 4. *Percentages of classification performance by sample sizes for 5-Fold Cross validity.*

			250	500	1000	2000
5K	TreeNet	Accurate Classification Rate	%71.20	%75.20	%75.20	%77.10
		Specificity	%67.54	%74.90	%75.45	%77.67
		Sensitivity	%74.26	%75.51	%74.95	%76.51
		Accuracy	%73.19	%74.30	%74.95	%76.66
		F1- score	%73.72	%74.90	%74.95	%76.58
		AUC value	%80.65	%82.53	%84.47	%85.30
	Random Forest	Accurate Classification Rate	%75.20	%74.20	%74.20	%76.55
		Specificity	%71.93	%75.69	%77.22	%79.53
		Sensitivity	%77.94	%72.65	%71.11	%73.44
		Accuracy	%76.81	%74.17	%75.37	%77.47
		F1- score	%77.37	%73.40	%73.18	%75.41
		AUC value	%82.79	%81.75	%83.92	%84.90

In Table 4, it is seen that for both analysis methods with 5-fold cross-validity, the Random Forest method is higher in the accurate classification rate in the sample size of 250 and that the TreeNet method is higher in the sample size of 500, 1000 and 2000. In terms of specificity, it has been demonstrated that Random Forest method is higher in all sample sizes. In terms of sensitivity, it is seen that the Random Forest method is higher in the sample size of 250 and the TreeNet method has been found to be higher in the sample sizes of 500, 1000 and 2000. Moreover, in terms of accuracy, it is seen that the Random Forest method is higher in sample size of 250 and the TreeNet method is higher in 500, 1000 and 2000 sample sizes. As for F1-

score, it is seen that the Random Forest method is higher in the sample size of 250 and TreeNet method is higher in the sample sizes of 500, 1000 and 2000. In terms of AUC value, it has been revealed that the Random Forest method is higher in the sample size of 250 and TreeNet method is higher in the sample sizes of 500, 1000 and 2000.

The classification performances obtained by both analysis methods as a result of 10-fold cross validation for each level of the sample size taken from the study group are presented in Table 5 as a percentage.

Table 5. *Percentages of classification performance by sample sizes for 10-Fold Cross validity*

			250	500	1000	2000
10K	TreeNet	Accurate Classification Rate	%75.20	%74.40	%76.60	%77.20
		Specificity	%76.32	%73.73	%75.84	%78.35
		Sensitivity	%74.26	%75.10	%77.37	%76.00
		Accuracy	%78.91	%73.31	%75.84	%77.10
		F1- score	%76.52	%74.19	%76.60	%76.54
		AUC value	%83.42	%82.92	%84.83	%84.80
	Random Forest	Accurate Classification Rate	%75.20	%74.20	%73.70	%76.35
		Specificity	%71.05	%75.69	%76.24	%79.82
		Sensitivity	%78.67	%72.65	%71.11	%72.72
		Accuracy	%76.43	%74.16	%74.58	%77.56
		F1- score	%77.53	%73.40	%72.80	%75.07
		AUC value	%83.12	%83.17	%83.49	%85.01

In Table 5, it has been demonstrated that both methods receive the same value in the sample size of 250 in terms of correct classification rate with 10-fold cross-validity; however, it has been seen that the TreeNet method is higher compared to the Random Forest method in the sample sizes of 500, 1000 and 2000. Nevertheless. in terms of specificity, it has been found that the TreeNet method is higher in the sample size of 250 and that the Random Forest method is higher in the sample size of 500, 1000 and 2000. As for sensitivity, it has been indicated that the Random Forest method is higher in the sample size of 250 and that TreeNet method is higher in the sample sizes of 500, 1000 and 2000. In terms of accuracy, it is seen that the TreeNet method is higher in the sample size of 250 and 1000 and the Random Forest method is higher in the sample size of 500 and 2000. Furthermore, In terms of F1-score, it is seen that the Random Forest method is higher in the sample size of 250 and TreeNet method is higher in the sample sizes of 500, 1000 and 2000. Finally, in terms of AUC value, it has been revealed that the TreeNet method is higher in the sample sizes of 250 and 1000 and the Random Forest method is higher in the sample sizes of 500 and 2000.

3.2. Findings on the TreeNet and Random Forest Methods Based on RMSE, MSE, MAD and MRAD Performance Measurements

The classification performances of RMSE, MSE, MAD and MRAD values obtained by both analysis methods for each level of sample size taken from the study group are presented in Table 6. As shown in Table 6, it is seen that the TreeNet method yields lower error values than the Random Forest method in all sample sizes. It has been shown that the error values of the TreeNet method, in itself, increase in all metrics towards the sample sizes of 250, 500 and 1000, and decrease in the sample size of 2000. In the Random Forest method. however, it has been revealed that the error values obtained in all metrics decrease as the sample size increases.

Table 6. *RMSE. MSE. MAD and MRAD Values of Both Methods in Each Sample Size.*

		250	500	1000	2000
TreeNet	RMSE	46.45	61.67	71.65	71.10
	MSE	2158.32	3803.72	5133.75	5056.28
	MAD	36.00	48.71	57.29	56.47
	MRAD	0.075	0.102	0.120	0.118
Random Forest	RMSE	96.65	93.35	92.91	90.32
	MSE	9342.81	8714.05	8633.24	8156.97
	MAD	83.72	79.61	78.70	75.69
	MRAD	0.175	0.166	0.165	0.159

4. DISCUSSION and CONCLUSION

In the current study. Bagging and Boosting algorithms were elaborated and the classification performances of TreeNet and Random Forest methods using these algorithms were compared through a real data set from a large-scale national assessment. In this section, the results yielded from both methods and the usefulness of both analysis methods in education have been discussed.

As the first result of the research, it was found that the performance measurements of TreeNet and Random Forest methods varied based on each sample size under 3, 5 and 10-fold cross validity. In its broadest sense, the TreeNet method yielded high values in accuracy, sensitivity rate, F1-score and AUC value in large samples whereas it takes high values in specificity and accuracy in smaller samples while it takes high values in specificity and accuracy in smaller samples. Furthermore, the Random Forest method takes high values in large samples in terms of specificity and accuracy although it yields high values in the smaller samples in the accuracy, sensitivity, F1-score and AUC value. In the performance measures listed above, it can be said that the Random forest method performs better in specificity and accuracy; however, the TreeNet method have a better performance in other metrics. Märker et al. (2011) noted that the TreeNet method performed better compared to the Random Forest method in terms of AUC value, Cohen's Kappa statistics and R2 value. In contrast, Mi et al. (2017) and Padmaja et al. (2021) reported in their study that the Random forest method performed better than the TreeNet method.

As the second result of the research, it has been found that with the increase in the number of samples within the TreeNet method the metric values expressing the error increase by the sample sizes of 1000 and 2000 and that it yield similar values in the sample sizes of 1000 and 2000. Instead of generating new classes through random selection from the data set, the Boosting algorithm learns from the errors and determines with which samples the incorrect classification process is performed and makes selections on these samples. In other words, considering that the Boosting algorithm acts sequentially with an iterative working principle with the logic of learning from errors by using the whole sample, the amount of error it produces in low data is reflected as less until the optimum number of trees is reached. In addition, as for the Random Forest method, it has been seen that the metric values that express the correct error from 250 samples to 2000 samples are reduced. Considering that the Bagging algorithm acts with an iterative working principle with the logic of learning from errors in order to use the random sample it yields from the data set to put back into place, it can be said that it can be said that these values decrease with the increase of the data it pulls randomly until it reaches the optimum number of trees to establish the final model. Finally, at all error rates for each sample size of the same data set, the TreeNet method has been shown to produce lower values than the Random Forest method. In this respect, it can be said that the TreeNet method produces more unbiased (Robust) results and performs better than the Random Forest method. Indeed, Padmaja et al. (2016) reported in their studies that the TreeNet method was more successful than the

Random forest method. In the same vein, in the study conducted by Subasi et al. (2022), it was reported that Stochastic Gradient Boosting method (another literature use of the TreeNet method) performed better compared to the Random Forest, Support Vector Machines, K-nearest neighbours algorithm and artificial neural networks for RMSE, MSE, MAE and RAE performance criteria. Moreover, Tuğ-Karaoğlu and Okut (2020) have stated that the Boosting algorithm is more successful than the Bagging algorithm in their study and the same authors have also drawn attention to the above-mentioned issues as the source of success. Likewise, Dietterich (2000b), Machová (2006) and Quinlan (1996) stated in their study that the Boosting algorithm was more successful than the Bagging algorithm.

When both analysis methods are compared internally, taking into account all conditions including sample size, cross-validity and performance criteria, it can be said that the TreeNet method shows higher classification and prediction performance than the Random Forest method. Märker et al. (2011) stated in their studies that the TreeNet method performed better than the Random Forest method in terms of classification performance. Similarly, Hastie et al. (2009) reported that boosting-based algorithms gave better results than bagging-based algorithms in most problem situations.

In conclusion, these conclusions have been yielded by the mathematics achievement test of the ABİDE implementation administered to 8th grade students. Further studies with higher actual and artificial data are recommended for the comparability of the results. Furthermore, it is recommended to use both analysis methods to give flexibility to the analysis of data sets obtained in the field of education, especially data that do not show parametric features.

Acknowledgments

This study was partly presented as an oral presentation at the Measurement and Evaluation Congress in Education and Psychology on 01-04 September 2021 in Ankara.

Declaration of Conflicting Interests and Ethics

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

Orcid

Hikmet Şevgin https://orcid.org/0000-0002-9727-5865

REFERENCES

Abdar, M., Zomorodi-Moghadam, M., & Zhou, X. (2018, 12-14, November). *An ensemble-based decision tree approach for educational data mining* [Conference presentation]. In 2018 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC), Kaohsiung, Taiwan. https://doi.org/10.1109/BESC.2018.8697318

Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., & Saeys, Y. (2010). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics, 26*(3). 392-398. https://doi.org/10.1093/bioinformatics/btp630

Abidi, S.M.R., Zhang, W., Haidery, S.A., Rizvi, S.S., Riaz, R., Ding, H., & Kwon, S.J. (2020). Educational sustainability through big data assimilation to quantify academic procrastination using ensemble classifiers. *Sustainability, 12*(15), 6074. https://doi.org/10.3390/su12156074

Aggarwal, D., Mittal, S., & Bali, V. (2021). Significance of non-academic parameters for predicting student performance using ensemble learning techniques. *International Journal of System Dynamics Applications*, *10*(3), 38-49. https://doi.org/10.4018/IJSDA.2021070103

Akman, M. (2010). *An overview of data mining techniques and analysis of Random Forests method: An application on medical field* [Unpublished master's thesis]. Ankara University.

Almasri, A., Celebi, E., & Alkhawaldeh, R.S. (2019). EMT: Ensemble meta-based tree model for predicting student performance. *Hindawi*, 1-13. https://doi.org/10.1155/2019/3610248

Amrieh, E.A., Hamtini, T., & Aljarah, I. (2016). Mining educational data to predict student's academic performance using ensemble methods. *International Journal of Database Theory and Application*, *9*(8), 119-136. http://dx.doi.org/10.14257/ijdta.2016.9.8.13

Ashraf, M., Zaman, M., & Ahmed, M. (2020). An intelligent prediction system for educational data mining based on ensemble and filtering approaches. *Procedia Computer Science*, *167*, 1471-1483. https://doi.org/10.1016/j.procs.2020.03.358

Ashraf, M., Salal, Y.K., & Abdullaev, S.M. (2021). *Educational Data Mining Using Base (Individual) and Ensemble Learning Approaches to Predict the Performance of Students*. In Data Science. Springer. https://doi.org/10.1007/978-981-16-1681-5_2

Arun, D.K., Namratha, V., Ramyashree, B.V., Jain, Y.P., & Choudhury, A.R. (2021, 27-29, January). *Student academic performance prediction using educational data mining* [Conference presentation]. In 2021 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India. https://doi.org/10.1109/ICCCI50826.2021.9457021

Baskin, I.I., Marcou, G., Horvath, D., & Varnek, A. (2017a). *Bagging and boosting of classification models*. Tutorials in Chemoinformatics, 241-247. John Wiley & Sons Ltd. https://doi.org/10.1002/9781119161110.ch15

Baskin, I.I., Marcou, G., Horvath, D., & Varnek, A. (2017b). *Bagging and boosting of regression models*. Tutorials in Chemoinformatics, 249-255. John Wiley & Sons Ltd. https://doi.org/10.1002/9781119161110.ch16

Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging. Boosting and variants. *Machine Learning. 36*(1), 105-139. https://doi.org/10.1023/A:1007515423169

Biau, G. (2012). Analysis of a Random Forest. *Journal of Machine Learning Research*, *13*(2012), 1063-1095. https://www.jmlr.org/papers/volume13/biau12a/biau12a.pdf

Biau, G., & Scornet, E., (2016). A random forest guided tour. *An Official Journal of the Spanish Society of Statistics and Operations Research*, *25*(2), 197-227. https://doi.org/10.1007/s11749-016-0481-7

Breiman, L. (1996). Bagging predictors. *Machine Learning 24*(2), 123-140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016, 13, August). *Xgboost: A scalable tree boosting system* [Conference presentation]. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco, CA, USA. http://dx.doi.org/10.1145/2939672.2939785

Clarke, B., Fokoue, E., & Zhang, H.H. (2009). *Principles and theory for data mining and machine learning*. Springer Science & Business Media. https://doi.org/10.1007/978-0-387-98135-2

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2012). *Multivariate statistics for social sciences: SPSS and LISREL applications* (2th edition). Pegem Academy.

Do-Nascimento, R.L., Fagundes, R.A., & Maciel, A.M. (2019, 15-18, July). *Prediction of School Efficiency Rates through Ensemble Regression Application* [Conference presentation]. In 2019 IEEE 19th International Conference on Advanced Learning Technologies, Maceio, Brazil. https://doi.org/10.1109/ICALT.2019.00050

Dietterich, T.G. (2000a). Ensemble Methods in Machine Learning. In: Multiple Classifier Systems. MCS 2000. *Lecture Notes in Computer Science, 1857*, 1-15. https://doi.org/10.1007/3-540-45014-9_1

Dietterich, T.G. (2000b). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning, 40*(2), 139-157. https://doi.org/10.1023/A:1007607513941

Dietterich, T.G. (2002). Ensemble learning. *The Handbook of Brain Theory and Neural Networks, 2*(1), 110-125. https://courses.cs.washington.edu/courses/cse446/12wi/tgd-ensembles.pdf

Efron, B., & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall/CRC.

Elish, M.O., & Elish, K.O. (2009, 24-27, March). *Application of treenet in predicting object-oriented software maintainability: A comparative study*. In 2009 13th European Conference on Software Maintenance and Reengineering, Kaiserslautern, Germany. https://doi.org/10.1109/CSMR.2009.57

Ferreira, A.J., & Figueiredo, M.A. (2012). Boosting algorithms: A review of methods, theory, and applications. *Ensemble machine learning* (1th edition, 35-85). Springer. https://doi.org/10.1007/978-1-4419-9326-7_2

Freund, Y., & Schapire, R.E. (1996, 3-6, July). *Experiments with a new boosting algorithm* [Conference presentation]. Proceedings of the Thirteenth International Conference on International Conference on Machine Learning, Bari Italy.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Annals of Statistics*, *28*(2), 337-407. https://doi.org/10.1214/aos/1016218223

Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, *29*(5) 1189-1232. https://www.jstor.org/stable/2699986

Friedman, J.H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38*(4), 367-378. https://doi.org/10.1016/S0167-9473(01)00065-2

Friedman, J.H., & Meulman, J. J. (2003). Multiple additive regression trees with application in epidemiology. *Statistics in Medicine*, *22*(9), 1365-1381. https://doi.org/10.1002/sim.1501

Geneur, R., Poggi, J.M., Tuleao Malot, C., & Villa-Vialaneix, N. (2017). Random forest for big data. *Big Data Research*, *9*, 28-46. https://doi.org/10.1016/j.bdr.2017.07.003

Geron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (1th edition). O'Reilly Media.

Guo, J., Bai, L., Yu, Z., Zhao, Z., & Wan, B. (2021). An AI-application-oriented in-class teaching evaluation model by using statistical modeling and ensemble learning. *Sensors, 21*(1), 241. https://doi.org/10.3390/s21010241

Han, J., Kamber, M., & Pei, J., (2012). *Data mining: concepts and techniques* (3th edition). Elsevier.

Hansen, L.K., & Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 12*(10), 993-1001. https://doi.org/10.1109/34.58871

Hastie, T., Tibshirani, R. & Friedman, J.H. (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer. https://doi.org/10.1007/978-0-387-21606-5

Hill, T., & Lewicki, P. (2006). *Statistics: methods and applications: a comprehensive reference for science, industry, and data mining* (1th edition). StatSoft, Inc.

Ho, T.K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(8), 832-844. https://doi.org/10.1109/34.709601

Huffer, F.W., & Park, C. (2020). A Simple Rule for Monitoring the Error Rate of Random Forest for Classification. *Quantitative Bio-Science, 39*(1), 1-15.

Injadat, M., Moubayed, A., Nassif, A.B., & Shami, A. (2020a). Systematic ensemble model selection approach for educational data mining. *Knowledge-Based Systems, 200*, 105992. https://doi.org/10.1016/j.knosys.2020.105992

Injadat, M., Moubayed, A., Nassif, A.B., & Shami, A. (2020b). Multi-split optimized bagging ensemble model selection for multi-class educational data mining. *Applied Intelligence, 50*(12), 4506-4528. https://doi.org/10.1007/s10489-020-01776-3

Kapucu, C., & Cubukcu, M. (2021). A supervised ensemble learning method for fault diagnosis in photovoltaic strings. *Energy*, 227, 1-12. https://doi.org/10.1016/j.energy.2021.120463

Karalar, H., Kapucu, C., & Gürüler, H. (2021). Predicting students at risk of academic failure using ensemble model during pandemic in a distance learning system. *International Journal of Educational Technology in Higher Education, 18*(1), 1-18. https://doi.org/10.1186/s41239-021-00300-y

Kausar, S., Oyelere, S., Salal, Y., Hussain, S., Cifci, M., Hilcenko, S., ... & Huahu, X. (2020). Mining smart learning analytics data using ensemble classifiers. *International Journal of Emerging Technologies in Learning, 15*(12), 81-102. https://www.learntechlib.org/p/217561/

Keser, S.B., & Aghalarova, S. (2022). HELA: A novel hybrid ensemble learning algorithm for predicting academic performance of students. *Education and Information Technologies, 27*(4), 4521-4552. https://doi.org/10.1007/s10639-021-10780-0

Kotsiantis, S., Patriarcheas, K., & Xenos, M. (2010). A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. *Knowledge-Based Systems*, *23*(6), 529-535. https://doi.org/10.1016/j.knosys.2010.03.010

Kumari, G. T. (2012). A Study of Bagging and Boosting approaches to develop meta-classifier. *Engineering Science and Technology: An International Journal*, *2*(5), 850-855.

Leedy, P.D., & Ormrod, J.E. (2005). *Practical research (Vol. 108).* Saddle River.

Lee, S.L.A., Kouzani, A.Z., & Hu, E. J. (2010). Random forest based lung nodule classification aided by clustering. *Computerized Medical Imaging and Graphics*, *34*(7), 535-542. https://doi.org/10.1016/j.compmedimag.2010.03.006

Li, B., Yu, Q., & Peng, L. (2022). Ensemble of fast learning stochastic gradient boosting. *Communications in Statistics-Simulation and Computation*, *51*(1), 40-52. https://doi.org/10.1080/03610918.2019.1645170

Machová, K., Puszta, M., Barčák, F., & Bednár, P. (2006). A comparison of the bagging and the boosting methods using the decision trees classifiers. *Computer Science and Information Systems*, *3*(2), 57-72. https://doi.org/10.2298/CSIS0602057M

Maclin, R., & Opitz, D. (1997, 27-31, July). *An empirical evaluation of bagging and boosting* [Conference presentation]. *AAAI-97:* Fourteenth National Conference on Artificial Intelligence, Rhode Island.

Märker, M., Pelacani, S., & Schröder, B. (2011). A functional entity approach to predict soil erosion processes in a small Plio-Pleistocene Mediterranean catchment in Northern Chianti, Italy. *Geomorphology, 125*(4), 530-540. https://doi.org/10.1016/j.geomorph.2010.10.022

Mi, C., Huettmann, F., Guo, Y., Han, X., & Wen, L. (2017). Why choose Random Forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence. *Peer J, 5*, e2849.

Mousavi, R., & Eftekhari, M. (2015). A new ensemble learning methodology based on hybridization of classifier ensemble selection approaches. *Applied Soft Computing, 37*, 652-666. https://doi.org/10.1016/j.asoc.2015.09.009

Nisbet, R., Elder, J., & Miner, G. (2009). *Handbook of statistical analysis and data mining applications* (1th edition). Academic Press.

Olson, D.L., & Delen, D. (2008). *Advanced data mining techniques.* Springer Science & Business Media.

Onan, A. (2015). On the performance of ensemble learning for automated diagnosis of breast cancer. R. Silhavy R. Senkerik, Z. K. Oplatkova, Z. Prokopova, & P. Silhavy (eds.), *In Artificial Intelligence Perspectives and Applications: Proceedings of the 4th Computer Science On-line Conference, Vol 1* (pp. 119-129). Springer International Publishing.. https://doi.org/10.1007/978-3-319-18476-0_13

Opitz, D.W., & Shavlik, J.W. (1996). Generating accurate and diverse members of a neural network ensemble. *Advances in Neural Information Processing Systems*, *8*, 535-541.

Padmaja, B., Prasad, V.R., & Sunitha, K.V.N. (2016). TreeNet analysis of human stress behavior using socio-mobile data. *Journal of Big Data*, *3*(1), 1-15. https://doi.org/10.1186/s40537-016-0054-3

Padmaja, B., Srinidhi, C., Sindhu, K., Vanaja, K., Deepika, N.M., & Patro, E.K.R. (2021). Early and accurate prediction of heart disease using machine learning model. *Turkish Journal of Computer and Mathematics Education, 12*(6), 4516-4528.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine, 6*(3). 21-45. https://doi.org/10.1109/MCAS.2006.1688199

Polikar, R. (2012). Ensemble learning. *In Ensemble machine learning* (1th edition pp. 1-34). Springer. https://doi.org/10.1007/978-1-4419-9326-7_1

Premalatha, N., & Sujatha, S. (2021, 15-17, September). *An Effective Ensemble Model to Predict Employment Status of Graduates in Higher Educational Institutions* [Conference presentation]. In 2021 Fourth International Conference on Electrical, Computer and Communication Technologies Erode, India. https://doi.org/10.1109/icecct52121.2021.9616952

Probst, P., & Boulesteix, A.L. (2017). To tune or not to tune the number of trees in random forest. *The Journal of Machine Learning Research*, *18*(1), 6673-6690. http://jmlr.org/papers/v18/17-269.html

Rokach, L. (2019). *Ensemble learning: Pattern classification using ensemble methods* (2th edition). World Scientific. https://doi.org/10.1142/9789811201967_0003

Pong-Inwong, C., & Kaewmak, K. (2016, 14-17, October). *Improved sentiment analysis for teaching evaluation using feature selection and voting ensemble learning integration* [Conference presentation]. In 2016 2nd IEEE international conference on computer and communications, Chengdu, China. https://doi.org/10.1109/CompComm.2016.7924899

Quinlan, J.R. (1996, 4-8, August). *Bagging, boosting, and C4. 5* [Conference presentation]. In 13th National Conference on Artificial Intelligence, Portland, Oregon, USA.

Saeys, Y., Abeel, T., & Peer, Y.V.D. (2008). Robust feature selection using ensemble feature selection techniques. W. Daelemans, B. Goethals & K. Morik (Eds.), *Machine learning and knowledge discovery in databases* (pp 313-325) Springer. https://doi.org/10.1007/978-3-540-87481-2_21

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. Wiley Interdisciplinary Reviews: *Data Mining and Knowledge Discovery. 8*(4). e1249. https://doi.org/10.1002/widm.1249

Schapire, R.E. (2003). The boosting approach to machine learning: An overview. *Nonlinear Estimation and Classification*, 149-171. https://doi.org/10.1007/978-0-387-21579-2_9

Schroeder, M.A., Lander, J., & Levine-Silverman, S. (1990). Diagnosing and dealing with multicollinearity. *Western Journal of Nursing Research*, *12*(2), 175-187. https://doi.org/10.1177/019394599001200204

Sinharay, S. (2016). An NCME instructional module on data mining methods for classification and regression. Educational Measurement: *Issues and Practice*, *35*(3), 38-54. https://doi.org/10.1111/emip.12115

Skurichina, M., & Duin, R.P. (2002). Bagging, boosting and the random subspace method for linear classifiers. *Pattern Analysis & Applications*, *5*(2), 121-135. https://doi.org/10.1007/s100440200011

Steinki, O., & Mohammad, Z. (2015). Introduction to ensemble learning. *Available at SSRN*, *1*(1), 1-9. http://dx.doi.org/10.2139/ssrn.2634092

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323. https://doi.org/10.1037/a0016973

Subasi, A., El-Amin, M.F., Darwich, T., & Dossary, M. (2022). Permeability prediction of petroleum reservoirs using stochastic gradient boosting regression. *Journal of Ambient Intelligence and Humanized Computing*, *13*, 3555-3564. https://doi.org/10.1007/s12652-020-01986-0

Sutton, C.D. (2005). Classification and regression trees, bagging, and boosting. *Handbook of Statistics*, *24*, 303-329. https://doi.org/10.1016/S0169-7161(04)24011-1

Şevgin, H. (2020*). Predicting the ABIDE 2016 science achievement: The comparison of MARS and BRT data mining methods [Unpublished Doctoral Thesis]*. Gazi University.

Şevgin, H., & Önen, E. (2022). Comparison of Classification Performances of MARS and BRT Data Mining Methods: ABİDE-2016 Case. *Education and Science*, *47*(211). http://dx.doi.org/10.15390/EB.2022.10575

Tabachnick, B.G., & Fidell, L.S. (2015). *Using multivariate statistics* (6th edition). (M. Baloğlu, Trans.). Nobel Publications. (Original work published 2012).

Ting, K. M. (2017). Confusion matrix. In C. Sammut & G. I. Webb (Eds.) *Encyclopedia of Machine Learning and Data Mining* (pp. 260–260). Springer.

Tuğ Karoğlu, T.T., & Okut, H., (2020). Classification of the placement success in the undergraduate placement examination according to decision trees with bagging and boosting methods. *Cumhuriyet Science Journal*, *41*(1), 93-105. https://doi.org/10.17776/csj.544639

Wang, Z., Wang, Y., & Srinivasan, R.S. (2018). A novel ensemble learning approach to support building energy use prediction. *Energy and Buildings*, *159*, 109-122. https://doi.org/10.1016/j.enbuild.2017.10.085

Yurdugül, H. (2006). The comparison of reliability coefficients in parallel, tau-equivalent, and congeneric measurements. *Ankara University Journal of Faculty of Educational Sciences*, *39*(1), 15-37. https://doi.org/10.1501/Egifak_0000000127

Zhang, C., & Ma, Y. (2012). *Ensemble machine learning: methods and applications.* Springer. https://doi.org/10.1007/978-1-4419-9326-7

Zhou Z.H. (2012). *Ensemble methods: foundations and algorithms.* Chapman and Hall/CRC.

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 3, 563–579

https://doi.org/10.21449/ijate.1290901

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

The impact of peer feedback on collaborative problem-solving skills in the online environment

Yeşim Karadağ[1,*], **Seher Yalçın**[2]

[1]Ministry of Education, Ankara, Türkiye
[2]Ankara University, Faculty of Educational Sciences, Department of Measurement and Assessment, Ankara, Türkiye

ARTICLE HISTORY

Received: May 03, 2023
Revised: Aug. 12, 2023
Accepted: Sep. 05, 2023

Keywords:
Collaborative problem solving,
Peer feedback,
21st-century skills,
Collaborative learning,
Information technologies and software.

Abstract: The aim of this study is to investigate the impact of online peer feedback in the Information Technologies and Software course on the improvement of collaborative problem-solving skills (CPS), which are considered essential skills for the 21st century. The impact of peer feedback on CPS was designed using a mixed-methods approach that combines quantitative and qualitative data in the research. The study employed a set of CPS tasks that were specifically designed to measure the target behaviors outlined in the 6th-grade Information Technologies and Software course curriculum. Additionally, the study made use of various instruments, including a peer feedback survey, a longitudinal rating scale for tracking the development of CPS, and a peer feedback interview protocol. As a result, a statistically significant difference was found between the CPS of the experimental and control groups. According to the results of the study, it can be expressed that the CPS and peer feedback skills of the students in the experimental group have improved and that they have a more positive attitude toward giving peer feedback and solving collaborative problems based on the results of the interview form.

1. INTRODUCTION

The increasing dependency on information technologies, changing societal trends, and the creation of knowledge by using technology have made it how to manipulate these technologies, and almost essential to focus on this question: how will we teach individuals to navigate available communication channels?" (Csapo & Funke, 2017). The situation has made the definition of 21st-century skills an intriguing subject on a global scale. National organizations, international consortia, teacher forums, and government initiatives have conducted several studies on workforce requirements, e-learning environments, digital technologies, and their applications in conventional classrooms. Additionally, they have explored topics like resolving significant global issues, among others. Some of the institutions that have conducted these studies include the Partnership for 21st Century Skills, the Collective Intelligence Center, and the National Academy's National Research Journal. (Csapo & Funke, 2017). Because of technological, economic, political, and societal changes in the 21st century, 21st-century skills have become almost a necessity in defining qualified individuals. These changes have not only

*CONTACT: Yeşim Karadağ ✉ yesimm.karadagg@gmail.com Ministry of Education, Ankara, Türkiye

led to changes in the qualifications sought in the workplace but also made it almost essential to make significant changes in the information, skills, and competencies that personal requirements to acquire through education (Cansoy, 2018; Fiore et al., 2017). The new tasks assigned to the education system have also brought about changes in the organizational structure of education and the roles played by those working in these structures. The change in the expected qualifications has made it almost essential to leave routine and repetitive tasks to technology and impart complicated skill sets like solving complicated problems and collaboration to individuals. The New Vision for Education Report by the World Economic Forum, published in March 2015, emphasizes the significance of 21st-century skills like collaboration and problem-solving based on an analysis of research conducted in approximately 100 countries. The report states that "there are clear signs that in the 21st century, many students could not be getting the education they should succeed, and countries are not finding the skilled workers they should compete with other countries and meet the demands of the times" (World Economic Forum, 2015).

The report "The Future of Jobs" (2018) by the World Economic Forum highlights that complicated problem-solving, collaboration, and emotional intelligence, critical thinking, creativity, and people management skills were prominent in 2020 (Tusiad, 2019). The same report also indicates that complicated problem-solving skills, leadership, and social skills were among the demanded skills in the commercial world in 2022 (World Economic Forum, 2018). In the 21st century, team-based projects are becoming more prominent and require individuals to work in groups, which in turn help develop their communication skills (Barron, 2000). Collaborative problem-solving (CPS) has become more and more significant in our education system due to both the expectations of the era and skills it provides individuals. According to Nelson's (1998) CPS theory, learning environments that require solving everyday problems help students develop critical thinking, questioning, creativity, decision-making, and complicated problem-solving skills, while also assisting in their socialization.

Buder et al. (2015) describe CPS as an intricate system that requires participants to coordinate their own problem solutions into a consistent sequence of events. Jennings and Wooldridge (1999) define CPS as the process in which individuals work collaboratively toward a common goal. On the other hand, Clewley at al. (2017) define it as the process in which two or more people share their knowledge, skills, readiness, and efforts to solve a problem. Cuevas et al. (2017) describe CPS as a communication process that requires individuals to share their resources and strategies with other teammates to achieve a common goal. Considering these definitions, CPS is an approach in which multiple individuals pool their resources and strategies to solve complicated problems, using both mental and interpersonal abilities.

If the aim is to solve complicated problems, focus on group work processes, and develop collaborative working habits in individuals, CPS is a useful educational and teaching tool (Csapo & Funke, 2017). CPS is defined as an uninterrupted problem-solving process with a common goal and team spirit in which teammates support each other and aim to improve individual relationships and communication (Flood & Lapp, 1989). Solving a problem requires individuals with different perspectives to evaluate their views within the context of the problem at hand. Students' ideas being appreciated and accepted by other teammates develop a sense of belonging and acceptance in the group (Ashman & Gillies, 2013). In a CPS environment, teammates are constantly exposed to different ideas, which allows each member to listen to others' solutions and reflect on those ideas. These solution proposals allow individuals to interpret and reconstruct them based on their own thought processes (Gardunio, 2001). Additionally, motivation loss is less of a problem in little clusters, and coordination increases during activities (Huber & Huber, 2008). One of the primary elements of CPS is for teammates to continuously communicate with each other and provide feedback to one another.

CPS involves constant communication and feedback among teammates, which is one of the vital elements of the process. Feedback, in general, includes information on the current learning state and performance of individuals engaged in a learning activity in relation to the desired learning outcomes and performance levels (Çevik, 2014). With increasing class sizes, the practicality of teacher-centered feedback is limited, leading to a problem of inadequate teacher feedback. Peer feedback can be a pragmatic solution to overcome this problem (Macfarlane-Dick & Nicol, 2006). Also, peer feedback enhances student learning because it allows for social sharing and interaction to construct information (Bijami et al., 2013). Peer feedback is a feedback type that can be considered the equivalent of teacher feedback, which has a shaping effect on learning, and involves students using each other as a source of information (Anwar et al., 2019). Peer feedback highly the attracts the attention of students due to its social dimension. According to Falchikov (2005), this is because receiving feedback from peers is less anxiety-provoking than from teachers. Peer feedback also contributes to transforming learning environments where the assessment process is limited to the teacher in a participatory learning culture (Çevik, 2014).

It can be quite difficult for students to monitor and track peer feedback in the classroom. However, diverse open source platforms and online learning environments like "Synergy" facilitate the online presentation of collaborative peer feedback (Er et al., 2020). In an online environment, students tend to be more open and constructive in their feedback (Carless & Liu, 2006). Compared to traditional peer feedback, online platforms allow students to contribute to their peers' work in a more structured manner without limitations (Bayat et al., 2020). Online peer feedback applications provide opportunities for enriching the environment by implementing various scripts and structuring the peer feedback process. Additionally, they offer students flexibility in changing feedback that may not be possible in face-to-face or paper-based feedback (Bayat et al., 2020).

In accordance with the literature , it was found that studies on CPS are generally focused on mathematics classes (Aydın, 2020; Hogan et al., 1996; Kittur & Tausczik, 2014) and aimed to understanding and improving the conceptual structure of CPS (Arıcı, 2019; Karakuş, 2020; Uzunosmanoğlu, 2013; Roschelle & Teasley, 1995; Barron, 2000; Rummel & Spada, 2005; Andrews-Todd et al., 2018; Molnár et al., 2021) as well as examining the influence of CPS on academic achievement (James & Johnston, 1996). No studies have been found on the development of measurement and evaluation approaches that are compatible with the structure of high-level abilities like collaboration, problem-solving, information, media, and technology literacy, which are among the 21st-century skills.

In studies related to peer feedback, it has been observed that peer feedback is focused on language education and writing abilities and a restricted number of studies have been conducted in this area (Temizkan, 2009; Özşavli, 2017; Patri, 2002; Nilson, 2003; O'Dowd & Ware, 2008; Dochy et al., 2010). Various studies have been conducted to develop collaboration and problem-solving skills, which are among the 21st-century skills (Genç, 2007; Gök, 2006; Uysal, 2010); however, no studies have been found that investigate the function of peer feedback in online CPS environments. Given the importance of this ability in both day-to-day life and the workplace, and the insufficient amount of research, and the inadequacy of teacher feedback in CPS practices carried out in crowded classes, it is thought that peer feedback will reduce the workload of teachers. This study was deemed essential to address the uncertainty about how CPS abilities, which have a significant role in the Information Technologies and Software course aimed at developing mathematical thinking, problem-solving, algorithmic thinking, and creativity skills, will change based on peer feedback in an online learning environment.

Research questions:

1. Do the online discussions among the students in the experimental group make a significant difference in the achievement scores obtained from the Problems Developed for Collaborative Problem Solving Skills (IPCPS)?
2. Has there been an improvement in collaborative problem-solving skills assessed based on the Collaborative Problem-Solving Skills Grading Key (CPSSGK) in the experimental group because of the experimental study?
3. Are the students in the experimental group competent in giving feedback before the training? How is the peer feedback practice from the perspective of the students in the experimental group after the training?

2. METHOD

2.1. Study Design

The research examining the effect of peer feedback given online on CPS skills was designed in a mixed model in which quantitative and qualitative data were handled together. The mixed model is a type of research in which qualitative elements and quantitative research approaches are combined to provide answers to the research question that will increase the depth of understanding and accuracy (Clark & Creswell, 2017). Experimental method with pretest-posttest control group was used to obtain quantitative data. Experimental designs are research designs that aim to discover cause and effect relationships between variables. In the pre-test-post-test control group studies, there are two groups, one control group and the other experimental group, which were created by unbiased assignment in order to keep extraneous variables under control. In both groups, the effect of the independent variable on the dependent variable is investigated by making measurements before and after the experiment (Karasar, 2018). With these measurements, in-group and between-group differences were analyzed. In order to obtain qualitative data, a questionnaire and interview form were applied to the students in the experimental group. The dependent variable of this study is CPS skills and success. The independent variable of the research is peer feedback in the four-week online CPS study applied to the experimental group.

2.2. Study Group

The study group consists of 32 randomly selected 6th-grade students who attended Şehit Onbaşı Ahmet Şükrü Karataş Boarding Middle School, Mehmet Akif Ersoy Middle School, and Bayraktar Middle School located in Karayazı district of Erzurum province during the 2021-2022 academic years. The purpose of focusing on 6th grade students as the study group in the research is that it is easier to acquire collaborative problem solving skills at an early age and that this skill to be acquired at an early age can be used in later years and in academic life. In addition, considering that the sample group in PISA 2015, in which collaborative problem solving skills were measured, was 15-year-old students, it is considered important to determine whether this skill is acquired at an earlier age (Karakuş, 2020; Türkeş Yazıcı, 2022). The students in the investigation group have been randomly assigned to 16 experimental and 16 control groups. Of the 16 students in the experimental group, nine were girls and seven were boys; eight were boarding students and eight were regular students. Of the 16 students in the control group, seven were girls and nine were boys; eight were boarding students and eight were receiving formal education. The students in the experimental and control groups were randomly divided into paired study groups, and the invariance of the groups was ensured throughout the implementation process. The interview form for collecting qualitative data was obtained by conducting one-on-one interviews with 16 volunteers who were part of the experimental group and willing to participate in the investigation.

2.3. Data Collection Tools

2.3.1. *Peer feedback survey*

The peer feedback survey was developed by the researcher to define students' ideas about what peer feedback is and their preferences and approaches to giving feedback. In the survey development process, a literature review was conducted first. Considering the information obtained because of the literature review, an item pool was created, and a draft survey was created by selecting items from the item pool. The draft of the prepared survey was presented to an expert in Turkish Language Teaching and an expert in Measurement and Evaluation in Education to ensure the content validity. The current version of the survey draft was prepared after it was reorganized by taking into account the feedback provided by the experts. The current version of the survey was administered to 12 students studying at the same grade level (6th grade) who could represent the study group. As a result of the pre-application, the survey was finalized by making the essential corrections on statements like self-confidence and being objective, which the students had difficulty in understanding. In the preliminary application, it was deemed appropriate to keep the questionnaire as a five-point scale by taking into account the students' ability to distinguish the difference between the scale scores, their academic levels and age groups (Büyüköztürk, 2005). In the light of the information obtained as a result of the pre-application, the questionnaire was developed as a five-point scale consisting of 14 items.

2.3.2. *CPS skill-based analytical rating scale (CPS-ARS)*

CPS-ARS was developed by Aydın (2020). Aydın (2020) used 19 sub-skill areas specific to intellectual and interpersonal skills as criteria in the rubric, adhering to the theoretical framework established in the ATC21S project. The social skills included in the scoring key were action, interaction, accomplishing a task /perseverance, response skills, audience awareness, compromise, self-assessment, shared memory, and responsibility initiative. Cognitive skills were organizing, goal-setting, resource allocation, uncertainty tolerance, open-mindedness, collecting pieces of information, regularity, regularity, rules, and hypotheses. In the behavioral indicators of these sub-skill areas, there were descriptions of behavioral indicators as 1-3 (low), 4-6 (medium), and 7-9 (high). In accordance with the problem situations and scope used in the investigation, the authors adapted the rubric to the study by taking expert beliefs from the 19 criteria developed by Aydın (2020) and using nine criteria specific to this study. In the adapted version of the grading rubric, under the social skills category, the skills of action, interaction, accomplishing a task/perseverance, responsiveness, audience awareness, compromise, and responsibility initiative were scored. Under the cognitive skills category, the students' CPS skills were scored using sub-skills of uncertainty tolerance and hypothesis.

2.3.3. *CPS skills enhanced problems (CPSEP)*

Semi-structured open-ended problems for collaborative problem solving skills were developed for the Information Technologies and Software course to be used in the research. The reason for developing semi-structured problems in the context of Information Technologies and Software course is that the course outcome is suitable for measuring collaborative problem solving skills and the researcher has the knowledge and skills that will be needed in the study process since he is an Information Technologies and Software teacher. Semi-structured open-ended problems require a limited number of solutions, rules and solutions to categorize the answers into scales. Due to the nature of the study, the semi-structured open-ended problems "Information and Technology Week Classroom Board" and "Creating Awareness of Fighting the Epidemic" were developed by the researcher and two Information Technologies and Software teachers, while Gülse's Story was inspired by the National Education 6th grade Information Technologies and Software book and made suitable for the study. In order to determine the appropriateness of the developed and adapted problems that make up the study, the beliefs of an Information Technologies and Software expert, a Turkish language expert, and a Measurement and Evaluation expert were taken and the problems were finalized thanks to the formal and content improvement feedback. In the first and fourth weeks of the study, the same problems were used in both the experimental and control groups, while in the second and third weeks, two different problem situations were addressed in the experimental group. The answers given by the experimental and control group students to the problems were evaluated by the researcher and the measurement and evaluation specialist using a scoring key on a 15-point scale.

2.3.4. *Peer feedback interview form*

To obtain the qualitative data of the study, a semi-structured interview form developed by Özşavli (2017) was used. The peer feedback interview form developed by Özşavli (2017) consisted of 11 items. In the interview form developed by Özşavli, essential arrangements were made with the opinions of the Turkish teacher, the Information Technologies and Software teacher, and the Measurement and Evaluation expert, and at draft of the form was created. Based on expert opinions, the expressions in the items that made up the form were adapted to the age of the students. The current version of the draft interview form was applied to five students studying at the same grade level (6th grade) who could represent the study group, and the comprehensibility of the statements was tested and the form was finalized after receiving expert opinion. The interview form, which was adapted for the study, was applied by interviewing 16 students in the experimental group one-on-one.

2.4. Data Collection

In the 21st century, there are technological tools and environments for measuring complicated skill like CPS and analyzing feedback, one of its most important elements. However, the study was conducted out using technological tools and software in the classroom environment in order to increase the continuity of the study, to minimize the possibility of missing data, to include teacher observations in the process, to enable students to conduct pair group work in a healthier way and to record the data. As mentioned in Heller and Heller's (2010) study, teams of two people each were formed in the experimental and control groups due to the process would proceed more efficiently if computerized, audio and video recordings could be analyzed, and the increase in the number of people in the group would complicate the cooperation, interaction and communication structure. The research process covered a period of four weeks. The weekly sessions were conducted in two class hours, with 20 minutes allocated for the solution generation phase and 60 minutes for students to participate in online discussions and provide feedback to their peers. In the research process, Ethics Committee Approval, Application Permission Certificate from the Provincial Directorate of National Education, and essential permissions with the Parental Informed Consent Form were obtained due to the participants of the research being under the age of 18.

2.5. Data Analysis

With the GPower program, the sample size required for non-parametric tests at 80% power with a margin of error of .05 was calculated as at least 14 people in both groups. The minimum sample size required for parametric tests was 21 people in each group. The sample size reached in the investigation was 16 people for both groups. In this context, it has been evaluated that the sample is sufficient for non-parametric tests. Moreover, based on the available sample sizes and the mean and standard deviation values for the dependent variable, the power value calculated for different non-parametric statistical tests ranges between 91% and 100%. This points to the accuracy and reliability of the decisions to be taken based on the data obtained. The data obtained qualitatively by analyzing the audio recordings and videos; audio recordings, video recordings and written documents obtained with the Peer Feedback Survey, CPS-ARS, Peer Feedback Interview Form and Pre-test Post-test were quantitatively analyzed quantitatively by using various descriptive statistics with the SPSS 25 package program. In the analyses, .05 was accepted as the significance level. Cohen's effect size was calculated for statistically significant results. In the interpretation of this value, a d value of less than 0.2 means a weak effect size, 0.5 means a medium effect size, and greater than 0.8 means a large effect size (Kılıç, 2014). To answer the first sub-objective of the study, "whether online discussions among students in the experimental group create a significant difference in achievement scores obtained from the CPSEP," the descriptive statistics results of the achievement scores obtained from the problem-solving tasks developed for CPS for the experimental and control group students are presented in Table 1. To answer the second sub-objective of the study, "whether there is an improvement in CPS skills evaluated based on the CPS-ARS in the experimental group after the experimental study," the descriptive statistics results of the CPS skill scores obtained from the CPS-ARS for the experimental and control group students are also presented in Table 1.

Table 1. *CPSEP and CPS-ARS descriptive statistics results of achievement scores.*

			N	x̄	Median	Mode	Kurtosis	Skewness
CPSEP Achievement Scores	Experimental Group	Pre-test	16	5.87	6.00	5.00	-1.10	-0.15
		Post-test	16	12.87	12.50	12.00	-0.24	0.99
	Control Group	Pre-test	16	6.87	7.00	5.00	-1.23	-0.85
		Post-test	16	5.37	5.00	3.00	-0.95	0.43
CPS-ARS Achievement Scores	Experimental Group	Pre-test	16	27.37	27.50	30.00	1.23	0.70
		Post-test	16	44.69	45.00	40.00	-0.94	-0.57
	Control Group	Pre-test	16	24.81	25.00	28.00	0.59	-0.13
		Post-test	16	23.25	23.50	17.00	-1.17	-0.01

When Table 1 is examined, owing to the fact that the arithmetic means and median values of the scores obtained are quite close to each other indicates that the data are normally distributed. When the kurtosis and skewness values of the obtained scores are analyzed, it is seen that the values are generally between -1 and +1. However, since the kurtosis values of the experimental group pretest scores and the control group pretest scores were greater than 1 and the group size was less than 21 (the number required for the minimum power calculated according to the Gpower program), they were analyzed using the Wilcoxon Signed Ranks Test. In addition, to determine whether there was a difference between the pre-test mean scores of the students in the experimental and control groups and whether there was a difference between the post-test mean scores of the students in the experimental and control groups, the Mann Whitney U Test was used since the parametric test assumptions were not met and the groups were independent of each other.

To examine the third sub-objective of the study, "Are the students in the experimental group competent in giving feedback before the training?", the responses to the peer feedback survey were analyzed by descriptive analysis. Furthermore, data obtained from the peer feedback interview form were analyzed using a descriptive analysis method to obtain the opinions of the students in the experimental group on peer feedback application after education, and the opinions of the students are included in the analysis.

3. RESULTS

3.1. The Effect of Online Discussions among the Students in the Experimental Group on Achievement Scores Obtained from the CPSEP

The results that the students in the experimental and control groups obtained from the problems developed for CPS skills before and after the application are given in Table 2.

Table 2. *Results of pretest - posttest scores by group.*

	Group	N	Rank Mean	Rank Total	U	*p*
Pre-test	Experimental	16	15.00	240.00	104	.36
	Control	16	18.00	288.00		
Post-test	Experimental	16	24.50	392.00	000	.00
	Control	16	8.50	136.00		

As seen in Table 2, the results of the analysis show that there is no significant difference between the mean ranks of the pretest scores of the experimental and control group students (U=104, *p*>.05). When the rank averages are taken into consideration, it is seen that there is no significant difference between the achievement scores of the students who will and will not participate in the program in which the effect of peer feedback on CPS is examined.

The results of the analysis in Table 2 show that there is a statistically significant difference in the mean ranks of the posttest scores of the experimental and control group students (U=000, *p*<.05). Considering the mean ranks, it was found that the students who participated in the program in which the effect of peer feedback on CPS was examined had higher mean ranks than the students who did not participate. The detected effect size was r=0.86, indicating a large effect and explaining 73% of the total variance (Cohen, 1992). This finding demonstrates the effectiveness of peer feedback in CPS as the experimental procedure applied. The results of the students in the experimental and control groups from the problems developed for CPS skills before and after the application are given in Table 3.

Table 3. *Results of experimental and control group pretest - posttest scores.*

	Pre-test- Post-test	N	Rank Mean	Rank Total	z	*p*
Experiment al	Negative Rank	16	8.50	136.00	3.53*	.00
	Positive Rank	0	.00	.00		
	Equal	0				
Control	Negative Rank	2	2.50	5.00	1.848*	.066
	Positive Rank	6	5.17	31.00		
	Equal	8				

** Based on negative ranks*

The findings presented in Table 3 reveal a significant difference in pretest and posttest scores for the students in the experimental group who took part in the intervention. (z=3.53, *p*<.05). When results are taken into consideration, it is seen that this observed difference is in favor of the negative ranks, that is, the posttest. The magnitude of this difference is r=0.89, the difference has a large effect and explains 79% of the total variance (Cohen, 1992). Based on these findings, it can be said that the program in which the effect of peer feedback on CPS was examined had a significant effect on improving the achievement scores of the experimental group students.

The analysis results in Table 3 indicate that there is no significant difference between the pretest and posttest scores of the control group students who participated in the investigation (z=1.85, *p*>.05). Based on these findings, it is seen that there is no significant difference between the achievement test scores of the control group students who did not participate in the program in which the effect of peer feedback on CPS was examined.

3.2. Improvement in Cooperative Problem Solving Skills Based on CPS-ARS in the Experimental Group

The results of the scores they received from the longitudinal rubric developed for CPS skills in the first and last applications of the study are given in Table 4.

Table 4. *Test results of pretest - posttest scores by group.*

	Group	N	Rank Mean	Row Total	U	*p*
Pre-test	Experimental	16	18.09	289.50	102.50	.34
	Control	16	14.91	238.50		
Post-test	Experimental	16	24.2	387.50	4.50	.00
	Control	16	8.78	140.50		

As seen in Table 4, the results of the analysis show that there is no significant difference between the mean ranks of the pretest scores of the experimental and control group students (U=102.50, *p*>.05). When the rank averages are taken into account, it is seen that there is no significant difference between the level of development of CPS skills of the students who participated in the program in which the effect of peer feedback on CPS was examined and those who did not.

According to Table 4, the results of the analysis show that there is a significant difference between the mean ranks of the posttest scores of the experimental and control group students (U=000, *p*<.05). Considering the mean ranks, it was determined that the students who participated in the program in which the effect of peer feedback on CPS was examined (387.50) had a higher mean rank than the students who did not participate (140.50). The magnitude of this difference was r=0.82, the difference had a large effect and explained 67% of the total variance (Cohen, 1992). According to the results of the longitudinal rubric based on CPS skills, the program contributed to the development of the CPS skills of the students in the experimental group. The results of the students in the experimental and control groups on the longitudinal rubric developed for CPS skills before and after the application are given in Table 5.

Table 5. *Test results of experimental and control group pretest - posttest scores.*

	Pretest – Posttest	N	Rank Mean	Rank Total	z	*p*
Experimental	Negative Rank	16	8.50	136.00	3.517*	.00
	Positive Rank	0	.00	.00		
	Equal	0				
Control	Negative Rank	6	8.17	49.00	-.785	.43
	Positive Rank	6	4.87	29.00		
	Equal	4				

** Based on negative ranks*

The analysis results in Table 5 show that the experimental group students who participated in the application there was a significant difference between the pretest and posttest scores (z=3.52, *p*<.05). When the results are considered, it is seen that this difference is in favor of the posttest, that is, the negative ranks. It is seen that the magnitude of this difference is r=0.88, the difference has a large effect and explains 77% of the total variance (Cohen, 1992). Based on these findings, it can be said that the program in which the effect of peer feedback on CPS was examined had a significant effect on the development of CPS skills of the experimental group students.

The analysis results in Table 5 show that there is no significant difference between the pretest and posttest scores of the control group students who participated in the intervention (z= -.79, *p*>.05). Based on these findings, it is seen that there is no significant difference in the CPS skills of the control group students who did not participate in the program in which the effect of peer feedback on CPS was examined.

3.3. Experimental Group Students' Efficiency in Giving Feedback and Their Opinions on Post-Training Peer Feedback Practice

Table 6 presents the descriptive results of the answers given by the students participating in the investigation to the survey questions about peer feedback before the experimental study.

Table 6. *Peer feedback survey descriptive analysis results.*

Item	Mean	Mode	Median	Standard Deviation
1	3.33	5	4	1.68
2	3.87	5	5	1.10
3	2.67	1	2	1.63
4	3.40	3	3	1.24
5	3.80	5	5	1.82
6	3.33	3	3	1.05
7	3.53	2	4	1.25
8	3.93	4	4	0.96
9	3.07	2	3	1.39
10	2.33	1	1	1.80
11	2.87	2	2	1.46
12	3.60	5	4	1.45
13	2.47	3	3	1.06
14	2.93	3	3	1.34

According to the results of the descriptive analysis in Table 6, the item with the highest mean score is item 2 (I feel nervous if I receive especially negative comments from my teachers), while the item with the lowest mean score is item 10 (I do not take my classmates' feedback seriously). According to the second item, it is understood that teachers' comments are very effective on students. In item 10, it is understood that they care about their classmates' feedback. However, they also stated that their friends' feedback was superficial (I8: My classmates' feedback is superficial). The mean of item 8 was found to be 3.93. It is also the item to which the students gave the most homogeneous response among all statements (sd: 0.961).

In the investigation, the Peer Feedback Interview Form (PFIF) was applied to learn the opinions of the students participating in the investigation about the peer feedback activity they carried out during the process. The Peer Feedback Interview Form was created by making some changes in the semi-structured interview form consisting of 11 open-ended items developed by Özşavli (2017). The four categories that stand out from the student opinions obtained through the semi-structured interview form are presented below.

a) Benefits of Peer Feedback: The opinions of the student coded E2, who thinks that it contributed to helping and solution-oriented thinking, are presented below:
E2: *"Yes, it was useful, and I experienced the feeling of helping them. Yes, it made me work solution-oriented."*

b) Contribution of Peer Feedback to the Development of Various Skills: The opinions of the student coded I, who thinks that it improves communication skills, are presented below:
I: *"Seeing my own mistakes affected my success positively. I did not have much contact with my friends, now I started to have more contact, I talked and communicated more with my groupmate, and I did not have much communication with my other friends, now I am better."*
c) The Contribution of Peer Feedback to IPC Skills: The opinions of the student coded H, who thought that it contributed to taking responsibility, are presented below:
H: *"While we were avoiding our responsibilities before the study, now we are trying to take responsibility with my groupmate. I think I have improved socially."*
d) Problems Experienced During Peer Feedback: The opinions of the student coded G, who stated that he had difficulty in going beyond his own perspective and not being open to other ideas, are presented below:
G: *"Yes, while evaluating my friends' work, I had difficulty both in going beyond my own viewpoint of view and in understanding my friends' points of view."*

In order to learn the opinions of the experimental group students participating in the investigation on peer feedback, the data obtained from the peer feedback survey before the application and the peer feedback interview form after the application were analyzed. According to the data obtained from the survey used before the implementation, students generally stated that receiving feedback increased their self-confidence, while according to the comments obtained from the post-implementation interview form, all students agreed that peer feedback increased their self-confidence. According to the peer feedback survey data, while there were students who thought that receiving peer feedback was superficial before the intervention, and after the intervention, all students stated that they took their peers' feedback into consideration and that the feedback contributed to them in many ways. According to the peer feedback survey data before the intervention, very few students thought that feedback from their peers would facilitate their learning and contribute to better learning, while after the intervention, all students stated that peer feedback was useful for their learning and improved their CPS skills in various ways.

4. DISCUSSION and CONCLUSION

In this study, two different data collection tools were used to examine the effect of peer feedback on CPS skills. When the data obtained because of the longitudinal rubric to gauge CPS skills and the answers given to the problems to gauge CPS skills were analyzed, it was seen that peer feedback had a significant effect on CPSs.

In the pretest results applied to the experimental and control groups, it was observed that there was no difference among the achievement scores of the students, but the achievement scores of the experimental group students who had online discussions for four weeks increased significantly, whereas there was no significant change in the achievement scores of the control group. This result may be evidence that students' feedback to each other in a study environment where there is no teacher feedback is effective in increasing achievement scores. This finding is consistent with the findings of Gu et al. (2015). Gu et al. (2015) stated that students tend to conduct research for reasons like seeking new information, clarifying their ideas, and justifying themselves in CPS studies, which enables students to realize new learning in the investigation process and increase their academic achievement. Schwartz (1995) stated that different problem-solving tasks improved problem-solving skills in collaborative groups. Additionally, many findings have been found that cooperative learning, which forms the basis of cooperative problem-solving, increases success (Açıkgöz, 1990; Baykara, 2000; Bonner et al., 2002; Genç, 2007; Kneivel et al., 2003). When the longitudinal rubric data on CPS skills based on the scoring of the communication of the experimental and control groups while working in pairs were analyzed, it was observed that while there was no difference among the scores of the two groups in the pretest study, a notable rise was observed in the achievement scores obtained from the rubric of the experimental group students who received feedback from both their groupmates and their peers who participated in the online discussion for four weeks during the CPS process. According to this result, online discussions and feedback improved students' social and cognitive CPS skills.

When the findings were evaluated in general, there was no significant difference among the CPS skills of the students in the experimental and control groups before the implementation. According to the posttest results conducted after the implementation process, there was a positive development in the CPS skills of the experimental group students who received peer feedback and participated in online discussions. In addition, there was no improvement in the CPS skills of the students in the control group. Some studies in the literature (Bulu & Pedersan, 2012; Ge & Land, 2003; Karakuş, 2020; Wegerif, 2006) also support this finding. According to these studies, as students face complicated cognitive, metacognitive, and strategic challenges during the solution of a collaborative problem, students stated that their CPS skills like organizing and retrieving information, modeling, and monitoring solutions, presenting persuasive ideas, evaluating and reflecting improved (Bulu & Pedersan, 2012; Ge & Land, 2003; Karakuş, 2020; Wegerif, 2006). In their study, Atar and Yavuz (2020) stated that Turkish students' CPS competencies can be improved through various applications. Accordingly, the OECD (2013) states that students' problem-solving competence can be improved through a high-quality education process. In this context, it is thought that it is important to design the educational environments and processes for developing these skills. According to the findings obtained from the post-intervention interview form, all students stated that peer feedback increased their self-confidence, was useful for their learning, and improved their CPS skills in various aspects. It was concluded that there was a positive change in the students' views on peer feedback before and after the application and that the feedback they received from their peers was critical in CPS skills.

According to the peer feedback survey data before the implementation, very few students thought that feedback from their peers would facilitate their learning and contribute to better learning, while after the implementation, all students stated that peer feedback was useful for their learning and improved their CPS skills in various ways. It was determined that the findings obtained from the interviews with the students were generally consistent with the findings obtained because of the studies conducted in the literature. Carnell's (2000) interviews with students showed that they liked receiving peer feedback. They stated that talking to friends was easier than talking to a teacher, that they felt freer when talking to their friends, and that they could say what they wanted. Ur (1996) found that students appreciated being consulted and often made a serious effort to give feedback. Smith and Tillema (2000) and Gijbels et al. (2008) found that students had positive attitudes toward receiving feedback and found it effective for their learning. In a study conducted in Hong Kong, those who reviewed their peers' texts reported that this inspired them to write their own essays (Berggren, 2015). In this context, it can be stated that student views on peer feedback overlap with the views in the related literature.

Based on the results obtained, some suggestions for researchers and practice are presented: i) Peer feedback can be integrated into the entire learning process as it improves students' affective skills like criticism, respecting different perspectives, defending their ideas in front of the community, and self-expression; ii) since it was found that students' CPS competencies can be improved through various practices, it is recommended that the educational environments and the process offered to students should be designed for developing these skills; iii) students can be trained in giving advanced feedback as it was shown that students were not at a sufficient level in giving advanced feedback, and feedback was effective in increasing achievement scores. In addition, other researchers can i) examine the differences among the achievement scores, CPS skills, and feedback-giving skills of students working with virtual and peer collaborators, ii) study the differences among students' achievement scores when solving problems individually and when solving problems collaboratively, iii) examine the relationship among CPS skills and the time students spend solving the problem, iv) the effect of peer feedback on collaborative problem solving skills can be examined in larger samples, in different courses, with students from different age groups and educational levels.

Acknowledgments

The authors would like to thank all the students who voluntarily participated in this study. This study was presented as a paper at CMEEP 2022 8th International Congress on Measurement and Evaluation in Education and Psychology. This paper was produced from the first author's master’s thesis prepared under the supervision of the second author.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Ankara University, 31.05.2021, 10/221.

Authorship Contribution Statement

Yeşim Karadağ: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing-original draft. **Seher Yalçın:** Methodology, Supervision, and Validation.

Orcid

Yeşim Karadağ https://orcid.org/0000-0002-5919-2081
Seher Yalçın https://orcid.org/0000-0003-0177-6727

REFERENCES

Açıkgöz, K. (1990). İşbirliğine dayalı öğrenme ve geleneksel öğretimin üniversite öğrencilerinin akademik başarısı, hatırda tutma düzeyleri ve duyuşsal özellikleri üzerindeki etkileri [The effects of cooperative learning and traditional teaching on academic achievement, retention levels and affective characteristics of university students]. *A.Ü. Eğitim Bilimleri Fakültesi: I. Ulusal Eğitim Bilimleri Kongresi*. Ankara.

Andrews-Todd, J., Fiore, S.M., Foltz, P.W., Graesser, A.C., Greiff, S., & Hesse, F.W. (2018). Advancing the science of collaborative problem solving. *Psychological Science in the Public Interest, 19*(2), 59-92. https://doi.org/10.1177/1529100618808244

Anwar, D., Rozimela, Y., & Suryani, R.W. (2019). Exploring the effect of peer feedback and the students’ perceptions of the feedback on students writing skill. *International Journal of Secondary Education, 7*(4), 116. https://doi.org/10.11648/j.ijsedu.20190704.14

Arıcı, Ö. (2019). *Türk öğrencilerin PISA 2015 sonuçlarına göre aracılık modelleriyle işbirlikçi problem çözme becerilerine ilişkin faktörlerin incelenmesi [Investigation of factors related to Turkish students' collaborative problem solving skills with mediation models according to PISA 2015 results]* [Unpublished master’s thesis]. Ankara Üniversitesi, Ankara, Türkiye.

Ashman, A.F., & Gillies, R.M. (2003). *Co-operative learning*. Psychology Press.

Atar, H.Y., & Yavuz, E. (2020). An examination of Turkish students’ PISA 2015 collaborative problem-solving competencies. *International Journal of Assessment Tools in Education, 7*(4), 588-606. https://doi.org/10.21449/ijate.682103

Aydın, F. (2020). *7. sınıf öğrencilerinin matematik dersinde işbirlikli problem çözme becerilerinin gelişiminin izlenmesinde kullanılabilecek boylamsal bir test deseni [A longitudinal test design for monitoring the development of 7th grade students’* *collaborative problem solving skills in mathematics course]* [Unpublished master's thesis]. Gazi Üniversitesi, Ankara, Türkiye.

Barron, B. (2000). Achieving coordination in collaborative problem-solving groups. *The Journal of The Learning Sciences, 9*(4), 403-436. https://doi.org/10.1207/S15327809JLS0904_2

Bayat, A., Biemans, H.J., Ginkel, S., Hatami, J., Mulder, M., & Noroozi, O. (2020). Students online argumentative peer feedback, essay writing, and content learning: Does gender matter?. *Interactive Learning Environments, 28*(6), 698-712. https://doi.org/10.1080/10494820.2018.1543200

Baykara, K. (2000). *İşbirliğine dayalı öğrenme teknikleri ve denetim odakları üzerine bir çalışma [A study on collaborative learning techniques and locus of control]* [Unpublished master's thesis]. Hacettepe Üniversitesi, Ankara, Türkiye.

Berggren, J. (2015). Learning from giving feedback: A study of secondary-level students. *ELT Journal, 69*(1), 58-70. https://doi.org/10.1093/elt/ccu036

Bijami, M., Kashef, S.H., & Nejad, M.S. (2013). Peer feedback in learning English writing: Advantages and disadvantages. *Journal of Studies in Education, 3*(4), 91-97. https://doi.org/10.5296/jse.v3i4.4314

Bonner, B.L., Laughlin, P.R., & Miner, A.G. (2002). Groups perform better than individuals on letters-to-numbers problems. *Organisational Behaviour and Human Decision Processes, 88*, 605–620. https://doi.org/10.1016/S0749-5978(02)00003-1

Bulu, S.T., & Pedersen, S. (2012). Supporting problem-solving performance in a hypermedia learning environment: The role of students prior knowledge and metacognitive skills. *Computers in Human Behavior, 28*(4), 1162-1169. https://doi.org/10.1016/j.chb.2012.01.026

Büyüköztürk, Ş. (2005). Anket geliştirme [Survey development]. *The Journal of Turkish Educational Sciences, 3*(2), 133-151. https://dergipark.org.tr/en/pub/tebd/issue/26124/275190

Cansoy, R. (2018). Uluslararası çerçevelere göre 21. yüzyıl becerileri ve eğitim sisteminde kazandırılması [21st century skills according to ınternational frameworks and their acquisition in the education system]. *Journal of Human and Social Sciences Research, 7*(4), 3112-3134. https://orcid.org/0000-0003-2768-9939

Care, E., & Griffin, P. (2012). A framework for teachable collaborative problem solving skills. In J. Buder, E. Care, P. Griffin, F. Hesse, & K. Sassenberg (Eds.), *Assessment and Teaching of 21st Century Skills* (pp. 37-56). Springer.

Carless, D., & Liu, N.F. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education, 11*(3), 279-290. https://doi.org/10.1080/13562510600680582

Carnell, E. (2000). Dialogue, discussion and secondary school students on how others help their learning. In S. Askew (Ed.). *Feedback for learning* (pp. 46-61). Routledge.

Clark, V.L.P., & Creswell, J.W. (2017). *Designing and conducting mixed methods research.* Sage publications.

Clewley, D., Dowell, N., & Graesser, A.C. (2017). *Assessing collaborative problem solving through conversational agents. In Innovative assessment of collaboration*. Springer.

Csapo, B., & Funke, J. (2017). *Educational research and innovation the nature of problem solving using research to inspire 21st century learning: Using research to inspire 21st century learning*. OECD Publishingcare.

Cuevas, H.M., Fiore, S.M., Scielzo, S., & Salas, E. (2002). Training individuals for distributed teams: Problem solving assessment for distributed mission research. *Computers in Human Behavior, 18*(6), 729-744. https://doi.org/10.1016/S0747-5632(02)00027-4

Çevik, Y.D. (2014). Dönüt alan mı memnun veren mi? Çevrimiçi akran dönütü ile ilgili öğrenci görüşleri [Is the receiver or the giver of feedback satisfied? Student views on online peer feedback]. *Journal of Instructional Technologies and Teacher Education, 3*(1), 10-23. https://dergipark.org.tr/en/pub/jitte/issue/25083/264720

Dochy, F., Gielen, S., Onghena, P., Peeters, E., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction, 20*(4), 304–315. https://doi.org/10.1016/j.learninstruc.2009.08.007

Er, E., Dimitriadis, Y., & Gašević, D. (2020). Collaborative peer feedback and learning analytics: theory-oriented design for supporting class-wide interventions. *Assessment & Evaluation in Higher Education, 46*(2), 169-190. https://doi.org/10.1080/02602938.2020.1764490

Falchikov, N. (2005). *Improving assessment through student ınvolvement: Practical solutions for aiding learning in higher and further education*. Routledge.

Fiore, S.M., Graesser, A., Greiff, S., Griffin, P., Gong, B., Kyllonen, P., & von Davier, A. (2017). In S. Fiore (Ed.). *Collaborative problem solving: Considerations for the national assessment of educational progress.* National Center for Education Statistics.

Flood, J., & Lapp, D. (1989). Cooperative problem solving: Enhancing learning in the secondary science classroom. *The American Biology Teacher*, *51*(2), 112-115. https://doi.org/10.2307/4448864

Gardunio, E.L.H. (2001). The influence of cooperative problem solving on gender differences in achievement, self-efficacy, and attitudes toward mathematics in gifted students. *Gifted Child Quarterly, 45*(4), 268-282. https://doi.org/10.1177/001698620104500405

Ge, X., & Land, S.M. (2003). Scaffolding students' problem-solving processes in an illstructured task using question prompts and peer interactions. *Educational Technology Research and Development, 51*(1), 21–38. https://doi.org/10.1007/BF02504515

Genç, M. (2007). *İşbirlikli öğrenmenin problem çözmeye ve başarıya etkisi [The effect of cooperative learning on problem solving and achievement]* [Unpublished doctoral thesis]. Marmara University.

Gök, T. (2006). *Fizik eğitiminde işbirlikli öğrenme gruplarında problem çözme stratejilerinin öğrenci başarısı, başarı güdüsü ve tutumu üzerindeki etkileri [The effects of problem solving strategies on student achievement, achievement motivation and attitude in cooperative learning groups in physics education]* [Unpublished doctoral thesis]. Dokuz Eylul University.

Gijbels, D., Segers, M.S.R., & Thurlings, M. (2008). The relationship between students' perceptions of portfolio assessment practice and their approaches to learning. *Educational Studies, 34*(1), 35-44. https://doi.org/10.1080/03055690701785269

Gu, X., Chen, S., Lin, L., & Zhu, W. (2015). An intervention framework designed to develop the collaborative problem-solving skills of primary school students. *Educational Technology Research and Development, 63*(1), 143-159. https://doi.org/10.1007/s11423-014-9365-2

Heller, K., & Heller, P. (2010). *Cooperative problem solving in physics a user's manual*. https://www.aapt.org/conferences/newfaculty/upload/coop-problem-solving-guide.pdf

Hogan, D.M., Tudge, J.R., & Winterhoff, P.A. (1996). The cognitive consequences of collaborative problem solving with and without feedback. *Child Development, 67*(6), 2892-2909. https://doi.org/10.1111/j.1467-8624.1996.tb01894.x

Huber, G.L., & Huber, A.A. (2008). *Structuring group interaction to promote thinking and learning during small group learning in high school settings* (pp. 110-131). Springer.

James, R.H., & Johnston, C.G. (1996). *An evaluation of the effectivess of collaborative problem-solving for learning economics* (No.534). The University of Melbourne: Melbourne.

Jennings, N.R., & Wooldridge, M. (1999). The cooperative problem-solving process. *Journal of Logic and Computation*, *9*(4), 563-592. https://doi.org/10.1093/logcom/9.4.563

Karakuş G. (2020). *İşbirlikli problem çözme öğretim programı tasarısının hazırlanması ve uygulanması [Preparation and ımplementation of cooperative problem solving curriculum design]* [Unpublished doctoral thesis]. Afyon Kocatepe University.

Karasar, N. (2018). *Bilimsel Araştırma Yöntemi [Scientific Research Method*] (33. Edition). Nobel Yayıncılık.

Kilic, S. (2014). Effect size. *Journal of Mood Disorders*, *4*(1), 44-46. https://doi.org/10.5455/jmood.20140228012836

Kittur, A., Kraut, R.E., & Tausczik, Y.R. (2014, February). *Collaborative problem solving: A study of mathoverflow*. In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing, USA.

Kneivel, E.M., Laughlin, P.R., Tan, T.K., & Zander, M.L. (2003). Groups perform better than the best individuals on the letters-to-numbers problems: Informative equations and effective strategies. *Journal of Personality and Social Psychology*, *85*, 684–694. https://doi.org/10.1037/0022-3514.85.4.684

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

Molnár, G., Pásztor-Kovács, A., & Pásztor, A. (2021). Measuring collaborative problem solving: research agenda and assessment instrument. *Interactive Learning Environments*, *1*(21). https://doi.org/10.1080/10494820.2021.1999273

Nelson, L.M. (1998). *Collaborative problem-solving: An instructional theory for learning through small group interaction* (pp. 1-143). ProQuest Dissertations Publishing.

Nilson, L.B. (2003). Improving student peer feedback. *College Teaching*, *51*(1), 34-38. https://doi.org/10.1080/87567550309596408

O'Dowd, R., & Ware, P. (2008). Peer feedback on language form in telecollaboration. *Language Learning & Technology*, *12*(1), 43-63. https://llt.msu.edu/vol12num1/wareodowd/

OECD (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. OECD publishing.

Özşavli, M. (2017). *Akran geri bildiriminin Türkçeyi yabancı dil olarak öğrenen öğrencilerin yazma becerisine etkisi [The effect of peer feedback on the writing skills of students learning Turkish as a foreign language]* [Unpublished master's thesis]. Mustafa Kemal University.

Patri, M. (2002). The influence of peer feedback on self-and peer-assessment of oral skills. *Language Testing*, *19*(2), 109-131. https://doi.org/10.1191/0265532202lt224oa

Roschelle, J., & Teasley, S.D. (1995). The construction of shared knowledge in collaborative problem solving. In E. Lehtinen (Ed.). *Computer Supported Collaborative Learning* (pp. 69-97). Springer.

Rummel, N., & Spada, H. (2005). Learning to collaborate: An instructional approach to promoting collaborative problem solving in computer-mediated settings. *The journal of the Learning Sciences*, *14*(2), 201-241. https://doi.org/10.1207/s15327809jls1402_2

Schwartz, D.L. (1995). The emergence of abstract representations in dyad problem solving. *The Journal of the Learning Sciences*, *4*, 321-354. https://doi.org/10.1207/s15327809jls0403_3

Smith, K., & Tillema, H.H. (2000). Learning from portfolios: Differential use of feedback in portfolio construction. *Studies in Educational Evaluation*, *26*(3), 193-210. https://doi.org/10.1016/S0191-491X(00)00015-8

Temizkan, M. (2009). Akran değerlendirmenin konuşma becerisinin geliştirilmesi üzerindeki etkisi [The effect of peer review on the development of speaking skills]. *Journal of Mustafa Kemal University Institute of Social Sciences*, *6*(12), 90-112. https://dergipark.org.tr/en/pub/mkusbed/issue/19557/208435

Türk Sanayicileri ve İşadamları Derneği, TUSIAD (2019). *Sosyal ve Duygusal Öğrenme Becerileri Raporu. (TÜSİAD-T/2019-11/609*). https://tusiad.org/tr/yayinlar/raporlar/item/10450-sosyal-ve-duygusal-ogrenme-becerileri

Türkeş Yazıcı, A. (2022*). Ortaokul 6. sınıf öğrencilerinin iş birlikli problem çözme becerilerinin incelenmesi [Examination of cooperative problem solving skills of middle school 6th grade students]* [Unpublished master's thesis]. Adnan Menderes University.

Ur, P. (1996). *Discussions that work: Task-centred fluency practice*. Cambridge University Press.

Uysal, G. (2010*). İlköğretim sosyal bilgiler dersinde işbirlikli öğrenmenin erişiye, problem çözme becerilerine, öğrenme stillerine etkisi ve öğrenci görüşleri [The effect of cooperative learning on achievement, problem solving skills, learning styles in elementary social studies course and students' opinions]* [Unpublished doctoral thesis]. Dokuz Eylul University.

Uzunosmanoğlu, S.D. (2013). *Examining computer supported collaborative problem solving processes using the dual-eye tracking paradigm* [Unpublished master's thesis]. Middle East Technical University.

Wegerif, R. (2006). A dialogic understanding of the relationship between CSCL and teaching thinking skills. *International Journal of Computer-Supported Collaborative Learning, 1*(1), 143-157. https://doi.org/10.1007/s11412-006-6840-8

World Economic Forum. (2015*). New vision for education: Unlocking the potential of technology*. Vancouver, BC: British Columbia Teachers' Federation.

World Economic Forum Boston Consulting Group (BCG). (2018). *Towards a reskilling revolution: A future of jobs for all.* World Economic Forum, Geneva, Switzerland.

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 3, 580–593

https://doi.org/10.21449/ijate.1262199

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

Application of the professional maturity scale as a computerized adaptive testing

Süleyman Demir[1,*], Derya Çobanoğlu Aktan[2], Neşe Güler[3]

[1]Sakarya University, Faculty of Education, Department of Educational Sciences, Sakarya, Türkiye
[2]Hacettepe University, Faculty of Education, Department of Educational Sciences, Ankara, Türkiye
[3]İzmir Democracy University, Faculty of Education, Department of Educational Sciences, İzmir, Türkiye

ARTICLE HISTORY

Received: Mar. 08, 2023

Revised: July 29, 2023

Accepted: Sep. 09, 2023

Keywords:

Computerized Adaptive Test (CAT),
CAT application,
Item selection method,
Stopping rule,
Vocational maturity.

Abstract: This study has two main purposes. Firstly, to compare the different item selection methods and stopping rules used in Computerized Adaptive Testing (CAT) applications with simulative data generated based on the item parameters of the Vocational Maturity Scale. Secondly, to test the validity of CAT application scores. For the first purpose, simulative data produced based on Vocational Maturity Scale item parameters were analyzed under different item selection methods (Maximum Fisher Information [MFI],Maximum Likelihood Weighted Information [MLWI] Maximum Posterior Weighted Information [MPWI] Maximum Expected Information [MEI] Minimum Expected Posterior Variance [MEPV] Maximum Expected Posterior Weighted Information [MEPWI]) and stopping rules (Standard Error [SE]<0.30, SE<0.50, SE <0.70, Number of Item [NI]=10, NI=20) by calculating the average number of items, standard error averages, correlation coefficients, bias, and RMSE statistics. For all the conditions of the item selection methods, standard error averages, correlation coefficients, bias, and RMSE statistics showed similar results. When the average number of items is considered, MFI and SE<0.30 were found as most appropriate methods to be used in CAT application. For the second purpose of the study, the paper-pencil form of the Vocational Maturity scale and CAT version were administered to 33 students. A moderate, positive, and statistically significant relationship was found between the CAT application scores and the paper-pencil form scores on the vocational maturity scale. As a result, it can be said that the vocational maturity scale can be applied as a computerized adaptive test and can be used in career guidance processes.

1. INTRODUCTION

The measurement results, which are the foundation of decisions to be made in education and psychology, must be reliable and valid. Decisions made with unreliable and invalid measurement results lead to erroneous evaluations of individuals, teaching methods, and programs. Validity is defined as the process of gathering evidence to support the decisions to be made based on the measurement results. Reliability, on the other hand, is expressed as the degree to which the results obtained from the measurement tool are free from random errors

*CONTACT: Süleyman Demir ✉ suleymand@sakarya.edu.tr Sakarya University, Faculty of Education, Department of Educational Sciences, Sakarya, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

(American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], (2014).

The high validity of the measurement results shows that, according to the definition of validity made by Thorndike and Hagen (1961), the measurement results are only related to the variable that is intended to be measured, and that no other feature interferes with the measurement results except for this variable (significantly) (Thorndike & Hagen, 1961; Turgut & Baykul, 2013). These features may be related to the variable to be measured, as well as to include constant, systematic and random errors. The fact that random error does not interfere with the measurement results positively affects both validity and reliability. Therefore, reliable results are required to ensure the results of the validity of the measurement tool. Because unreliable measurement results cannot be valid, the measurement tool should be as reliable as possible with as little random error as possible. According to classical test theory, increasing the number of items in the measurement tool and controlling the sources of random errors as much as possible increases the reliability and thus the validity of the measurement results. Although the number of items in the measurement tool increases the reliability, this increase may cause individuals to lose motivation and fatigue. As a result of this situation, individual-related random errors in measurement occur.

In education and psychology, measurement methods with fewer items have been developed to reduce random errors caused by individuals. Computerized Adaptive Test (CAT) applications are one of these measurement methods. CAT applications can estimate ability levels with fewer items than traditional paper-pencil tests (Gardner et al., 2004; Gibbons et al., 2016; Hol et al., 2008; Kaskatı, 2011; Penfield, 2006; Stochl et al., 2016; Petersen et al., 2016). In traditional paper-pencil tests, individuals answer all items, while in CAT applications they only answer the items relevant to their ability level. The instantaneous ability level is calculated in CAT applications after each item that the individual answers while taking the test. The final ability level calculated for the individual as a result of the CAT application is expected to be similar to the actual ability level. In CAT applications, while the individual answers items based on his or her ability level, he or she is not required to answer items that do not provide information about himself or herself, in other words, items that are higher or lower than his or her ability level (Linacre, 2000; Reckase, 1989; van der Linden, 1998). CAT applications are composed of five basic components: an item pool, a test initiation method, an item selection method, an ability estimation method, and a stopping rule (Dodd et al., 1995; Reckase, 1989; Thompson & Weiss, 2011; Wise & Kingsbury, 2000).

For the reliability and effectiveness of CAT applications to be high, the appropriate components must be used. Monte Carlo simulation studies have been conducted on simulative item parameters and post hoc simulation studies conducted with true item parameters in the literature, methods that allow obtaining measurement results with a high level of validity have been tried to be specified. Furthermore, it is seen that the item selection method is the focus of the vast majority of these studies (Choi & Swartz, 2009; Penfield, 2006; van der Linden, 1998; Veldkamp, 2003).

The item selection method component was defined by Choi and Swartz (2009) as the core of CAT applications, and it was stated that administering items appropriate for the individual's ability level will increase the effectiveness of CAT applications. Item selection methods are examined in two categories: traditional methods and Bayesian methods. Bayesian methods perform item selection methods based on the final distribution, while traditional methods perform item selection based on the item information function.

The Maximum Fisher Information (MFI) method is one of the traditional methods for the item selection in CAT applications. In the MFI method, the item that provides the most information for the instantaneous ability level estimated based on the individual's responses is administered.

In cases where the instantaneous ability level and the true ability level differ, the standard error amount increases because the item used will not be suitable for the true ability level (Hambleton et al., 1991; Penfield, 2006; Thissen & Mislevy, 2000; van der Linden & Pashley, 2000). The MFI method is defined by Lord and Novick (1968) as the Attenuation Paradox, the condition that the reliability and therefore the validity of the measurement results are low despite applying items with maximum information for the instantaneous ability level of individuals.

Figure 1. *Representation of the Attenuation Paradox.*

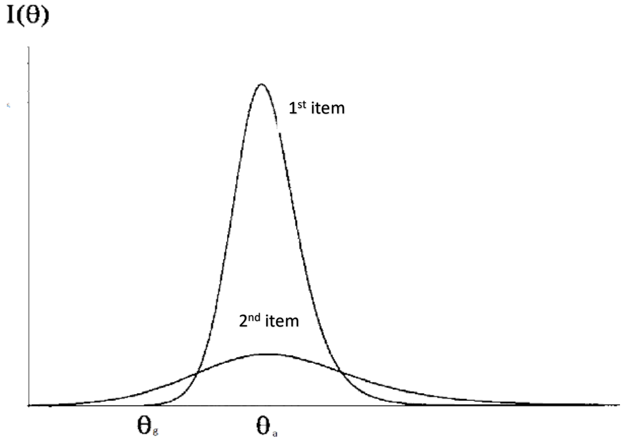


Figure 1 shows the information functions of two different items. θ_g shows the individual's true ability level, while θ_a shows the instantaneous ability level. The first item provides less information on the individual's true ability level (θ_g) than the second item, while the second item provides more information on the individual's instantaneous ability level (θ_a). Therefore, despite providing less information about the actual ability level, the MFI method favors the first item. In this case, the measurement results will be inaccurate due to the application of the item, which provides little or no information on the individual's true ability level.

Another disadvantage of the MFI method is that it results in excessive use of these items due to constant use. Excessive use of some items in maximum information-based methods causes all items not be used and the measurement precision is very high (Davis, 2002; Davis & Dodd, 2008). To avoid the Attenuation paradox and excessive item use, Bayesian statistical approaches to item selection have been developed rather than methods based solely on instead of on information function. In the study conducted by Boztunç-Öztürk and Doğan (2015), it was found that whether item exposure is controlled in maximum information-based and Bayesian item selection methods does not make a significant difference in terms of measurement precision but when item exposure is not controlled in maximum information-based methods, all of the items in the item pool are not used. Not using all of the items is related to item pool size as much as item exposure control (Leroux et al., 2019). In the study conducted by Leroux et al. (2019), it was found that when the item pool is small, all of the items in the item pool are used when even the item exposure is not controlled. Considering this information available in the literature, it can be said that using item selection methods based on maximum information when the item pool is small will not result in high measurement precision or not using all items.

Most CAT research with polytomous models, comparing Bayesian item selection methods and traditional methods appears to be post hoc simulation studies (Choi & Swartz, 2009; Passos et al., 2007; Penfield, 2006; van Rijn et al., 2002; Veldkamp, 2003). While it is expected that measurement results in CAT applications that use the ability estimation method and stopping rules, particularly the item selection method studied in a simulative environment, will have a low level of error, studies on polytomous CAT applications have been limited by method comparisons in a simulative environment, and the number of application-oriented CAT studies has been limited (Aybek & Demirtaşlı, 2017). Given that measurement tools in the fields of education and psychology are mostly used with paper and pencil in Türkiye, it is believed that increasing the use of CAT applications will be more beneficial to both researchers and participants.

This study compares different item selection methods with the actual application of the "Occupational Maturity Scale" a commonly used educational measurement tool with 40 items and one dimension (Akdaş & Ekinci, 2016; Aktuğ & Birol, 2011; Kutlu, 2012; Orhan & Ültanır, 2011; Sahranç, 2000; Sürücü, 2005; Ulaş & Yıldırım, 2015; Ürün, 2010). The concept of occupational maturity is defined by Super (1957) as meeting the requirements of each professional development step, being ready for the next development step, and having basic abilities that can overcome the difficulties that may be encountered (Kuzgun & Bacanlı, 1995). The feasibility of the Occupational Maturity Scale as a computerized adaptive test is tested, and the amount of error, bias, and correlation coefficients are calculated using the CAT application's simulation under various item selection methods and stopping rules. The relationship between the scores obtained from the CAT application and the paper-pencil test was investigated using minimum error methods with these coefficients.

While data obtained with the scale are more objective, valid, reliable, and useful than data obtained through non-test techniques (such as observation, interview, etc.), there may be random errors in the measurement results due to factors such as low motivation in individuals in answering the scale items, social desirability, psychological fatigue, and the length of the scale. As a result of this situation, researchers are focusing on alternative data collection methods rather than traditional paper-pencil methods. One such method is CAT applications, which can estimate the ability associated with the actual ability level at a high level with a much smaller number of items. This estimate is based on the application of items appropriate to the individual's own ability level. Therefore, selecting the appropriate item for the individual is critical to the effectiveness of CAT applications. However, there is limited research on the comparison between traditional item selection methods and Bayesian item selection methods in real CAT applications and post-hoc simulation studies (Aybek & Demirtaşlı, 2017; Choi & Swartz, 2009; Passos et al., 2007; Penfield, 2006; van der Linden, 1998; van Rijn et al., 2002; Veldkamp, 2003). Thus, this study's results are expected to contribute to both the occupational guidance process and the usability of CAT in scientific studies.

1.1. Research Problems

This study aims to address the following research questions using CAT applications:

1. Does the correlation coefficient between the simulatively estimated occupational maturity level and the actual occupational maturity level differ depending on the item selection method and stopping rules used, mean number of items administered, standard error means, bias and RMSE values?
2. Is there a relationship between the CAT application and the occupational maturity levels obtained from the paper-pencil test application using the determined item selection method and stopping rule?

2. METHOD

2.1. Participants

Before starting to collect data, permission was obtained from Hacettepe University Ethics Commission with the decision dated 24.10.2017 and numbered 433-3695. This study's data were gathered from two distinct groups of participants. The first group's data were employed to determine the item parameters of the Occupational Maturity Scale and to test the IRT assumptions. The second group's data was used to test the validity of the CAT application of the Occupational Maturity Scale. Table 1 shows the demographic information of the first and second groups.

Table 1. *Demographic information of the participants.*

			Frequency	Percentage
The first group	Gender	Female	366	54.06%
		Male	311	45.94%
	Class	11th Grade	510	75.33%
		12th Grade	167	24.67%
The second group	Gender	Female	16	48.48%
		Male	17	52.52%

The first group consisted of 677 students in the 11th and 12th grades from Adapazarı, Erenler, Hendek and Serdivan districts of Sakarya. For this data group, firstly, permission was obtained from Sakarya National Education Directorate. 12 high schools were determined by cluster sampling from high schools located in Serdivan, Erenler, Hendek and Adapazarı. Data were collected from 677 students on a voluntary basis from 11th and 12th grade students studying in these high schools. Of these 677 students, 366 were female and 311 were male; 510 of them are in the 11th grade and 167 are in the 12th grade.

The second group consisted of 33 students in the 11th and 12th grade from Private School of Sakarya University Foundation. 33 students on a voluntary basis CAT application of the Occupational Maturity Scale and the paper-pencil test application were carried out. When the literature is examined, it is recommended to be 1-2 weeks between the two applications (Bardhoshi & Erford 2017; Cattell, 1986; Cattell et al., 1970; Deyo et al., 1991; Nunnally & Bernstein, 1994), and the application was made by leaving 10 days between the CAT and paper-pencil test applications.

2.2. Data Collection Tools and Methods

The occupational Maturity Scale used in this study was developed by Kuzgun and Bacanlı (2005). The scale consists of 40 items with one dimension and was reported to have an internal consistency coefficient of 0.89, and a test-retest reliability coefficient of 0.82. The scale was administered to the first group of participants to obtain the item parameters of the Occupational Maturity Scale and to test the IRT assumptions.

Based on these data, simulative data were produced according to different item selection methods and stopping rules with the FIRESTAR program using the item parameters of the Occupational Maturity Scale, and the CAT application was prepared with the CONCERTO platform.

2.3. Data Collection

To create an item pool, there must be at least 24-30 items that can provide information at all ability levels. However, having a certain number of items does not necessarily mean that the CAT application will be sufficient. The item information and test information functions are also critical for CAT applications (Dodd et al., 1995).

The item parameters of the Occupational Maturity Scale were determined using the IRTPRO package program. To determine the item parameters, a one-dimensional Item Response Theory (IRT) analysis was performed under a graded response model. The analysis revealed that, the step parameters of the four items (2nd, 4th, 6th and 20th) were outside the ranges (-4.00, +4.00), which are the lower and upper limits for the IRT models. Four items were removed from the item pool because the a_i parameter was less than 0.60, the information functions were weak, and data production did not occur when these items were in the item pool while generating simulative data.

With the use of the simulated data generated by the FIRESTAR software, various item selection strategies and stopping rules were explored to assure the greatest effect from the CAT application of the Occupational Maturity Scale, whose item parameters were established. Using the simulative data, the average number of items, the correlation between the CAT application and the occupational maturity levels determined by the paper-pencil exam, the bias and the RMSE statistics were calculated. Table 2 lists the methods for the item selection and the stopping rules.

Table 2. *The methods of selection of items used in the simulation and the rules of termination.*

Manipulated Variable	Methods	Number of Conditions
Item Selection Method	Maximum Fisher Information (MFI) Maximum Likelihood Weighted Information (MLWI) Maximum Posterior Weighted Information (MPWI) Maximum Expected Information (MEI) Minimum Expected Posterior Variance (MEPV)	6
Stopping Rule	Standard Error < 0.30 Standard Error < 0.50 Standard Error < 0.70	3

	NI=10
	NI=20
		2

Table 2, shows that a total of 30 different conditions have occurred under different item selection methods and different stopping rules. A total of 30,000 people's simulated data were produced, with 1,000 people in each situation. In the study, while the item selection method and the stopping rule were manipulated, other variables held constant in the study are listed below.

- Selection of the first item: 0 ability level (θ=0.00)
- Sample mean and standard deviation: *sd*=1.00
- Frequency of item use control: Not used (coded as 1).
- Ability estimation method: Expected Final Estimation Method
- Minimum and maximum ability levels: -4.00, +4.00
- IRT model: Graded Response Model
- Scaling: 1.7
- Ability increase value: 0.10
- Standard error calculation method: Final distribution
- A priori distribution: $\bar{X} = 0.00$, sd=1.00

2.4. Computerized Adaptive Testing Application

During the spring semester of the 2017-2018 academic year, the paper-pencil form and the CAT version of the occupational maturity scale were administered to 33 (11th and 12th grade) students. The paper-pencil test was administered in the students' own classrooms at the beginning of their course by the researcher. Ten days after the paper-pencil form was administered to the students, the CAT version was administered to the participants. The CAT was carried out by a researcher in the computer laboratory located in the same building as the classrooms at the school. Students' transportation from their classrooms to the laboratory was provided by guidance counsellors and course teachers. Since the test was online, all computers and the internet connection were checked and the relevant web page was opened and made available for students' use. Students logged in by entering their student numbers and gender to give feedback to the CAT application and to match it with the paper-pencil test. After entering the necessary information, the students clicked on the "Continue" button to begin. The student's answers to the items in the CAT application were recorded in the database. If a student wanted to pass an item without answering, a warning message "Please do not pass without answering the Item!" appeared. At the end of the CAT application, which continued until the specified condition was met, an information screen about the Occupational Maturity ability level was displayed.

2.5. Analysis of the Data

To answer the first research problem, correlation coefficient, standard error mean, bias and RMSE statistics were calculated using IRTPRO, Excel and SPSS 17.0. Then these values were examined to determine whether they differed.

• The IRTPRO package program was used to determine the item parameters of the Occupational Maturity Scale. To use the IRTPRO program, a 15-day trial version was rented from Scientific Software International by e-mail at 2018. As a result of the analysis performed under the Progressive Response Model, a_i and β_{ij} were calculated.

• To analyze the simulated data, the correlation coefficient (Pearson Product Moments Correlation Coefficient), and the average number of items applied, the standard error mean (SE), bias, and RMSE statistics were calculated between the simulatively estimated and actual occupational maturity levels for each condition by using Excel and SPSS 17.0. High correlation coefficient, low standard error, bias, and low RMSE statistics (close to 0) indicate that there is no difference (deviation) between individuals' true ability level and estimated ability level. The methodology for calculating bias, RMSE, and standard error averages—three statistics used to compare various stopping rules—is described here:

• The standard error can be calculated in two different ways according to the IRT, depending on the information function and depending on the final distribution. During the production of simulative data, standard error calculation was performed depending on the final distribution.

$$SE_{post} = \sqrt{\mathrm{var}(g(\theta)|U)}$$

• The bias statistic is equal to the average of the difference between the actual value of a parameter and the estimated value.

$$BIAS = \frac{\sum_{i=1}^{n}(\theta_{ig}-\theta_{ik})}{n}$$

• The RMSE statistic is the average of the squares of the difference between the true value and the predicted value of a parameter.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\theta_{ig}-\theta_{ik})^2}{n}}$$

To answer the second research problem, the correlation coefficient was calculated. The correlation coefficient between the CAT application and the paper-pencil application was calculated using the SPSS 17.0 program.

3. FINDINGS

3.1. Findings Related to the First Research Problem

Table 3 presents the findings for the first research problem based on different stopping rules.

Table 3. *The average item obtained under different item selection methods for different stopping rules, mean of SE, correlation coefficient, bias and RMSE statistics.*

Stopping Rule	Item Selection Method	Mean of Item	Mean of SE	r	Bias	RMSE
SE<0.30	MFI	5.44	0.175	0.952	0.147	0.101
	MLWI	5.43	0.174	0.929	0.165	0.157
	MPWI	5.16	0.186	0.944	0.169	0.120
	MEI	5.64	0.171	0.949	0.152	0.111
	MEPV	5.36	0.174	0.951	0.148	0.106
	MEPWI	5.31	0.179	0.945	0.164	0.119
SE<0.50	MFI	4.07	0.259	0.936	0.224	0.136
	MLWI	4.00	0.264	0.936	0.228	0.136
	MPWI	4.01	0.255	0.936	0.223	0.135
	MEI	4.01	0.259	0.937	0.223	0.133
	MEPV	4.01	0.257	0.939	0.217	0.130
	MEPWI	4.01	0.255	0.936	0.223	0.135
SE<0.70	MFI	4.00	0.267	0.933	0.233	0.141
	MLWI	4.00	0.264	0.936	0.228	0.136
	MPWI	4.00	0.257	0.936	0.225	0.135
	MEI	4.00	0.260	0.937	0.225	0.134
	MEPV	4.00	0.257	0.939	0.217	0.129
	MEPWI	4.00	0.255	0.936	0.223	0.135
NI=10	MFI	10.00	0.136	0.965	0.124	0.076
	MLWI	10.00	0.137	0.962	0.130	0.082
	MPWI	10.00	0.139	0.962	0.131	0.082
	MEI	10.00	0.134	0.963	0.128	0.080
	MEPV	10.00	0.125	0.969	0.108	0.066
	MEPWI	10.00	0.139	0.962	0.131	0.082
NI=20	MFI	20.00	0.080	0.986	0.067	0.080
	MLWI	20.00	0.085	0.983	0.065	0.033
	MPWI	20.00	0.076	0.985	0.062	0.033
	MEI	20.00	0.080	0.987	0.057	0.029
	MEPV	20.00	0.077	0.988	0.054	0.027
	MEPWI	20.00	0.079	0.985	0.061	0.033

The results show that the correlation coefficient between the estimated and true occupational maturity level of individuals produced simulatively under different item selection methods and stopping rules, the average number of items applied, standard error averages, bias, and RMSE statistics provide similar outcomes. In addition, when using variable-length stopping rules (SE<0.30, SE<0.50, SE<0.70), the test is completed using an average of between 4 to 5.64 items. In this case, the number of items decreases by 84% to 90% compared to the original scale.

Under different item selection methods and stopping rules, the lowest mean standard error was determined as 0.076 (for the condition MPWI; NI=20), while the highest mean standard error was determined as 0.267 (for the condition MFI; SE<0.70). The highest correlation coefficient between the predicted occupational maturity level and the actual occupational maturity level was determined as 0.988 (for the condition MEPV; NI=20), and the lowest correlation coefficient was determined as 0.929 (for the condition MLWI; SE<0.30). The lowest bias statistic was calculated as 0.054 (for MEPV; NI=20 condition) and the highest bias statistic was calculated as 0.233 (for MFI; SE<0.70 condition). The lowest RMSE statistic was obtained as 0.027 (for MEPV; NI= 20 condition) and the highest RMSE statistic was obtained as 0.141 (for MFI; SE<0.70 condition).

Overall, it is seen that the most appropriate stopping rule is NI=20, and the most appropriate item selection method is MEPV. When the stopping rule is set as NI=20, the CAT application is expected to end with 45% fewer items than the original scale, while when the stopping rule is SE<0.30, the CAT application is expected to end with 85% fewer items than the original scale. Therefore, it is suggested that the SE<0.30 stopping rule should be used for real CAT applications, considering the low level of differences between correlation coefficients, bias, and RMSE statistics, and the significant decrease in the number of items.

When the stopping rule was determined as SE<0.30, the highest correlation coefficient (0.952), the lowest bias (0.147) and RMSE statistics (0.101) were obtained based on the MFI method. Thus, it was predicted that it would be more appropriate to determine the item selection method as MFI and the stopping rule as SE<0.30 in the actual CAT application.

3.2. Findings Related to the Second Research Problem

Table 4 presents descriptive statistics of the occupational maturity levels of individuals obtained from the CAT application and the paper-pencil test administration.

Table 4. *Descriptive statistics of occupational maturity levels obtained from CAT and paper-pencil test applications.*

	N	Minimum	Maximum	Mean	*sd*
CAT	33	-0.19	2.46	1.283	0.626
Paper-Pencil	33	-0.06	2.26	0.812	0.600

The results show that a minimum of -0.19 and a maximum of 2.46 occupational maturity level were estimated from the CAT application. The average of the occupational maturity levels obtained from the CAT application was 1.28, while the standard deviation was 0.63. On the other hand, the minimum and maximum occupational maturity levels estimated from the paper-pencil test application were -0.06 and 2.26, respectively. The average of the occupational maturity levels obtained from the CAT application was 0.81, while the standard deviation was 0.60. Furthermore, there was a moderate (r=0.535) positive and statistically significant relationship between the CAT application and the occupational maturity levels obtained from the paper-pencil test application (p<0.05).

4. DISCUSSION and CONCLUSION

The results of the study indicate that the correlation coefficient between the estimated occupational maturity level and the true occupational maturity level of individuals produced simulatively under different item selection methods and stopping rules, the average number of items applied, standard error averages, bias and RMSE statistics provide similar results. This finding is consistent with previous studies that compared different methods of item selection and stopping rules (Aybek & Demirtaşlı, 2017; Choi & Swartz, 2009; Ho, 2010; Veldkamp, 2003). However, the results of this study differ from Penfield’s (2006) study, which found that the MPWI and MEI methods had a higher level of measurement precision than the MFI method (in case the information functions of the items in the item pool were flat), while the MPWI and MEI methods were not different from each other.

In a study by van der Linden's (1998) that compared the item selection methods and stopping rules under the 2-parameter model, the MFI and MPWI methods had the highest bias statistics (NI=5 and 10), while the other three item selection methods (MEI, MEPV and MEPWI) had lower bias.

The study also found that when variable length stopping rules are used (SE<0.30; SE<0.50; SE<0.70), ability estimation can be made with 84%-90% fewer items than the original scale. This finding is in parallel with the advantage of the following ratios in the number of items: the rate of 73.3% was reached as a result of the study of Gardner et al. (2004); the rates of 36%-65% reached as a result of the study of Smits et al. (2011); the 50.86% rate reached as a result of Aybek and Demirtaşlı's (2017) study; the rate of 75% obtained in the study of Gibbons et al. (2016); the rate of 67% obtained in the study by Stochl et al. (2016); the 50%-85% rates obtained in the study by Petersen et al. (2016); the 30%-71% rates obtained in the study by Choi and McClenen (2020); the rate of 75% obtained in the study by Harrison et al. (2020); the 50%-63% rates obtained in the study by Yasuda et al. (2021); the 62%-96% rates obtained in the study by Liu et al. (2022); the rate of 78% obtained in the study by Giordano et al. (2023). In the studies conducted by Smits et al. (2011) and Aybek and Demirtaşlı (2018), it can be said that the low ratio in test lengths is because there is a more limited pool of items compared to other studies.

Furthermore, there was a moderate (r=0.535) positive and statistically significant relationship between the CAT application and the occupational maturity levels obtained from the paper-pencil test application ($p<0.05$). The correlation to be obtained from CAT and paper-pencil application is expected to be high as in simulation studies. Compared to the correlation coefficient obtained in the simulation CAT study (r=0.952; MFI, SE<0.30; see Table 3), the correlation coefficient obtained in the real CAT study (r=0.535) is lower. It can be said that there are several reasons for this. Firstly, the correlation coefficient obtained from the real CAT application is relatively lower than the correlation coefficient obtained from the simulation data, due to the sample size. Because the sample size is effective in calculating the correlation coefficient (Green, 1991; Harris, 1985; Tabachnick & Fidell, 1996; Wilson & Morgan, 2007). In addition, the application of CAT to students in the computer laboratory instead of the classroom may have caused random errors to be mixed in the measurement results. In this case, the difference in the measurement results obtained from the CAT and paper-pencil application caused the correlation coefficient to be low. In addition, in studies with dichotomous and polytomous measurement tools, it is seen that the correlation coefficient obtained from the real CAT study is lower than the correlation coefficient obtained from the simulation CAT study (Aybek & Demirtaşlı, 2018; Şahin & Gelbal, 2020).

Acknowledgments

This study was produced from the doctoral thesis titled "Investigation of different item selection methods in terms of stopping rules in polytomous computerized adaptive testing".

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**: Hacettepe University, 24.10.2017, 433-3695.

Authorship Contribution Statement

Süleyman Demir: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing-original draft. **Derya Çobanoğlu Aktan**: Methodology, Supervision, and Validation. **Neşe Güler**: Methodology, Supervision, and Validation.

Orcid

Süleyman Demir https://orcid.org/0000-0003-3136-0423
Derya Çobanoğlu Aktan https://orcid.org/0000-0002-8292-3815
Neşe Güler https://orcid.org/0000-0002-2836-3132

REFERENCES

AERA, APA, & NCME (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Akdaş, G., & Ekinci, M. (2016). Sağlık meslek lisesi öğrencilerinin mesleki olgunluk düzeylerinin ve algıladıkları aile desteğinin incelenmesi [Analysis of vocational school of health students' professional maturity and family support perception levels]. *Uluslararası Hakemli Psikiyatri ve Psikoloji Araştırmaları Dergisi 7, 83-100.* https://doi.org/10.17360/UHPPD.2016723147

Akıntuğ, Y., & Birol, C. (2011). Lise öğrencilerinin mesleki olgunluk ve karar verme stratejilerine yönelik karşılaştırmalı analiz [Comparative analysis of vocational maturity and decision making strategies of high school students]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 41,* 1-12. http://efdergi.hacettepe.edu.tr/yonetim/icerik/makaleler/675-published.pdf

Aybek, E.C., & Demirtaşlı, R.N. (2017). Computerized adaptive test (CAT) applications and item response theory models for polytomous items. *International Journal of Research in Education and Science, 3*(2), 475-487. https://doi.org/10.21890/ijres.327907

Aybek, E.C., & Çıkrıkçı, R.N. (2018). Kendini Değerlendirme Envanteri'nin bilgisayar ortamında bireye uyarlanmış test olarak uygulanabilirliği [Applicability of the Self-Assessment Inventory as a computerized adaptive test]. *Turkish Psychological Counseling and Guidance Journal, 8*(50), 117-141. https://dergipark.org.tr/tr/pub/tpdrd/issue/40299/481364

Bardhoshi G., & Erford B.T. (2017). Processes and procedures for estimating score reliability and precision. *Measurement and Evaluation in Counseling and Development, 50*(4), 256-263. https://doi.org/10.1080/07481756.2017.1388680

Cattell R.B. (1986). The psychometric properties of tests: Consistency, validity, and efficiency. In Cattell R.B., Johnson R.C. (Eds.), *Functional psychological testing* (pp. 54-78). Brunner/Mazel.

Cattell R.B., Eber H.W., & Tatsuoka M.M. (1970). *Handbook for the Sixteen Personality Factor Questionnaire (16PF)*. Institute for Personality and Ability Testing.

Choi, S.W. (2009). Firestar: Computerized adaptive testing simulation program for polytomous item response theory models. *Applied Psychological Measurement, 33(1),* 644-645. https://doi.org/10.1177/0146621608329892

Choi, S.W., & Swartz, R.J. (2009). Comparison of CAT item selection criteria for polytomous items. *Applied Psychological Measurement, 33(1),* 419-440. https://doi.org/10.1177/0146621608327801

Choi, Y., & McClenen, C. (2020). Development of adaptive formative assessment system using computerized adaptive testing and dynamic bayesian networks. *Applied Sciences, 10*(22), 81-96. https://doi.org/10.3390/app10228196

Davis, L.L. (2002). *Strategies for controlling item exposure in computerized adaptive testing with polytomously scored items* [Unpublished doctoral dissertation]. The University of Texas.

Davis, L.L., & Dodd, B.G. (2008). Strategies for controlling item exposure in computerized adaptive testing with the partial credit model. *Journal of Applied Measurement*, *9*(1), 1-17. https://pubmed.ncbi.nlm.nih.gov/18180546/

Deyo, R.A., Diehr, P., & Patrick, D.L. (1991). Reproducibility and responsiveness of health status measures: Statistics and strategies for evaluation. *Controlled Clinical Trials*, *12*(4), 142-158. https://doi.org/10.1016/S0197-2456(05)80019-4

Dodd, B.G., De Ayala, R.J., & Koch, W.R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement*, *19*(1), 5-22. https://doi.org/10.1177/014662169501900103

Gardner, W., Shear, K., Kelleher, K.J., Pajer, K.A., Mammen, O., Buysse, D., & Frank, E., (2004). Computerized adaptive measurement of depression: A simulation study. *BMC Psychiatry*, *4*(13).

Gibbons C., Bower P., Lovell K., Valderas J., & Skevington S. (2016). Electronic quality of life assessment using computer-adaptive testing. *Journal of Medical Internet Research*, 18. https://doi.org/10.2196/jmir.6053

Giordano, A., Testa, S., Bassi, M. et al. (2023). Applying multidimensional computerized adaptive testing to the MSQOL-54: a simulation study. *Health Qual Life Outcomes* 21, 61 https://doi.org/10.1186/s12955-023-02152-8

Green, S.B. (1991). How many subjects does it take to do a regression analysis?. *Multivariate Behavioral Research*, *26*, 499-510.

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

Harris, R.J. (1985). *A primer of multivariate statistics*. Academic Press.

Harrison, C., Loe, B.S., Lis, P., & Sidey-Gibbons, C. (2020). Maximizing the potential of patient-reported assessments by using the open-source concerto platform with computerized adaptive testing and machine learning. *Journal of Medical Internet Research*, *22*(10), 1–8. https://doi.org/10.2196/20950

Ho, T. (2010). *A Comparison of item selection procedures using different ability estimation methods in computerized adaptive testing based on the Generalized Partial Credit Model*, [Unpublished doctoral dissertation]. University of Texas.

Kutlu, M. (2012). Anadolu ve genel lise öğrencilerinin çeşitli değişkenlere göre mesleki olgunluk düzeylerinin incelenmesi [An analysis of vocational maturity levels of anatolian and general high school students in terms of some variables]. *İnönü Üniversitesi Eğitim Fakültesi Dergisi*, *13*(1), 23-41. https://dergipark.org.tr/tr/download/article-file/92233

Kuzgun, Y., & Bacanlı, F. (2005). Mesleki Olgunluk Ölçeği el kitabı [Professional Maturity Scale handbook]. MEB Basımevi.

Linacre, J.M. (2000). Computer-adaptive testing: A methodology whose time has come. In Chae, S., Kang, U., Jeon E., & Linacre J.M. (Eds.), *Development of computerized middle school achievement test (in Korean)*. Komesa Press.

Liu, K., Zhang, L., Tu, D., & Cai, Y. (2022). Developing an Item bank of computerized adaptive testing for eating disorders in Chinese University students. *SAGE Open*, *12*(4). https://doi.org/10.1177/21582440221141273

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Nunnally, J., & Bernstein, I.H. (1994). *Psychometric theory*. McGraw-Hill.

Orhan, A.A., & Ültanır, E. (2014). Lise öğrencilerinin mesleki olgunluk düzeyleri ile karar verme düzeyleri [Vocational maturity level and decision making strategies of high school students]. *Ufuk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 3*(5), 43-55. https://dergipark.org.tr/tr/download/article-file/1358749

Passos, V., Berger, M.P.F., & Tan, F.E. (2007). Test design optimization in CAT early stage with the Nominal Response Model. *Applied Psychological Measurement*, *31*(3), 213–232. https://doi.org/10.1177/01466216062915

Penfield, R.D. (2006). Applied Bayesian item selection approaches to adaptive tests using polytomous items. *Applied Measurement in Education, 19*, 1-20. https://doi.org/10.1207/s15324818ame1901_1

Petersen, M.A., Gamper, E.M., Costantini, A., Giesinger, J.M., Holzner, B., Johnson, C., Sztankay, M., Young, T., Groenvold, M. (2016). An emotional functioning item bank of 24 items for computerized adaptive testing (CAT) was established. *Journal of Clinical Epidemiology, 70*, 90–100. https://doi.org/10.1016/j.jclinepi.2015.09.002

Reckase, M.D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement Issues and Practice, 8*, 11-15. https://doi.org/10.1111/j.1745-3992.1989.tb00326.x

Sahranç, Ü. (2000). *Lise öğrencilerinin mesleki olgunluk düzeylerinin denetim odaklarına göre bazı değişkenler açısından incelenmesi [A Study on some variables affecting career maturity levels of high school students depending on their locus of control]*, [Unpublished master dissertation]. Hacettepe University.

Smits, N., Cuijpers, P., & van Straten, A. (2011). Applying computerized adaptive testing to the CES-D scale: A simulation study. *Psychiatry Research*, *188*(1), 147-155. https://doi.org/10.1016/j.psychres.2010.12.001

Stochl, J., Böhnke, J.R., Pickett, K.E., & Croudace, T.J. (2016). An evaluation of computerized adaptive testing for general psychological distress: Combining GHQ-12 and Affectometer-2 in an item bank for public mental health research. *BMC Medical Research Methodology, 16*(1), 1-15. https://doi.org/10.1186/s12874-016-0158-7

Super, D.E. (1957). *The psychology of careers*. Harper.

Sürücü, M. (2005). *Lise öğrencilerinin mesleki olgunluk ve algıladıkları sosyal destek düzeylerinin incelenmesi [High school students' vocational maturity and perceived social support level]*, [Unpublished master dissertation]. Gazi University.

Şahin, M.D., & Gelbal, S. (2020). Development of a multidimensional computerized adaptive test based on the bifactor model. *International Journal of Assessment Tools in Education, 7*(3), 323-342. https://doi.org/10.21449/ijate.707199

Tabachnick, B.G., & Fidell, L.S. (1996). *Using multivariate statistics*. HarperCollins.

Thissen, D., & Mislevy, R.J. (2000). Testing algorithms. In H. Wainer (Ed.). *Computerized adaptive testing*, (101-135). Lawrence Erlbaum Assc.

Thompson, N.A., & Weiss, D.J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research and Evaluation, 16*(1), 1-9. https://doi.org/10.7275/wqzt-9427

Thorndike R.L., & Hagen, E. (1961). *Measurement and evaluation in psychology and education.* John Wiley and sons.

Turgut, M.F., & Baykul, Y. (2013). *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education]*. PegemA Yayıncılık.

Ulaş, Ö., & Yıldırım, İ. (2015). Lise öğrencilerinde mesleki olgunluğun yordayıcıları [Predictors of career maturity among high school students]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 30*(2), 151-165. http://www.efdergi.hacettepe.edu.tr/yonetim/icerik/makaleler/14-published.pdf

Ürün, A.E. (2010). *Lise öğrencilerinin kendine saygı düzeyleri ile mesleki olgunlukları arasındaki ilişki [The relationship between the self-esteem level and the vocational maturity of high school students]* [Unpublished master dissertation]. Balıkesir University.

van der Linden, W.J. (1998). Bayesian item-selection criteria for adaptive testing. *Psychometrika*, *62*, 201–216. https://link.springer.com/article/10.1007/BF02294775

van der Linden, W.J., & Pashley, P.J. (2000). Item selection and ability estimation in adaptive testing. In W.J. van der Linden & C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 1–25). Kluwer.

Van Rijn, P., Eggen, T.J., Hemker, B.T., & Sanders, P.F. (2002). Evaluation of selection procedures for computerized adaptive testing with polytomous items. *Applied Psychological Measurement 26*, 393- 411. https://doi.org/10.1177/014662102237796

Veldkamp, B.P. (2003). Item selection in Polytomous CAT. In H., Yanai, A., Okada, K., Shigemasu, Y., Kano & J.J. Meulman (Eds.), *New Developments in Psychometrics* (pp. 207-214). Springer Verlag.

Wilson, C.R., & Morgan, B.L. (2007). Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology*, *3*(2), 43-50. https://doi.org/10.20982/tqmp.03.2.p043

Wise S.L., & Kingsbury G.G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica*, *21*, 135-155. https://www.uv.es/revispsi/articulos1y2.00/wise.pdf

Yasuda, J., Mae, N., Hull, M.M., & Taniguchi, M., (2021). Optimizing the length of computerized adaptive testing for the force concept inventory. *Physical Review Physics Education Research*, *17*(1), 1-15. https://doi.org/10.1103/PhysRevPhysEducRes.17.010115

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 3, 594–612

https://doi.org/10.21449/ijate.1335260

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

The mediating role of academic self-efficacy between the answer-copying tendency and the fear of negative evaluation

Müge Uluman Mert[1,*], **Emine Burcu Tunç**[1]

[1]Marmara University, Faculty of Education, Department of Educational Measurement and Evaluation

ARTICLE HISTORY

Received: July 31, 2023
Revised: Sep. 07, 2023
Accepted: Sep. 13, 2023

Keywords:
Answer-copying tendency,
Academic self-efficacy,
Fear of negative evaluation,
Structural equation modelling.

Abstract: The aim of this research study was to analyse the relationship among answer-copying tendency, academic self-efficacy, and fear of negative evaluation. To this end, we formed a structural equivalence model, and we evaluated the mediating role of academic self-efficacy between answer-copying tendency and fear of negative evaluation. A total of 562 university students participated in the study. We used the following as data collection tools: The Brief Fear of Negative Evaluation Scale, Academic Self-Efficacy Scale, and Answer-Copying Tendency Scale. For the analysis of the data, first the measurement model was tested, then the Structural Equation Model was established and estimations were made with Maximum Probability Estimation. According to the results, academic self-efficacy plays the role of a mediatory variable between fear of negative evaluation and answer-copying tendency. As can be seen from the impact of fear of negative evaluation on answer-copying tendency, there is a meaningful, positive correlation between the two variables. When we included academic self-efficacy in the model as a mediatory variable, we observed that the relationship between fear of negative evaluation and answer-copying tendency weakened and became less noteworthy. In the light of these observations, we can assert that the tendency of individuals with high academic self-efficacy to cheat in academic contexts is lower even if they have a fear of negative evaluation.

1. INTRODUCTION

All around the world, answer copying or academic dishonesty, in general, have been controversial issues for decades. Answer copying is defined as the act of using unallowed sources during an exam or in the preparation of academic assignments, having some other people answer the questions in an exam or do an assignment (Evans et al., 1993), or the attempt to answer questions in an exam by illicitly using the materials that have been prepared by those who took the same exam previously (O'Rourke et al., 2010). While answer copying is regarded as a subcategory of academic dishonesty (Kibler et al., 1988), it can often be used as a synonym of academic dishonesty as well (Carpenter et al., 2006; Harding et al., 2004). Though there is no unanimous definition of answer copying, the term in this study hereby is used to refer to a test-taker's getting the answers from another source during an in-class assessment practice (Demir, 2018).

*CONTACT: Müge Uluman MERT ✉ mugeulumann@gmail.com Marmara University, Faculty of Education, Department of Educational Measurement and Evaluation, istanbul, Türkiye

Answer copying has a technical dimension that affects reliability and validity. Answer copying negatively affects the reliability and thus the validity of a test as it increases the scope of errors in assessment (Angoff, 1974; Holland, 1996). Therefore, it is imperative that answer copying behaviour, which poses a threat to the psychometric features of a test, is well understood and be minimized to the extent possible. Previous studies (Gerdeman, 2000; Hughes & McCabe, 2006) have shown that to understand the nature of answer copying behaviour, one needs to closely observe all the relevant factors. At this point, working with variables that affect an individual's answer copying behavior will enable more reliable and valid measurement results to be obtained. However, it is considered extremely important to study the ethical dimensions that affect answer copying behavior.

As one of the ethical dimensions of the reasons for answer copying, the reasons stemming from the education system are stated. (McCabe & Trevino, 1996). It has been asserted that the fact that learners are assessed based on their exam scores rather than their performance during the learning process may lead them to display cheating behaviour (Alkan, 2008; Küçüktepe & Eminoğlu-Küçüktepe, 2014; Mert, 2012; Özden et al., 2015). In addition to that, the assumption that what one learns throughout a given course is of no use in practical life has been cited among the reasons why test-takers cheat (Mert, 2012).

Yet another reason for cheating in exams is related to the instructor of the course in question (Eminoğlu & Nartgün, 2009; Mert, 2012; Özden et al., 2015; Seven & Engin, 2008). The following factors have been listed as reasons for cheating: the teacher’s use of items at lower cognitive levels in the exams s/he prepares for assessment purposes, the teacher’s failure to administer the assessment process in an ethical manner, the tendency to use multiple-choice task type (Koç, 2018), and the lack of communication between the teacher and the student (Mert, 2012).

Except for the reasons related to the education system and the instructor, individual factors are also cited among the reasons for answer copying (Anderman & Murdock, 2007; Bacon et al., 2020; Kayiş, 2013; Lemons & Seaton, 2011; Özden et al., 2015; Polat, 2017; Seven & Engin, 2008). It has been stated that answer copying tendency of students who have attendance issues is higher than others who regularly attend classes, that answer copying tendency of those who aspire to be a faculty member is lower than others (Çeliköz, 2016; Sevgi & Memduhoğlu, 2021), and that answer copying tendency of the students with a high grade point average is lower than others (Tümkaya, 2019).

We have observed that previous research on answer copying has focused on the test-taker’s attitude, perception, and tendencies (Hughes & McCabe, 2006; McCabe & Trevino, 1997) and has dealt with concepts such as self-efficacy, academic procrastination, motivation, perfectionism, academic success, and ethical values (Polat, 2017). Studies show that there is a negative relationship between answer copying tendency and academic self-efficacy. Even if a student has studied enough for the exam, it is known that if the perception of academic self-efficacy is low, the tendency to answer copying is high (Duran, 2020; Özden, Özdemir-Özden & Biçer, 2015; Saylık et al., 2021). However, most of the studies are related to self-efficacy and answer copying tendency. In this study hereby, one of the concepts that we worked on in relation to answer copying tendency is the concept of academic self-efficacy.

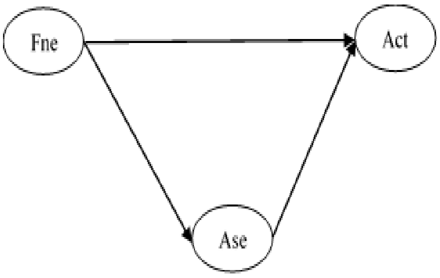
Academic self-efficacy is a prominent concept when learning activities based on self-efficacy sources are taken into consideration (Ekici, 2009; Tabancalı & Çelik, 2013). The term self-efficacy was first put forward by Bandura (1977) and was defined as the ability to fulfil an academic task successfully and one’s belief in the capability to reach a certain goal that one sets for himself or herself (Pajares, 2012; Yılmaz et al., 2007; Zimmerman 2000). An individual whose self-efficacy is high allocates more time to studying and uses this time more efficiently (Linnenbrink & Pintrich, 2003; Usher & Pajares, 2008), is more successful (Altun & Yazıcı, 2013; Bahar, 2019; Chemers et al., 2001; Choi, 2005; Robbins et al., 2004; Zajocava et al.,2005) and has a higher level of motivation (Aktaş, 2017; Eroğlu et al., 2017; Pajares & Schunk, 2001; Schunk & Pajares, 2002; Schunk & Mullen, 2012; Şeker, 2017), compared to an individual whose self-efficacy is low. When the related literature is reviewed, we can see that the number of studies that have been conducted on teacher candidates is high in number, and that the concept of self-efficacy has been studied by taking into account certain demographic variables (Bong, 2004; Ekici, 2012; Eroğlu & Yıldırım, 2018; Durdukoca, 2010; Oğuz, 2012; Polat et al., 2015). However, we can also observe that the relationship between academic efficacy and the following has been studied: various hidden variables (i.e., academic procrastination) (Albayrak, 2014; Ay et al.,2019; Nurbanu & Kumcağız, 2019; Odacı & Çelik, 2011), academic motivation (Alemdağ et al., 2014; Koca & Dadandı, 2019; Yıldız & Kardaş, 2021), self-esteem and self-compassion (Yıldırım & Demir, 2017), and anxiety about one's social appearance (Tekeli, 2017). When we consider the research studies focusing on both academic dishonesty and academic self-efficacy (Duran, 2020; Saylık et al., 2021) and those on academic dishonesty and efficacy jointly (Amelia & Usman, 2020; Büyükgöze, 2017; Karimah & Khairani, 2020; Mustika et al., 2021; Nora & Zhang, 2010; Permatasari, 2017), we can observe that a negative correlation exists between the two.

Another variable thought to have an impact on an individual's answer copying tendency is fear of negative evaluation (Bozdağ 2021; Bozdoğan & Öztürk, 2008; Kıral & Saracaloğlu; Ömür et al., 2014). Fear of negative evaluation refers to one's constant and excessive worry that he/she may be criticized harshly by others (Carleton et al., 2006; Weeks et al., 2009). These individuals, who think that people expect an outstanding performance of them feel a high level of apprehension. They have a fear of being ostracized by others because of the mistakes they may make, and owing to their fear of negative evaluation they tend to avoid engaging in activities which they do not believe they are excellent at (Frost et al., 2010). Those with a fear of negative evaluation consider themselves to be inferior to others, avoid creating an undesirable impression on them, and do not want to be alienated socially (Weeks et al., 2009). While some studies regard fear of negative evaluation as part of social anxiety (La Greca & Lopez, 1998), some others consider this fear in isolation from social anxiety (Kocovski & Endler, 2000). Although fear and anxiety are two different concepts, they are related to each other (Sylvers et al., 2011).

In previous research studies, fear of negative evaluation has been studied in relation to the following concepts or terms: the tendency towards academic dishonesty (Bozdağ, 2021; Kıral & Saracaloğlu; Ömür et al., 2014), grade orientation (Özgüngör, 2006), success rate (Alkan, 2015; Sevimli, 2009), social anxiety (Bilge & Kelecioğlu, 2008; Downing et al., 2020; Liu et al., 2020; Totan et al., 2009), depression and timidity (Bilge & Kelecioğlu, 2008), introversion (Watson, 2009), the level of boldness, (Erdoğan & Uçukoğlu, 2011) etc. It has been stated that there is a meaningful relationship between fear of negative evaluation and the tendency for academic dishonesty (Bozdağ, 2021; Kıral & Saracaloğlu; Ömür et al., 2014). We can see that there is a negative correlation between fear of negative evaluation and academic self-efficacy (Elcanlar, 2009; Han & Elçiçek, 2021).

Answer copying tendency behavior is one of the variables that threaten the psychometric properties of test scores, and it was stated that the way to understand the nature of this behavior is to consider the relevant factors. For this reason, answer copying tendency was considered as the dependent variable in this study. Based on the literature mentioned above, the present research suggests the relationship among fear of negative evaluation, academic self-efficacy, and answer-copying tendency within the framework of structural equation modeling. Therefore, the aim of this study was to investigate the dynamics behind the relationship between fear of negative evaluation and answer-copying tendency. That is, the mediator role in this relationship by academic self-efficacy was expected to be illuminated. The present study proposed that answer-copying tendency could be the result of fear of negative evaluation via the effect of academic self-efficacy. Recently, models that investigate into the motives behind answer copying have emerged (Babanejad et al., 2021; Mih &amp; Mih, 2016; Sabbagh; 2021; Yu et al., 2017; Yu et al., 2021), but no research study has been found that tests the mediatory role of academic self-efficacy (Ase) between fear of negative evaluation (Fne) and answer-copying tendency (Act). In order to reveal the relationship among these concepts, we used a structural equation modeling and examined the mediation role of academic self-efficacy between fear of negative evaluation and answer-copying tendency (Figure 1).

Figure 1. *Illustration of the mining model.*



2. METHOD

The main purpose of this research is to reveal the mediating role of academic self-efficacy in the relationship between the answer-copying tendency and the fear of negative evaluation. For this purpose, we used the relational screening model, which is designed to determine the presence and degree of change between variables thought to be related (Christensen et al., 2015).

2.1. Study Group

We carried out the study through an online data collection platform. Considering the variables used in the study, the purpose of the study, and the accessibility of the participants, we selected a total of 562 university students studying at Marmara University Atatürk Faculty of Education as participants. After obtaining ethical approval from Marmara University Social Sciences and Humanities Research and Publication Ethics Committee (Decision number: 2023-553006), during the selection process, we sought diversity at the highest possible level and took care to ensure that the subjects participated in the study of their own free will. Of the study group, 74.55% are women and 25.45% are men; 24.2% foreign languages (English - German Teaching), 30.3% psychological counselling and guidance, 29.0% Science (Science - Chemistry - Biology - Physics Teaching), 16.5% Social Studies (Social Studies - History - Geography Teaching); 16.3% 1st grade; 37.1% 2nd grade; 31.8% 3rd grade; 14.8% consists of 4th grade students.

2.2. Data Collection Tools

For the purpose of the study, Negative Evaluation Scale, Short Fear Scale, Academic Self-Efficacy Scale and Answer-Copying Tendency Scale were used.

2.2.1. *Short Fear of the Negative Evaluation Scale*

The Fear of Negative Evaluation Scale was developed by Leary (1983) to measure the fear of negative evaluation. The scale was developed in a 5-point Likert type, scored from 1 (Not at all appropriate) to 5 (Totally appropriate). There are 11 items in the scale. A minimum of 12 points and a maximum of 60 points can be obtained from the scale. Items 2, 7, and 11 in the scale are scored in reverse. The total score is obtained by adding the scores obtained from the scale items. An increase in the scores obtained from the scale indicates that the level of fear of negative evaluation increases; decrease indicates that the level of fear of negative evaluation decreases.

The validity and reliability study of the scale was carried out by Çetin et al., (2010). Construct validity and criterion-related validity methods were used to determine the validity of the Fear of Negative Evaluation Scale. As a result of the exploratory factor analysis, the KMO coefficient was calculated as .88 and the Bartlett test χ^2 value was calculated as 1095.56 (*p*<.001). 40.19% of the total variance of the scale. It has been determined that it has a one-dimensional structure that explains the Item 4 was removed from the scale due to the low correlation between the item and the total score of the 4th item in the scale. The scale was subjected to validity and reliability analysis with 11 items. As a result of confirmatory factor analysis, Fit index values were calculated as RMSEA=0.062, NFI=0.96, CFI=0.98, IFI=0.98, RFI=0.95, GFI=0.95 and AGFI=0.92. The internal consistency reliability coefficient of the scale was calculated as .84, the test-retest reliability coefficient as .82 and the test-half reliability coefficient as .83.

2.2.2. *Academic Self-Efficacy Scale*

Perceived academic self-efficacy is defined as a student's belief that he or she can successfully complete an academic task. The Turkish version of the "Academic Self-Efficacy Scale" developed by Jerusalem and Schwarzer in 1981 was made by the researchers. The original language of the scale was German and the Cronbach alpha reliability value was .87. The translation of the scale into Turkish was carried out by linguistic experts and its suitability to Turkish was evaluated by experts in terms of content and evaluation. In line with the analyses, it was revealed that the scale adapted to Turkish was one-dimensional like the original scale and consisted of seven items in total. The Cronbach alpha reliability value of the scale was determined as .79.

2.2.3. *Answer-Copy Tendency Scale in University Students*

The Answer-Copy Tendency Scale in University Students is a scale developed to reveal the potential of students to detect suspicious answer patterns. The total scores and item score distributions of the scale consisting of two factors and 20 items were normal. The item discrimination index was 0.40 or higher. α inconsistency coefficient was 0.88 or higher, while test-retest reliability coefficient was 0.80. No significant and serious differential function was detected on the substances. Goodness of fit statistics show at least acceptable model-data fit (χ^2/sd=2.79, RMSEA=0.056, SRMR=0.036, GFI=0.92, NFI=0.98, CFI=0.99). The results show that the validity and reliability levels of the scale are quite high and can be used to understand the nature of response replication.

2.3. Data Analysis

In order to determine the relationship between the concepts, a structural equation model was created and the mediating role of academic self-efficacy between fear of negative evaluation and answer-copying tendency was investigated. For all analyses Lisrel 8.51 was used.

First, descriptive statistics and correlation analyses were made, and then the pre-SEM measurement model was tested. After the measurement model, predictions were made in the structural model. SEM estimates were made using Maximum Probability Estimation. This tool was chosen because it is less likely to affect fit values from sample size and distribution (Anderson & Gerbing, 1988; Hu & Bentler, 1998).

The Fear of Negative Evaluation and Academic Self-Efficacy scales used in the research study are one-dimensional. Item parcellation is one of the important methods used to normalize the distribution of variables observed on the scales with a single factor structure and to increase the reliability of these indicators. When the literature is examined, it can be said that there are different parcellation methods (Matsunaga, 2008; Wu & Wen, 2011). Among these methods, we used the relatively frequently used parceling method in the parcellation of the Fear of

Negative Evaluation and Academic Self-Efficacy scales. We sorted items according to the parceling method by the size of the item-total correlation and created plot indicators by adding item sets to obtain equivalent indicators. Therefore, in order to increase the chances of obtaining relatively equivalent indicators, we spread the "better" and "worse" items on different parcels. We made analyses by creating two parcels of both scales. For Fear of Negative Evaluation scale the items in the first parcel of the scale are respectively; 6th, 3rd, 11th, 12th and 7th items, in the second parcel are respectively; 9th, 8th, 5th, 1st, 2nd and 10th items. For Academic Self-Efficacy scales the items in the first parcel of the scale are respectively; 4th, 6th and 5th items, in the second parcel are respectively; 3rd, 2nd, 1st and 7th items. First, the measurement model must show an acceptable fit, then the structural model must be tested (Anderson & Gerbing, 1988). We analysed the distribution of variables using skewness, the curtose value and skewness - kurtosis value divided by standard error. These obtained values are given in Table 1.

Table 1. *Descriptive statistics for sub-dimension and parcels.*

		Statistic	Std. Error	Statistic / Std. Error
FNE1PRCL	Skewness	-.011	.114	-0.09
	Kurtosis	-.411	.228	-1.80
FNE2PRCL	Skewness	.011	.114	0.09
	Kurtosis	-.275	.228	-1.20
EV	Skewness	.212	.114	1.85
	Kurtosis	-.364	.228	-1.60
NPEG	Skewness	.165	.114	1.44
	Kurtosis	-.315	.228	-1.38
ASE1PRCL	Skewness	.037	.114	0.32
	Kurtosis	-.339	.228	-1.49
ASE2PRCL	Skewness	-.166	.114	-1.45
	Kurtosis	-.106	.228	-0.46

When we examined the Table 1, all values obtained as a result of dividing the skewness and kurtosis values by the standard error range from -1.96 to 1.96, which is the critical value. In addition to these values, we used one of the normality tests, the Kolmogorov-Smirnov test ($p >$ .05). Based on these results, we can argue that all variables are normally distributed in the sample. For multicollinearity problem such as Variance inflation factor (VIF) and condition Index (CI) (Alin, 2010) were determined. In the current study VIF and CI values were lower than the critical values, 10 and 30, respectively. Findings demonstrated that there were no multicollinearity issues.

Bootstrap analysis was applied to examine the mediating role of academic self-efficacy between fear of negative evaluation and tendency to copy answers. This analysis was performed with 5000 bootstrap samples and 95% confidence intervals. The absence of a "0" value between the Bottom (BootLLCI) and Upper (BootULCI) Bootstrap values is interpreted as the effect of the factor variable. In the literature, it is stated that the bootstrap method is much stronger and gives better results than other methods such as Sobel Test (Creedon & Hayes, 2015; Hayes, 2009; Ecclesiastes & Kelley, 2011). With this method, a small rehearsal of the population is made by repeatedly burying it over the existing dataset. If the confidence interval calculated after this procedure does not contain zero, we can safely say that there is an indirect effect (Bollen & Stine, 1990; Ecclesiastes & Hayes, 2008; Shrout & Bolger, 2002).

3. RESULTS

3.1. Measurement Model Testing

This study had three latent variables and six indicators of these variables. First, we examined the descriptive statistics and correlation values of each indicator, the values of which are given in Table 2. The measurement model was tested using indicators for each of the three hidden variables.

Table 2. *Means, standard deviations and correlations of observed variables.*

Observed variables	M	*sd*	1	2	3	4	5	6
Fear of Negative Evaluation								
1 FNE1PRCL	14.22	4.89	1.00					
2 FNE2PRCL	18.19	5.07	.86**	1.00				
Academic Self-Efficacy								
3 ASE1PRCL	7.83	2.31	-.32**	-.32**	1.00			
4 ASE2PRCL	12.70	2.27	-.30**	-.29**	.57**	1.00		
Answer-Copying Tendency								
5 EV	17.79	7.89	-.03**	-.04**	-.11**	-.17**	1.00	
6 NPEG	27.95	12.72	.15**	.11**	-.18**	-.26**	.61**	1.00

Notes: N=562. $^{**}p$ <0.01

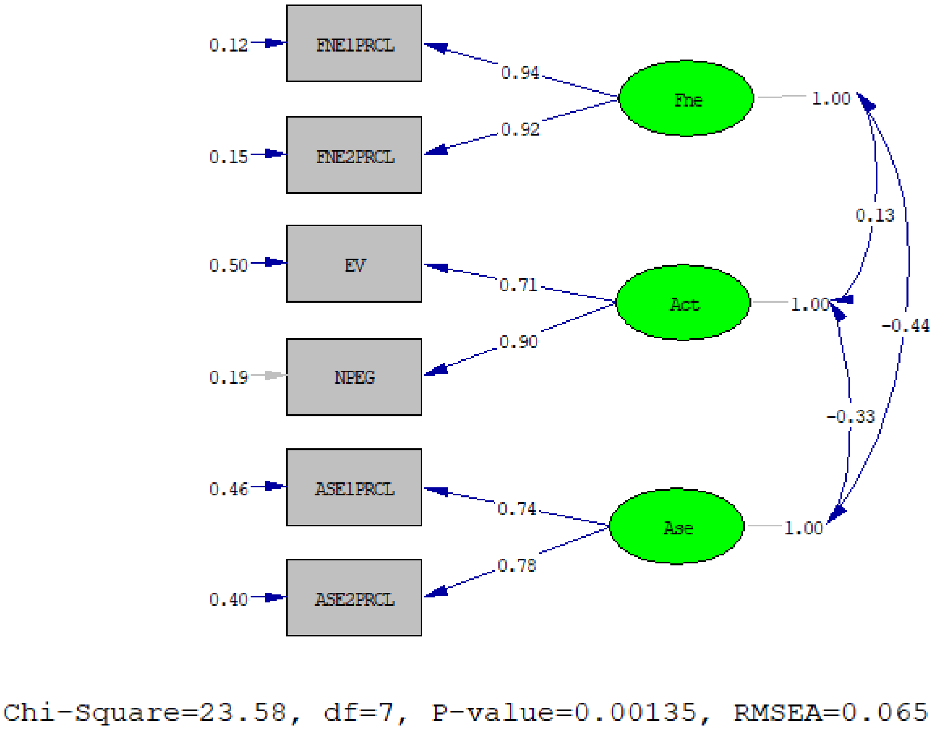
We checked for correlations between all indicator variables in the model and found them all to be statistically significant before testing the measurement model (p<.01, see Table 2). After descriptive statistics and correlation values, we tested the measurement model. The factor loads, standard errors, and *t*-values for the measurement model are shown in Table 3.

Table 3. *Factor loads, standard errors and t-values for the measurement model.*

Measure and variable	Unstandardized factor loading	*SE*	*t*	Standardized factor loading
Fear of Negative Evaluation				
1 FNE1PRCL	4.58	1.23	22.46	0.94
2 FNE2PRCL	4.68	1.29	22.12	0.92
Academic Self-Efficacy				
3 ASE1PRCL	1.71	0.31	14.69	0.74
4 ASE2PRCL	1.77	0.32	15.24	0.78
Answer-Copying Tendency				
5 EV	5.56	1.87	18.14	0.71
6 NPEG	11.44	1.87	26.74	0.90

As seen in Table 3, standardized factor loading varies between .71 and .94. The *t* values were found to be between 14.69 and 26.74 and significant. Standardized parameter estimates for the measurement model are given in Figure 2.

Figure 2. *Standardized parameter estimates for the measurement model.*



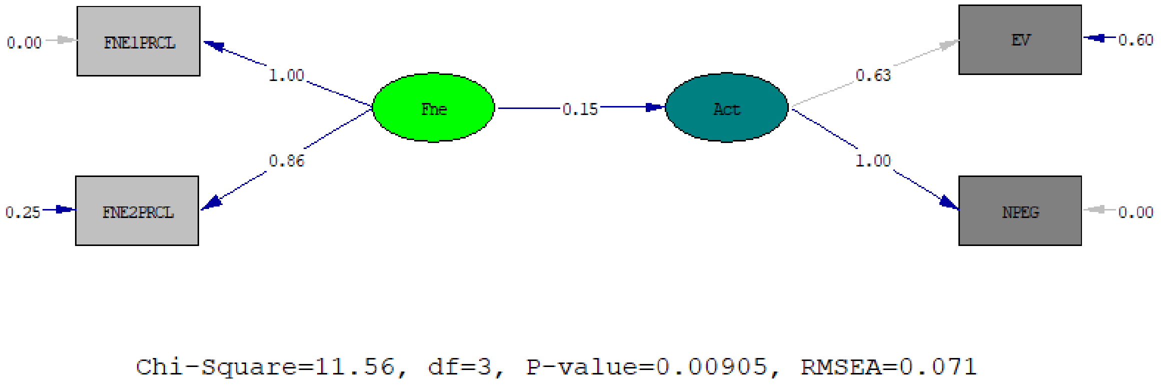
Notes: FNE1PRCL-FNE2PRCL = fear of negative evaluation; ASE1PRCL- ASE2PRCL = Academic Self-efficacy; EV (Ethical Value) – NPEG (Negative Perception of Test and Grade) = Tendency to Answer-Copy

Testing of the measurement model resulted in an acceptable fit to the data, as indicated by the goodness of the following fit statistics: χ^2(7, N=562)= 23.58; Root Mean Square Approximation Error (RMSEA)=0.065; 90 percent confidence interval for RMSEA=(0.037; 0.095); Compliance Goodness Index (GFI)=0.99; Comparative Fit Index (CFI)=0.99; Standardized Root Mean Square Meter Residue (SRMR)=0.038; Incremental Adjustment Index (IFI)=0.99; Non-normative Compliance Index (NNFI)=0.97. As shown in Table 3, all the loads of the sub-dimensions and parcels on hidden structures were statistically significant.

3.2. Testing of Structural Models

Within the scope of the research, we first tested the direct relationship between the fear of negative evaluation and the tendency to copy-answer. The results are shown in Figure 3.

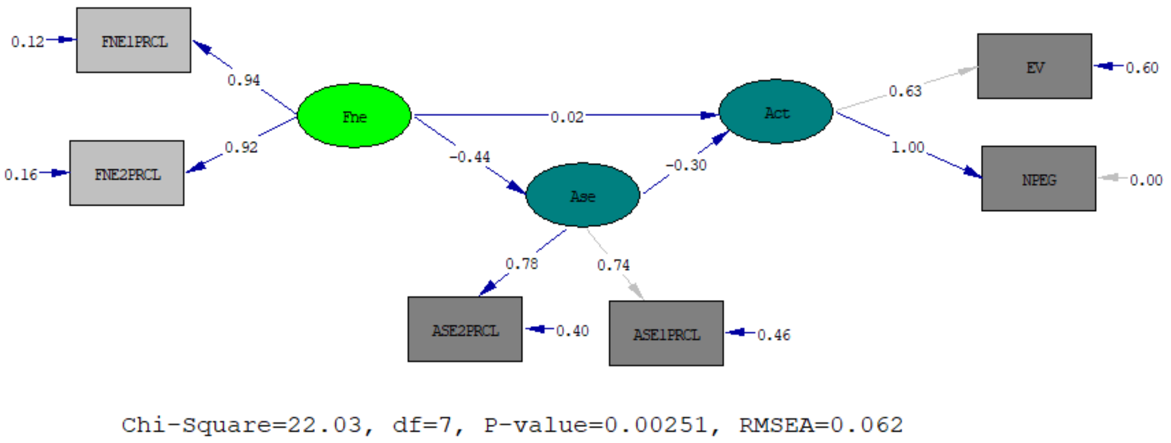
Figure 3. *Baseline model - Standardized parameter estimates for the direct relationship between fear of negative evaluation and tendency to copy answers.*



The test of the direct relationship between fear of negative evaluation and the pattern of the tendency to copy the response found an acceptable fit to the data, as indicated by the goodness of the following fit statistics: χ^2(3, N=252)=11.56; Root Mean Square Approximation Error (RMSEA)= 0.071; 90 percent confidence interval for RMSEA=(0.031; 0.12); Compliance Goodness Index (GFI)=0.99; Comparative Fit Index (CFI)=0.99; Standardized Root Mean Square Meter Residue (SRMR)= 0.048; Incremental Adjustment Index (IFI)=0.99; Non-normative Compliance Index (NNFI)=0.97.

After the direct relationship between fear of negative evaluation and tendency to copy answers, we investigated the mediating role of academic self-efficacy between fear of negative evaluation and tendency to copy answers. The results are shown in Figure 4.

Figure 4. *Standardized parameter estimates for the mediating role of academic self-efficacy between fear of negative evaluation and answer-copy tendency.*



Testing the mediating role of academic self-efficacy between fear of negative evaluation and the response copying tendency model found an acceptable fit for the data, as indicated by the goodness of the following fit statistics: χ^2(7, N=562)=22.03; Root Mean Square Proximity Error (RMSEA)=0.062; 90 percent confidence interval for RMSEA=(0.034; 0.092); Compliance Goodness Index (GFI)=0.99; Comparative Adjustment Index (CFI)=0.99; Standardized Root Mean Square Meter Residue (SRMR)=0.037; Incremental Adjustment Index (IFI)=0.99; Non-normative Compliance Index (NNFI)=0.97.

In the basic model, the path coefficient between the fear of negative evaluation and the response-copying tendency decreases from 0.15 to 0.02 in the mediation model. In the mediation model, the relationship between fear of negative evaluation and the tendency to copy-answer decreased in this way; However, when the mediation variable was added to the model, the relationship between the fear of negative evaluation and the tendency to copy the answer became meaningless. According to Baron & Kenny's (1986) method, this shows the full mediating effect of academic self-efficacy between these two variables.

3.3. Bootstrap Analysis

According to the findings of the study, the structural model showed an acceptable fit to the data. In addition, bootstrap confidence intervals were calculated for mediation. We aimed to test the importance of indirect pathways, i.e. from fear of negative evaluation (independent variable) to academic self-efficacy (mediator) and from academic self-efficacy to response-copying tendency (dependent variable) using the Bootsrap method. In the study, we plotted 5000 bootstrap samples and examined the upper and lower limits of 95% CI.

The results of the Bootstrap analysis, which was used to determine whether the mediating role between fear of negative evaluation of academic self-efficacy and the response-copying tendency was statistically significant, are given in Table 4.

Table 4. *Bootstrap analysis results on the indirect effect of academic self-reliance.*

Standardized indirect impact	Boot standard error	BootLLCI (Low value)	BootULCI (Upper value)
0.0445	0.0092	0.0273	0.0636

The standardized value for the lower value is 0.0273 and the upper value is 0.0636. Significant mediation is specified when the upper and lower limits of 95% CI do not contain zeros." 0" is not between these two values, so we can say that the mediating role of academic self-efficacy between fear of negative evaluation and tendency to copy responses is statistically significant. According to Gürbüz (2019), if the K^2 value is close to 0.01, it is interpreted as low effect, if the K^2 value is close to 0.09, it is considered as medium effect, if the K^2 value is close to 0.25, it is interpreted as high effect. When the fully standardized effect size of the mediation effect (K^2=0.0414; S.H.=0.0083; 95% CI [0.0258, 0.0588]) is considered, it is seen that this value indicates a medium effect level of mediation. And also confidence intervals of the effect size value significant because it does not cover 0 (zero).

In line with this finding, the relationship between the fear of negative evaluation and answer-copy tendency differs when the academic self-efficacy variable is included in the model. In other words, although there is a low correlation between the fear of negative evaluation and and answer-copy tendency, the relationship between these two variables is based on academic self-efficacy, since full mediation was detected.

4. DISCUSSION and CONCLUSION

Although there are different reasons for cheating, it is seen that the reasons originating from the individual are mostly studied (Bacon et al., 2020; Strap, 2013; Lemon & Seaton, 2011; Özden et al., 2015; Polat, 2017; Seven & Engin, 2008). It is important to examine the variables linked to individuals themselves, because such a study will lead to a deeper understanding of the tendency to copy responses and provide insight into ways to reduce this tendency. When the relevant literature is examined, it is seen that the copying of answers is examined in relation to concepts such as academic procrastination, self-efficacy, motivation, perfectionism, academic success, ethical values (Polat, 2017). Similarly, in this study, we examined the tendency to answer-copy along with the following variables: academic self-efficacy and fear of negative evaluation. According to the results of the research, we have determined that academic self-efficacy is a variable that clearly has a mediating role between the fear of negative evaluation and the tendency to copy answers. When we consider the direct relationship between the fear of negative evaluation and the tendency to copy answers, we observe that there is a significant positive relationship between the two variables. We observe that when academic self-efficacy enters the model as a mediating variable, the relationship between fear of negative evaluation and tendency to copy responses weakens and, therefore, the relationship becomes less meaningful. Based on this, we can say that although individuals with high academic self-efficacy have high fear of negative evaluation, they have a low tendency to copy answers.

When we reviewed the relevant literature, we found no previous research that examined the variables of response copying tendency, academic self-efficacy, and fear of negative evaluation together. Therefore, we interpreted these variables based on studies that compared two of the three variables listed.

According to the results of the research, the relationship between the fear of negative evaluation and the tendency to copy the answer was found to be significant. There are other studies in the literature that support this conclusion. Bozdoğan & Öztürk (2008) stated in their study on teacher candidates that those who had a fear of failure in some courses cheated in exams. Ömür et al. (2014) found a positive relationship, although not very strong, between the fear of negative evaluation and the tendency of teacher candidates to copy answers. When the sub-dimensions are examined, we can see that the sub-dimension with the strongest relationship with the fear of negative evaluation is the tendency to dishonesty in research and reporting. King & Saracaloğlu (2018) reaches similar conclusions in her studies with undergraduate and graduate students: There is a weak but significant relationship between the tendency to academic dishonesty and the fear of negative evaluation. Wu et al. (2019), in their study on individuals aged 17-62 years, stated that there was a negative, moderate and significant relationship between fear of negative evaluation and dishonesty. In his study on university students, Bozdağ (2021) identified a weak but positive relationship between the fear of negative evaluation and the tendency to academic dishonesty, and stated that the higher the students' fear of negative evaluation, the higher the tendency to academic dishonesty.

According to the results of this study, there is a negative, medium and significant relationship between fear of negative evaluation and academic self-efficacy. While we haven't found a study that focuses on the relationship between fear of negative evaluation and academic self-efficacy, there are a few studies that deal with fear of negative evaluation and self-efficacy. In previous studies (Elcanlar, 2009; Han & Elçiçek, 2021), it is stated that individuals with high levels of self-efficacy have a relatively lower level of fear of negative evaluation. Roomman & Özcan (2019) found that academic procrastination among students is associated with fear of negative evaluation and this relationship is mediated by academic self-efficacy. The findings suggest that improving students' academic self-efficacy may play an important role in reducing procrastination behavior. Sook-Cho & Hee-Kyung (2015) found that fear of negative evaluation has a negative impact on the academic self-efficacy and academic achievement of secondary school students. These results highlight the importance of students' academic self-efficacy and fear of negative evaluation. Additionally, the article suggests that increasing students' self-efficacy may help reduce fears of negative evaluation and increase their academic success.

The results of the study show that there is a negative and significant relationship between the tendency to copy answers and academic self-efficacy. There may be studies supporting this conclusion in the literature. Gordon & Demment (1993) examined the relationship between academic self-efficacy, coping strategies, and academic performance among college students. The study found that academic self-efficacy determines college students' coping strategies, and these strategies influence their academic performance. The results suggest that improving college students' academic self-efficacy may help improve their ability to cope with stress and ultimately improve their academic performance. Nora & Zhang (2010), in their study of students, stated that those with low levels of self-efficacy tended to copy a stronger response. Büyükgöz (2017) found a moderate and negative relationship between academic dishonesty tendency and self-efficacy levels in her study on teacher candidates. In a similar way, Akyüz *et al*. (2016) stated that there is a negative and significant relationship between a person's perception of academic self-efficacy and unethical behavior. Permatasari (2017) stated that there is a significant negative relationship between self-efficacy and cheating behavior in vocational high schools. Similar results have been obtained in recent studies. In the structural equivalence model they created, Sabzian & Mirderikvand (2020) and Sabzian & Mirderikvand (2018) stated that academic self-efficacy directly affects academic cheating behaviors. In their study of high school and college students, Amelia & Usman (2020) found that self-efficacy plays a role in response copying behavior. Karimah & Khairani (2020) found a negative, moderate, and significant relationship between self-efficacy and cheating behavior. Saylık et al. In their (2021) study, they noted that students who felt a high level of effectiveness in academic life had a weak tendency to have a positive attitude toward copying answers. Similarly, Mustika et al. (2021) revealed that there is a negative, moderate and significant relationship between self-efficacy and academic cheating.

Although some models have been developed in recent years on the causes of response copying behavior, no other studies have been conducted testing the mediating role of fear of negative evaluation between academic self-efficacy and the tendency to copy answers. On the other hand, this study has some limitations. The results of this study were obtained by using self-reporting scales. The study was limited in that it saw fear of negative evaluation as the predictive variable predicting the tendency to copy answers and academic self-efficacy as the mediator variable. In future studies, different forecasting and mediation variables can be developed and tested. In this study, university students were used as participants. The same pattern can be tested on students at different stages of training. 25% of the cohort consisted of male participants, so the same study could be carried out with more men included. This study provides information to all stakeholders in the field of education on how the level of academic self-efficacy affects the strength of the tendency to copy answers. Qualitative data can be studied in other studies as to why the academic self-efficacy variable is a full mediator. The same research can be carried out at different educational levels. It is recommended that activities to increase students' academic self-efficacy should be designed to curb the tendency to copy answers.

Acknowledgments

The authors would like to thank the blind reviewers for their useful comments and insightful suggestions.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Marmara University/Institution, 23.05.2023/05-05 - 553006.

Authorship Contribution Statement

Müge Uluman Mert: Investigation, Resources, Methodology, Visualization, Software, Formal Analysis, and Writing-original draft. **Emine Burcu Tunç**: Methodology, Supervision, and Validation.

Orcid

Müge Uluman Mert https://orcid.org/0000-0003-4155-3114
Emine Burcu Tunç https://orcid.org/0000-0002-8225-9299

REFERENCES

Akyüz, B., Kesen, M., & Oğrak, A. (2016). Örgütsel güven ve akademik özyeterlik algısının genel sinizm ve etik dışı davranışlara etkisi [The effect of organizational trust and academic self-efficacy perception on generalism and unethical behavior]. *Çankırı Karatekin Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *7*(1), 85-106.

Aktaş, H. (2017). Akademik güdülenme ile akademik öz yeterlik arasındaki ilişki: İlahiyat fakültesi öğrencileri üzerine ampirik bir araştırma [The relationship between academic motivation and academic self-efficacy: An empirical research on theology faculty students]. *Journal of the Human & Social Science Researches*, *6*(3), 1376-1398.

Albayrak, E. (2014). *Üniversite öğrencilerinde beş faktör kişilik, akademik öz-yeterlik, akademik kontrol odağı ve akademik erteleme [He big five personality, academic self-*

efficacy, academic locus of control and academic procrastination among university students] [Unpublished Master Thesis]. Teknik University.

Alemdağ, C., Erman, Ö., & Yılmaz, A.K. (2014). *Beden eğitimi öğretmeni adaylarının akademik motivasyon ve akademik öz-yeterlikleri* [Academic motivation and academic self-efficacy of physical education teacher candidates]. *Spor Bilimleri Dergisi*, *25*(1), 23-35.

Alkan, Ş. (2008). *İlköğretim ikinci kademe ile ortaöğretim öğrencilerinin ve öğretmenlerinin kopya çekmeye ilişkin görüşleri* [*Opinions of primary and secondary school students and teachers about cheating*] [Unpublished Master Thesis]. Fırat University.

Alkan, V. (2015). *Akademik ortamlarda olumsuz değerlendirilme korkusu ölçeğinin geliştirilmesi [Developing a scale for fear of negative evaluation in academic settings]* [Yüksek lisans tezi]. Ankara University.

Alin, A. (2010). Multicollinearity. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(3), 370-374.

Altun, F., & Yazıcı, H. (2013). Ergenlerin benlik algılarının yordayıcıları olarak: akademik öz-yeterlik inancı ve akademik başarı [As predictors of adolescents' self-perceptions: academic self-efficacy beliefs and academic achievement]. *Kastamonu Üniversitesi Eğitim Dergisi*, *21*(1), 145-156.

Amelia, D., & Usman, O. (2020). The Influence of Self Efficacy, Peer Conformity, Parenting Style, and Academic Procrastination on Student Cheating Behavior. Peer Conformity, Parenting Style, and Academic Procrastination on Student Cheating Behavior (January 1, 2020). http://dx.doi.org/10.2139/ssrn.3512423

Anderson, J.C., & Gerbing, D.W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*(3), 411.

Anderman, E.M., & Murdock, T.B. (Ed.). (2007). *Psychology of academic cheating*. Elsevier Academic Press

Angoff, W.H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association*, *69*, 44–49.

Ay, Ş.Ç., Arslan, F.Z., Adıgüzel, İ., & Çoban, K. (2019). Lise öğrencilerinin akademik öz-yeterlik algısı ve akademik erteleme davranışı arasındaki ilişki [The relationship between high school students' academic self-efficacy perception and academic procrastination behavior]. *Düzce Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *9*(1), 116-126.

Babanejad Nigjeh, F., Shams Esfandabad, H., & Namvar, H. (2021). Investigating the mediating role of academic motivation in the relationship between basic psychological needs, educational justice, and cheating behavior. *International Journal of Pediatrics*, *9*(9), 14446-14456.

Bacon, A.M., McDaid, C., Williams, N., & Corr, P.J. (2020). What motivates academic dishonesty in students? a reinforcement sensitivity theory explanation. *British Journal of Educational Psychology*, *90*(1), 152-166.

Baron, R.M., & Kenny, D.A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173.

Bahar, H.H. (2019). Sınıf öğretmen adaylarında akademik öz-yeterlik algısının akademik başarıyı yordama gücü [The predictive power of academic self-efficacy perception of primary school teacher candidates on academic success]. *Ilkogretim Online, 18(1), 149-157.*

Bilge, F., & Kelecioğlu, H. (2008). Olumsuz değerlendirilme korkusu ölçeği - Türkçe formunun psikometrik özellikleri [Fear of Negative Evaluation Scale - Psychometric properties of the Turkish version]. *Eurasian Journal of Educational Research*, *32(1)*, 23-30.

Bollen, K.A., & Stine, R. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociological Methodology*, *20*(1), 15-140.

Bong, M. (2004). Academic motivation in self-efficacy, task value, achievement goal orientations, and attributional beliefs. *The Journal of Educational Research*, *97*(6), 287-298.

Bozdağ, B. (2021). Examination of university students' fear of negative evaluation and academic dishonesty tendencies. *Participatory Educational Research*, *8*(3), 176-187.

Bozdoğan, A.E., &Öztürk, Ç. (2008). Öğretmen adayları neden kopya çeker [Why teacher candidates cheat]. *İlköğretim Online*, *7*(1), 141-149.

Büyükgöze, H. (2017). Öğretmen adaylarının akademik sahtekârlık eğilimlerinde öz yeterlik ve akademik kontrol odağının rolü [The role of self-efficacy and academic locus of control in teacher candidates' academic dishonesty tendencies]. *Celal Bayar Üniversitesi Sosyal Bilimler Dergisi*, *15*(01), 801-822.

Carleton, N., McCreary, D., Norton, P., & Asmundson, G. (2006). Brief fear of negative evaluation scale revised. *Depression and Anxiety*, *23*(5), 297-303.

Carpenter, D.D., Harding, T.S., Finelli, C.J., Montgomery, S.M., & Passow, H.J. (2006). Engineering students' perceptions of and attitudes towards cheating. *Journal of Engineering Education*, *95*(3), 181-194.

Chemers, M., Hu, L., & Garcia, B.F. (2001). Academic self-efficacy and first year college student performance and adjustment. *Journal of Educational Psychology*, *93*(1), 55-64.

Christensen, L.B., Johnson, R.B., & Turner, L.A. (2015). *Nitel ve karma yöntem araştırmaları*. Sever (Çev.) Research Methods Design and Analysis. A. Aypay (Çev. Ed.). Anı Publishing.

Choi, N. (2005). Self-efficacy and self-concept as predictors of college students' academic performance. *Psychology in the Schools*, *42*(2), 197-205.

Creedon, P.S., & Hayes, A.F. (2015). Small sample mediation analysis: How far can we push the bootstrap. *In Annual conference of the Association for Psychological Science*, *21*(1), 9–19.

Çeliköz, M. (2016). Öğretmen adaylarının kopya çekmeye yönelik tutumları ve kopya çekme nedenleri [Attitudes of teacher candidates towards cheating and reasons for cheating]. *Eğitim ve Öğretim Araştırmaları Dergisi*, *5*(2), 241-251.

Çetin, B., Doğan, T., & Sapmaz, F. (2010). Olumsuz değerlendirilme korkusu ölçeği kısa formu'nun Türkçe uyarlaması: Geçerlik ve güvenirlik çalışması [Turkish version of the short form of fear of negative evaluation scale: A validity and reliability study]. *Eğitim ve Bilim*, *35(156), 205-216.*

Demir, E. (2018). As a potential source of error, measuring the tendency of university students to copy the answers: a scale development study. *Eurasian Journal of Educational Research*, *18*(75), 37-58.

Downing, V.R., Cooper, K.M., Cala, J.M., Gin, L.E., & Brownell, S.E. (2020). Fear of negative evaluation and student anxiety in community college active-learning science courses. CBE-*Life Sciences Education*, *19(2), 20-28.*

Duran, A. (2020). *Kopya çekme eğilimleri ile akademik başarı, akademik özyeterlik ve akademik ertelemecilik arasındaki ilişkiler* [*Relationships between cheating tendencies and academic achievement, academic self-efficacy and academic procrastination*] [Unpublished Master Thesis]. Ankara University.

Durdukoca, Ş.F. (2010). Sınıf öğretmeni adaylarının akademik özyeterlik algılarının çeşitli değişkenler açısından incelenmesi [Examining the academic self-efficacy perceptions of primary school teacher candidates in terms of various variables]. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, *10*(1), 69-77.

Ekici, G. (2009). Biyoloji öz-yeterlik ölçeğinin Türkçeye uyarlanması [Adaptation of biology self-efficacy scale into Turkish]. *Kastamonu Eğitim Dergisi*, *17*(1), 111-124.

Ekici, G. (2012). Akademik öz-yeterlik ölçeği: Türkçeye uyarlama, geçerlik ve güvenirlik çalışması [Academic self-efficacy scale: Turkish adaptation, validity and reliability study]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 43, 174-185.

Eminoğlu, E., & Nartgün, Z. (2009). Üniversite öğrencilerinin akademik sahtekarlık eğilimlerinin ölçülmesine yönelik bir ölçek geliştirme çalışması [A scale development study to measure the academic dishonesty tendencies of university students]. *Uluslararası İnsan Bilimleri Dergisi*, *6*(1), 215-240.

Ercanlar, M. (2019). Fransızca yabancı dil öğrencilerinin olumsuz değerlendirilme korkusu ve özyeterlilik inançları [Fear of negative evaluation and self-efficacy beliefs of French foreign language students]. *Anadolu University Journal of Education Faculty*, *3*(4), 239-252.

Erdoğan, Ö., & Uçukoğlu, H. (2011). İlköğretim okulu öğrencilerinin anne-baba tutumu algıları ile atılganlık ve olumsuz değerlendirilmekten korkma düzeyleri arasındaki ilişkiler [The relationships between primary school students' perceptions of parental attitudes and their assertiveness and fear of negative evaluation.]. *Kastamonu Eğitim Dergisi*, *19*(1), 51-72.

Eroğlu, O., & Yıldırım, Y. (2018). Beden eğitimi ve spor öğretmeni adaylarının akademik öz-yeterlik düzeylerinin belirlenmesi [Determination of academic self-efficacy levels of physical education and sports teacher candidates]. *Türkiye Spor Bilimleri Dergisi*, *2*(2), 67-73.

Eroğlu, O., Yıldırım, Y., & Şahan, H. (2017). Spor bilimleri fakültesindeki öğrencilerin akademik öz-yeterlik ve akademik güdülenme düzeyleri arasındaki ilişkinin incelenmesi: Akdeniz Üniversitesi örneği [Examining the relationship between academic self-efficacy and academic motivation levels of students in the faculty of sports sciences: The case of Akdeniz University]. *Türkiye Spor Bilimleri*, *1*(1), 38-47.

Evans, E.D., Craig, D., & Mietzel, G. (1993). Adolescents' cognitions and attributions for academic cheating: a cross-national study. *Journal of Psychology*, *127*, 585- 602.

Frost, R.O., Glossner, K., & Maxner, S. (2010). *Social anxiety disorder and its relationship to perfectionism*. Hofmann, S.G. & DiBartolo, P.M. (Ed.). Social anxiety: Clinical, developmental, and social perspectives (s. 119-145). Elsevier Academic Press.

Gerdeman, R.D. (2000). *Academic dishonesty and the community college*. ERIC Digest.

Gordon, S.M., & Demment, M.L. (1993). Academic Self-Efficacy, Coping, and Academic Performance in College Students. *Journal of Educational Psychology*, *91*(1), 55-64.

Han, B., & Elçiçek, Z., (2021). Öğretmen adaylarının olumsuz değerlendirilme korkusu ile özyeterlik algıları arasındaki ilişkinin incelenmesi [Examining the relationship between teacher candidates' fear of negative evaluation and their self-efficacy perceptions]. *Dumlupınar Üniversitesi Eğitim Bilimleri Enstitüsü Dergisi*, *5*(2), 59-73.

Harding, T.S., Carpenter, D.D., Finelli, C.J., & Passow, H.J. (2004). Does academic dishonesty relate to unethical behavior in professional practice? An exploratory study. *Science and Engineering Ethics*, *10*(2), 311-324.

Hayes A.F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millenium. *Communication Monographs*, *76*(4), 408–420.

Holland, P.W. (1996). *Assessing unusual agreement between the incorrect answers of two examinees using the K-index: Statistical theory and empirical support* (Research Report RR 96-7). Educational Testing Service.

Hu, L.T., & Bentler, P.M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424.

Hughes, J.M.C., & McCabe, D.L. (2006). Understanding academic misconduct. *Canadian Journal of Higher Education*, *36*(1), 49 – 63.

Koca., F., & Dadandi, İ. (2019). Akademik öz-yeterlik ile akademik başarı arasındaki ilişkide sınav kaygısı ve akademik motivasyonun aracı rolü [The mediating role of test anxiety and academic motivation in the relationship between academic self-efficacy and academic achievement]. *İlköğretim Online, 18*(1), 241-252.

Karimah, H., & Khairani, K. (2020). The relationship of self efficacy with cheating behavior and ımplications for guidance and counseling services. *Journal Neo Konseling, 2*(4).

Kayiş, A.R. (2013). *Üniversite öğrencilerinin başarı yönelimlerinin incelenmesi* [*Examining the achievement orientations of university students*] [Unpublished Master Thesis]. Anadolu University.

Kıral, B., & Saracaloğlu, S. (2018). Akademik sahtekârlık eğilimi ile olumsuz değerlendirilme korkusu arasındaki ilişki [The relationship between academic dishonesty tendency and fear of negative evaluation]. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi, 15*(1), 323-359.

Kibler, W.L., Nuss, E.M., Paterson, B.G., & Pavela, G. (1988). *Academic integrity and student development: Legal issues and policy perspectives*. College Administration Publications.

Kocovski, N.L., & Endler, N.S. (2000). Social anxiety, self-regulation, and fear of negative evaluation. *European Journal of Personality, 14*(4), 347-358.

Koç, S. (2018). *Üniversite öğrencilerinin kopya çekmeye yönelik eğilimleri ve planlanmış davranış teorisi bağlamında kopya çekme davranışına yönelik model sınaması* [University students' cheating tendencies and model testing for cheating behavior in the context of planned behavior theory] [Unpublished Master Thesis]. Yüzüncü Yıl University.

Küçüktepe, C., & Eminoğlu Küçüktepe, S. (2014). Üniversite öğrencilerinin kopya çekme davranışlarının öğrenci görüşlerine göre incelenmesi [Examining the cheating behaviors of university students according to student opinions]. *Eğitim ve Öğretim Araştırmaları Dergisi, 3*(3), 253-270.

La Greca, A.M., & Lopez, N. (1998). Social anxiety among adolescents: Linkages with peer relations and friendships. *Journal of Abnormal Child Psychology, 26*(2), 83-94.

Lemons, M.A., & Seaton, J.L. (2011). Justice in the classroom: Does fairness determine student cheating behaviors? *Journal of Academic Administration in Higher Education, 7*(1).

Linnenbrink, E.A., Pintrich, P.R. (2003). The role of self-efficacy beliefs in student engagement and learning in the classroom. *Reading & Writing Quarterly, 19*, 119- 137.

Liu, X., Yang, Y., Wu, H., Kong, X., & Cui, L. (2020). The roles of fear of negative evaluation and social anxiety in the relationship between self-compassion and loneliness: a serial mediation model. *Current Psychology, 41*, 5249–5257.

Matsunaga, M. (2008). Item parceling in structural equation modeling: A primer. *Communication Methods and Measures, 2*(4), 260-293.

McCabe, D.L., & Trevino, L.K. (1996). What we know about cheating in college: longitudinal trends and recent developments. *Change, 28*(1), 28-33.

McCabe, D.L., & Trevino, L.K. (1997). Individual and contextual ınfluences on academic dishonesty: a multi-campus investigation. *Research in Higher Education, 38*(3), 379-396.

Mert, E.L. (2012). Temel işlevi bilim insanı yetiştirmek olan bazı bölümlerde kopya [In some departments whose main function is to train scientists, copy]. *Turkish Studies (Elektronik), 7*(3 B), 1813-1829.

Mih, C., & Mih, V. (2016). Fear of failure, disaffectıon and procrastination as mediators between controlled motivation and academic cheating. *Cognitie, Creier, Comportament/Cognition, Brain, Behavior, 20*(2), 117-132.

Mustika, M., Hasmayni, B., & Sani, Z.N. (2021). The relationship between self-efficacies to academic cheating in Madrasah Aliyah Islamiyah Sunggal. Budapest International

Research and Critics Institute (BIRCI-Journal): *Humanities and Social Sciences, 4*(2), 2800-2815.

Nurbanu, S., & Kumcağız, H. (2019). Ergenlerin akademik erteleme davranışları, akademik öz yeterlik inançları ve mükemmeliyetçilik [Adolescents' academic procrastination behaviors, academic self-efficacy beliefs, and perfectionism]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 35*(2), 375-386.

Odacı, H., & Çelik, Ç.B. (2011). Üniversite öğrencilerinin problemli internet kullanımlarının akademik öz-yeterlik, akademik erteleme ve yeme tutumları ile ilişkisi [The relationship between university students' problematic internet use and academic self-efficacy, academic procrastination and eating attitudes]. *Education sciences, 7*(1), 389-403.

Odacı, H., & Özcan, Ö. (2019). The Mediating Role of Academic Self-Efficacy in the Relationship Between Fear of Negative Evaluation and Academic Procrastination. *Journal of Rational-Emotive & Cognitive-Behavior Therapy*, 37(4), 112-128.

Oğuz., A. (2012). Sınıf öğretmeni adaylarının akademik öz yeterlik inançları [Academic self-efficacy beliefs of classroom teacher candidates]. *Anadolu Journal of Educational Sciences International, 2*(2), 15-28.

O'Rourke, J., Barnes, J., Deaton, A., Fulks, K., Ryan, K., & Rettinger, D.A. (2010). Imıtation is the sincerest form of cheating: the influence of direct knowledge and attitudes on academic dishonesty. *Ethics and Behavior, 20,* 47-64.

Ömür, Y.E., Aydın, R., & Argon, T. (2014). Olumsuz değerlendirilme korkusu ve akademik sahtekârlık [Fear of negative evaluation and academic dishonesty]. *Eğitim ve İnsani Bilimler Dergisi, 5*(9), 131-149.

Özden, M., Uçansoy Baştürk, A., & Demir, M. (2015). Kopya çektim, çünkü…: bir olgu bilim çalışması [I cheated because…: a phenomenology study]. *Turkish Online Journal of Qualitative Inquiry, 6*(4),57-89.

Özden, M., Özdemir Özden, D., & Biçer, B. (2015). Akademik usulsüzlük: sınıf öğretmeni adaylarının deneyimleri. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi, 45*, 130-143.

Özgüngör, S. (2006). Öz bilinç, olumsuz değerlendirilme korkusu, performans odaklı sınıf algısı ve not yönelimi [Self-awareness, fear of negative evaluation, performance-oriented classroom perception and grade orientation]. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi*, *19*, 1-8.

Pajares, F. (2012). *Motivational role of self-efficacy beliefs in self-regulated learning*. D.H. Schunk & B.J. Zimmerman (Eds.), Motivation and self-regulated learning. Routledge.

Pajares, F., & Schunk, D.H. (2001). Self-beliefs and school success: selfefficacy. *Self-Perception, 11*(2), 239-266.

Permatasari, D.P. (2017, Ekim). *Correlation between self–efficacy and cheating behavior on vocational high school students*. 8th International Conference on Language, Innovation, Culture, and Education, London.

Polat, M. (2017). Türkiye'de öğrenciler neden kopya çeker? Bir meta-sentez çalışması [Why do students cheat in Turkey? A meta-synthesis study]. *Eğitim Bilimleri Araştırmaları Dergisi, 7*(1), 223-242.

Polat, M., Dilekmen, M., & Yasul, A.F. (2015). Öğretmen adaylarında okula yabancılaşma ve akademik öz-yeterlik: Bir chaid analizi incelemesi [School alienation and academic self-efficacy in teacher candidates: A chaid analysis review]. *Uluslararası Eğitim Bilimleri Dergisi, 2*(4), 214-232.

Preacher, K.J., & Hayes, A.F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*(3), 879–891.

Preacher, K.J., & Kelley, K. (2011). Effect size measures for mediation models: quantitative strategies for communicating indirect effects. *Psychological Methods, 16*(2), 93.

Robbins, S.B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130*(2), 261-283.

Sabbagh, C. (2021). Self-reported academic performance and academic cheating: exploring the role of the perceived classroom (in) justice mediators. *British Journal of Educational Psychology*, *91*(4), 1517-1536.

Sabzian, S., Ghadampour, E., & Mirderikvand, F. (2018). Providing a causal model for perceptions of emotional climate and flexibility of family with academic dishonesty: the mediating role of academic self-efficacy. *Quarterly Journal of Social Work*, *7*(3), 32-43.

Sabzian, S., Ghadampour, E., & Mirderikvand, F. (2020). Presenting a causal model of academic engagement and academic ethics with academic cheating: The mediating role of academic self-efficacy. *Journal of School Psychology*, *8*(4), 131-155.

Sadeghi, M., Ghaampour, E., & Ghare Veysi, S. (2022). The effect of research self-efficacy on academic cheating in graduate students: the mediating role of academic locus of control. *Knowledge & Research in Applied Psychology*. https://doi.org/10.30486/jsrp.2020.1890771.2250

Saylık, A., Altay, E., & Gezici-Yalçın, M. (2021). Akademik alan memnuniyeti, öz-yeterlik ve kontrol odağının kopya çekmeye yönelik tutumun yordayıcıları olarak incelenmesi [Examining academic field satisfaction, self-efficacy and locus of control as predictors of cheating attitude]. *Kalem Eğitim ve İnsan Bilimleri Dergisi, 11*(1), 289-329.

Schunk, D.H., & Mullen, C.A. (2012). *Self-efficacy as an engaged learner. In Handbook of research on student engagement* (pp. 219-235). Springer.

Schunk, D.H., & Pajares, F. (2002). *The Development of Academic Self-Efficacy. In Development of achievement motivation* (pp. 15-31). Academic Press.

Shrout, P.E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychological Methods*, *7*(4), 422.

Seven, M.A., & Engin, A.O. (2008). Eğitim fakültesi öğrencilerinin kopya çekmeye duydukları ihtiyaç ve kopya çekme sebepleri [The need of education faculty students to cheat and the reasons for cheating]. *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 1*, 121-136.

Sevgi, K., & Memduhoğlu, H.B. (2021). Üniversite öğrencilerinin kopya çekmeye yönelik genel eğilimlerinin belirlenmesi [Determining the general tendencies of university students towards cheating]. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 21*(1), 194-221.

Sevimli, D. (2009). Beden eğitimi ve spor yüksekokulu özel yetenek sınavına katılan adayların olumsuz değerlendirilme korkusunun araştırılması [Investigation of the fear of negative evaluation of the candidates who participated in the physical education and sports school special talent exam]. *Türkiye Klinikleri, 1*(2), 88-94.

Sook Cho, K., & Hee-Kyung, L., (2015). A Longitudinal Study of the Relationships Among Fear of Negative Evaluation, Academic Self-Efficacy, and Academic Achievement in Middle School Students. *Social Behavior and Personality: An International Journal, 30*(4), 551-556.

Sylvers, P., Lilienfeld, S.O., & LaPrairie, J.L. (2011). Differences between trait fear and trait anxiety: Implications for psychopathology. *Clinical Psychology Review, 31*(1), 122-137.

Şeker, S.S. (2017). Müzik eğitimi bölümü öğretmen adaylarının akademik güdülenme ve akademik öz-yeterlik düzeylerinin incelenmesi [Examination of academic motivation and academic self-efficacy levels of music education department teacher candidates]. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi, 17*(3), 1465-1484.

Tabancalı, E., & Çelik, K. (2013). Öğretmen adaylarının akademik öz-yeterlikleri ile öğretmen öz-yeterlilikleri arasındaki ilişki [The relationship between teacher candidates' academic self-efficacy and teacher self-efficacy]. *Journal of Human Sciences, 10*(1), 1167-1184.

Tekeli, Ş.C. (2017). *Beden eğitimi ve spor öğretmeni adayları ile diğer öğretmen adaylarının sosyal görünüş kaygısı ve akademik öz-yeterlik düzeylerinin karşılaştırılması* [*Comparison of social appearance anxiety and academic self-efficacy levels of physical education and sports teacher candidates and other teacher candidates*] [Unpublished Master Thesis]. Bartın University.

Totan, T., Doğan T., Sapmaz, F., & Katmancıoğlu, A., (2009). *Üniversite öğrencilerinde sosyal kaygının olumsuz değerlendirilme korkusu ve iyimserlikle ilişkisi* [*The relationship of social anxiety with fear of negative evaluation and optimism in university students*], IV. Sosyal Bilimler Eğitimi Kongresi Bildiri Kitabı [IV. Social Sciences Education Congress Proceedings], İstanbul.

Tümkaya, S. (2019). Sınıf öğretmenliği öğrencilerinin kopya çekme tutumları, görüşleri ve benlik saygısının incelenmesi [Examining the cheating attitudes, opinions and self-esteem of classroom teacher students]. SDU *International Journal of Educational Studies, 6*(2), 15-34.

Usher, E.L., & Pajares, F. (2008). Self-efficacy for self-regulated learning a validation study. *Educational and Psychological Measurement, 68,* 3, 443-463.

Watson, F.S. (2009). Shyness in the context of reduced fear of negative evaluation and self-focus: a mixed methods case study. [Unpublished Dissertation]. University of South Florida.

Weeks, J.W., Rodebaugh, T.L., Heimberg, R.G., Norton, P.J., & Jakatdar, T.A. (2009). To avoid evaluation, withdraw: Fears of evaluation and depressive cognitions lead to social anxiety and submissive withdrawal. *Cognitive Therapy and Research, 33*, 375-389.

Wu, S., Liang, J., Lin, J., & Cai, W. (2019). Oneself is more important: Exploring the role of narcissism and fear of negative evaluation in the relationship between subjective social class and dishonesty. *PLoS One, 14*(6), e0218076.

Wu, Y., & Wen, Z.L. (2011). Item parceling strategies in structural equation modeling. *Advances in Psychological Science, 19*(12), 1859-1867.

Yıldırım, F.B., & Demir, A. (2017). Kendini engellemenin yordayıcıları olarak öz saygı, öz anlayış ve akademik özyeterlik [Self-esteem, self-understanding, and academic self-efficacy as predictors of self-handicapping]. *Ege Eğitim Dergisi, 18*(2), 676-701.

Yıldız, F.N.Y. & Kardaş, F. (2021). Ergenlerde akademik öz-yeterlik, içsel motivasyon, azim ve psikolojik dayanıklılığın iyi oluş ile ilişkisinin incelenmesi [Examination of the relationship between academic self-efficacy, intrinsic motivation, perseverance, and psychological resilience and well-being in adolescents]. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi, 18*(1), 1073-1099.

Yılmaz, M., Gürçay, D., & Ekici, G. (2007). Akademik öz-yeterlik ölçeğinin Türkçe'ye uyarlanması [Adaptation of the academic self-efficacy scale into Turkish]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 33*(33), 253-259.

Yu, H., Glanzer, P.L., & Johnson, B.R. (2021). Examining the relationship between student attitude and academic cheating. *Ethics & Behavior, 31*(7), 475-487.

Yu, H., Glanzer, P.L., Sriram, R., Johnson, B.R., & Moore, B. (2017). What contributes to college students' cheating? A study of individual factors. *Ethics & Behavior, 27*(5), 401-422.

Zajocava, A., Lynch, S.M., & Espenshade, T.J. (2005). Self-efficacy, stres and academic in college. *Research in Higher Education, 46*(6), 677-706.

Zimmerman, B.J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology, 25*(1), 82-91.

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 3, 613–637

https://doi.org/10.21449/ijate.1286819

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

The Contemporary Leadership Behavior of School Principals Scale according to teacher's perceptions: Validity and reliability study

İbrahim Gül[1,*], Selda Örs Özdil[1]

[1]Ondokuz Mayıs University, Faculty of Education, Department of Educational Sciences, Samsun, Türkiye

ARTICLE HISTORY

Received: Apr. 24, 2023
Revised: Aug. 22, 2023
Accepted: Sep. 12, 2023

Keywords:

School management,
School principal,
Educational leader,
Contemporary leadership behaviors,
Scale development.

Abstract: School principals are expected to possess modern leadership abilities that place an emphasis on ideas like collaboration, innovation, technological advancement, and egalitarianism. The objective of this study was to establish the psychometric features of the scale of current leadership behaviors of school principals (SCLBSP) developed in accordance with teacher perspectives. For the scale's content validity, eight experts were contacted, and for each item, the content validity ratio (CVR) and content validity index (CVI) were computed. Two different samples served as the basis for the scale development investigation. 253 teachers' worth of data were utilized in the exploratory factor analysis (EFA), whereas 215 teachers' worth of data were used in the confirmatory factor analysis (CFA) and measurement invariance studies. Cronbach's alpha, McDonald's omega, Split-Half method coefficients, composite reliability (CR), and average variance extracted (AVE) values were determined to determine the scale's reliability. The studies produced a single-factor structure with 34 items that explained 74.4% of the total variation. All SCLBSP items were found to have high levels of discrimination, and the reliability and validity of the entire scale were also found to be high.

1. INTRODUCTION

According to Werner (1993), a leader is someone who supports a group's efforts to accomplish organizational goals, and according to Northouse (2007) and Yukl (2010), leadership is the process of motivating others to work toward an organization's objectives. The modern leader is development- and learning-focused, inventive, egalitarian, and collaborative. Only modern leaders who are flexible will survive in today's world of fast change, according to Drucker (2000). When we examine leadership theories from the past and now, we can find that they always follow the same procedure: The earliest trait theories (Bass, 1990; Stogdill, 1948), behavioral techniques (Bakan et al., 2010; Hemphill & Coons, 1957; Stogdill, 1963), and situational leadership approaches (Catano & Stronge, 2007; Fairholm, 2002; Klingborg et al., 2006) have all since supplanted the original trait theories. Eventually, leadership was divided into two categories: traditional leadership and transformative leadership (Bass, 1990; Conger, 1999; Silver, 1990). The concept of servant leadership was first articulated in the 2000s (Stone & Pattarson, 2005). Researchers (Northouse, 2007; Sharma & Shilpa, 2013; Werner, 1993; Yukl, 2010) concentrated on how a group may accomplish organizational goals more

*CONTACT: İbrahim Gül ✉ igul@omu.edu.tr 🖃 Ondokuz Mayıs University, Faculty of Education, Department of Educational Sciences, Samsun, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

successfully despite the ongoing development of new leadership theories. As a result, modern leadership strategies were established. In contrast to traditional leadership styles, contemporary leadership emphasizes modern traits including cooperation, collaboration, communication, innovation, and digital technologies (Day & Antonakis, 2012; Erer & Demirel, 2018; Gronn, 2002; Northouse, 2007).

Based on the Social Exchange Theory, transformational leadership is considered among contemporary approaches (Burns, 1978). Transformational leaders are extroverted, adaptable, emotionally balanced, responsible and open to experiences (Judge & Bono, 2000). Many theories present different dimensions and values while presenting the detailed characteristics of leadership, but never provide a coherent definition of the structure itself. Contemporary leadership theory and leadership, especially in the last decade, are characterized by a number of critical themes, and the common elements in these themes are not always conceptualized in a similar way by researchers (Komuves & Dugan, 2010).

Seeing school administrators not only as administrators but also as leaders is an important factor in the development and success of schools. For this reason, today, the concept of school administrator has been replaced by educational leadership (Bennis, 2009; Bhattacharyya, 2018; Froiland, 2019). As leaders of educational institutions, school principals are expected to have contemporary leadership skills that emphasize concepts such as collaborative, innovative, technology-following, and egalitarian in a rapidly changing world in recent years to effectively manage their schools and increase student development (Department of Basic Education [DBE], 2019; Hargreaves & Fink, 2019; Leithwood et al., 2004, Liu et al., 2016; National Association of Elementary School Principals [NAESP], 2001; Pont et al., 2008). The contemporary school leader, who plays a key role in increasing student achievement and quality of education, should have different leadership approaches such as educational, instructional, strategic, visionary, transformational, charismatic, servant, social, authentic, spiritual, organizational, ethical and cultural leadership (Campbell, 2012; Fry, 2003; Hırlak & Taşlıyan 2018; Ireland & Hitt, 2005; John & Cole, 1999; Stodd, 2022; Sutherland & Gosling, 2010; Taylor et al., 2014; Wart, 2013; Wildavsky 2006). The contemporary leader works in close relationships with teachers, students, and parents. By consulting and supporting teachers, he or she helps them develop innovative strategies to improve students' learning experiences (Council of Chief State School Officers [CCSSO], 1996; Delaware Department of Education [DDE], 1998; Interstate School Leaders Licensure Consortium [ISLLC], 2008). They encourage teachers and students to understand the diverse cultural and socioeconomic backgrounds of teachers and students and to ensure that all students have equal opportunities to learn. Contemporary school leaders create opportunities to continuously develop themselves and their teachers. The basis of contemporary school leadership is student achievement and its effective maintenance (Jones & Harris, 2014; Leithwood et al., 2004; Sezer, 2018). By developing these contemporary leadership skills, school principals can help students and staff realize their full potential. In other words, there are increasing expectations for schools to be managed by principals with contemporary leadership behaviors, and therefore, the interest in the contemporary leadership behaviors of school principals is also increasing.

1.1. Professional Standards for Educational Leaders

The global economy, global jobs, and 21st-century skills that schools need to prepare students for necessitate a change in schools and education, and thus a change in school leaders (National Policy Board for Educational Administration [NPBEA], 2015). In 2015, the National Policy Board for Educational Administration (NPBEA) updated the professional standards for educational leadership to help ensure that every student is well-educated and prepared for the 21st century. As educational leaders, principals' achievement of the standards described below will strengthen the belief that every student will succeed academically and personally.

1.1.1. *Mission, vision, and core values*

The contemporary school principal attaches importance to values, that is, he/she determines the basic goals to guide decisions (Hodgkinson, 2008; Sabuncuoğlu & Tüz, 2001). They create a vision and mission for the school based on the core values of the organization (Chopra & Sehgal, 2019). Vision is the goal that set the direction for the future success of the school, which reveals what the school will do and where it will go (Lissack & Roos, 2001). Mission, on the other hand, is a general statement of the school's purpose, outlines the boundaries of the organization (Cornelissen, 2004), and includes the norms that hold the school together (Campbell & Yeung, 1991).

1.1.2. *Ethical and professional norms*

The contemporary school principal takes ethical and professional norms into account. They define ethical standards, professional responsibilities, and ethical issues related to the profession (Mantiri, 2011). Ethical behavior helps everyone to do their job with honesty and integrity (Freeman & Stewart, 2006; Menbarrow, 2021; Yukl, 2010). The ethical values of the organization support the mission and vision (Bowen, 2016), but some leaders with ethical values may cause unethical outcomes due to their incompetence (Ciulla, 2005).

1.1.3. *Cultural sensitivity*

The contemporary school principal should have the competence to explore cultural differences and promote equity (Rengi, 2014). They should understand the different cultural, linguistic, and socio-economic backgrounds of teachers and students and work to ensure that all students have equal learning opportunities. In a culturally responsive school, there is collaboration and competence (Bennett & Bennett, 2004). With cultural sensitivity, employees develop positive feelings towards each other (Chen & Starosta, 2000). In a way, it is the ability to understand and interpret others (Moran et al., 2007).

1.1.4. *Curriculum, instruction, and assessment*

The contemporary school principal is responsible for the effective implementation and supervision of the curriculum that shows what students will be taught, fulfills instructional leadership roles (McDonald et al., 2013; Murphy, 2005) and provides feedback to teachers as an instructional leader (Gülbahar, 2014). They support the development of coherent curricula, instruction, and assessment systems to improve the academic achievement and well-being of every student (NPBEA, 2015).

1.1.5. *Building student communities*

Creating a learning community at school is critical to students' learning and development (Verbiest et al., 2005). A professional learning community helps build the pedagogical content knowledge necessary for effective learning (Cheng, 2009). Research highlights the importance of student communities in increasing achievement, equity, and social inclusion in schools (Maier et al., 2017). The role of the contemporary school principal as an organizational leader is limited by the characteristics and dynamics of the system (Zaccaro & Klimoski, 2001). The contemporary school principal should use a combination of contemporary leadership practices to improve the quality of education and student achievement. On the other hand, he/she should involve teachers, parents, and other stakeholders in the decision-making processes.

1.1.6. *Professional development of school staff*

Learning and development play an important role in contemporary leadership. The contemporary school principal leads the learning-teaching process by giving importance to the professional development of teachers (Şişman, 2009). At the same time, he/she creates a qualified school environment (Hess & Kelly, 2005). They see schools as "learning organizations" (Okutan, 2003). The contemporary school principal should create opportunities for teachers and themselves to continuously improve. They should also ensure that students and staff have opportunities for learning and development.

1.1.7. *Creating unity among employees*

The contemporary school principal should be a team leader to increase cooperation among the staff and teachers working in the school. In such a situation, the principal should create an environment of participation and trust among teachers and manage relationships well (Manzoor et al., 2011; NPBEA, 2015). Thus, people who assume different roles in the school serve a common purpose (Elma, 2004; Merriam-Webster, 2023).

1.1.8. *Total participation in school*

Although education services are usually provided by the state, the school is not an institution detached from society (Adams, 1998). Community and family support are needed to provide resources to the school and to solve some problems (Kurt, 2005). The school interacts with its environment and this interaction necessitates cooperation with the community. The contemporary school principal creates a culture of cooperation and shared responsibility by involving teachers, parents, and other stakeholders in decision-making processes (Spillane et al., 2007). Thus, the educational leader can improve teacher collaboration and teaching practices and increase student achievement (Leithwood et al., 2002).

1.1.9. *School business and management*

Since the school is a bureaucratic institution, it has a business aspect (Taymaz, 2021). School management improves the quality of other services in the school. Effective execution of the school's services in this area helps to increase student achievement (Hoy & Miskel, 2012). Effectiveness in educational institutions is understood as the successful operation of administrators, teachers, and other employees in terms of awareness of organizational missions (Jacob & Shari, 2013). Effectiveness and efficiency in schools are complementary phenomena and can be increased through technological and scientific developments (Antonijević, 2018).

1.1.10. *School improvement*

Success in school improvement depends on the correct management of change (Heck & Hallinger, 2010; Penlington et al., 2008). School improvement is the use of various strategies and techniques to improve the quality of education in schools, increase students' academic achievement and improve communication between students, teachers, and school management (Leithwood et al., 2020). In the process of school improvement, it is necessary to get the opinions of different segments for a comprehensive situation analysis (Ministry of Education, 2007). This requires the contemporary school principal to be a transformational leader (Sun & Leithwood, 2012; Leithwood et al., 2004).

School administrators play a key role in improving student achievement and the quality of education. However, in today's rapidly changing world, the leadership approaches of school administrators also need to change. For this reason, contemporary leadership approaches that emphasize the importance of collaboration, innovation, and development are receiving increasing attention and research (Hargreaves & Fink, 2006). As leaders of educational institutions, school principals are expected to possess these contemporary leadership skills to effectively manage their schools and ensure student.

1.2. Measurement Tools to Determine the Leadership of School Principals in Türkiye

The extent to which school principals exhibit leadership behaviors is a critical issue that is frequently researched in national and international literature. In Türkiye, some scale development and adaptation studies have been conducted to determine the different leadership approaches of school principals. Scale development studies on various leadership approaches of school principals have been carried out by different people in different years or scales developed by others have been adapted into Turkish. Summary information about these scales developed in Türkiye or adapted into Turkish is given in Table 1 with their various characteristics.

Table 1. *Scale development and adaptation studies on leadership in Türkiye.*

Scale Developer	Yılmaz (2006)	Durnalı (2018)	Sezer (2018)	Dursun et al. (2019)	İlğan & Ekiz (2020)	Akyürek & Karabay (2022)
Scale Dimensions	-Communicative ethics -Climatic ethics -Ethics in decision making -Behavioral ethics.	-Motivation -Referral -Law -Infrastructure	-School development -Ensuring professional commitment -Administrative practices -Vision and mission -School-family	-Political Leadership -Human-Based Leadership -Charismatic Leadership -Structural Leadership	-Respect for private life -Professional management ethics -Creating a democratically based working environment -Role model behavior display	-One dimensional
Scale Name and Number of Items	Ethical Leadership Scale 44 items	School Principals Technological Leadership Scale 30 items	Educational Leadership Standards Scale 53 items	Multifaceted Leadership Orientations Scale 19 items	School Principals' Display of Ethical Leadership Behaviors Scale 51 items	Innovative School Leadership Scale 28 items
Scale Adaptation	Turan & Ebiçoğlu (2002)	Doğan-Kılıç et al., (2011)	Bellibaş et al., (2016)	Cerit et al., (2018)	Zorlu & Korkmaz (2020)	Yalçın & Atasoy (2021)
Scale Dimensions	-Excitement -Communication -Having a vision -To be trustworthy, to trust -Setting an example -Being democratic and tolerant -Being positive -Consistency	-Vision development -Creating an audience -Sharing vision -Monitoring the process -Conclusion -Teamwork	-Determining the Mission of the school -Training Program Management -Developing a Positive Learning Climate	-Self-management -Manage time -Effect -Comfort -Decision making -Commitment Communication -Empathy	-One dimensional	-Direction -Human development - Organizational development -Curriculum development
Scale Name, From Whom it was adapted, and Number of Items	Effective Leadership Scale Burwash (1997) Key to Leadership 40 items	Effective Leadership Scale in Learning Organizations Kabacoff (1998) 36 items	Principal Instructional Management Rating Scale Hallinger and Murphy (1985) 44 items	Effective Leadership Qualities Scale, Sun, Wang, and Sharma (2014) 16 items	Sustainable Leadership Scale Dalati et al (2017) 10 items	School Leadership Scale Leithwood and McCullough (2017) 22 items

When Table 1 is examined, it is seen that some of the scales that can be used to determine the leadership of school principals in Türkiye focus on the behaviors of school principals regarding a single leadership aspect, i.e., instructional, ethical, technological leadership, etc. (Akyürek &

Karabay, 2022; Bellibaş et al., 2016; Durnalı, 2018; Yılmaz, 2005). Although some scales are multidimensional, it is understood that they do not reveal the all-round and inclusive contemporary educational leadership behaviors of the school principal (Bellibaş et al., 2016; Cerit et al., 2018; Doğan-Kılıç et al., 2011; Dursun et al., 2019; Turan & Ebiçoğlu, 2002; Yalçın & Atasoy, 2021). Although the behaviors exhibited by the school principal are expressed in different ways, they complement each other, that is, they aim to increase the success of the students and have a homogeneous feature. In other words, a person's leadership is revealed by the combination of various aspects of her and her evaluation as a whole. Homogeneity and unidimensionality are synonymous concepts and can be seen as a feature that item groups have or do not have (Mcdonald, 1981). Unidimensionality is that the feature/ability to be measured shows a single structure in a measurement process. In other words, it means that the items measure a single dimension, a single feature (Hambleton et al., 1991). In this respect, it seems possible to evaluate the leadership behaviors of the school principal, which are defined by using different names, in a one-dimensional structure.

Research on leadership necessitates the need for contemporary and holistic school leadership. Therefore, there is a need for an up-to-date and useful measurement tool that addresses the extent to which school principals exhibit contemporary leadership behaviors as a whole and covers all aspects of contemporary leadership. The inclusion of such an up-to-date measurement tool in the Turkish literature is important in terms of determining the extent to which school principals in Türkiye demonstrate contemporary leadership behaviors and examining the relationships between contemporary leadership and various variables.

In light of this information, this study aims to develop a scale of contemporary leadership behaviors of school principals (SCLBSP), whose theoretical structure is conceptualized based on NPBEA's (2015) professional standards of educational leadership, and to determine its psychometric properties.

2. METHOD

2.1. Study Group

This section should indicate the study’s design, the sampling, the data collection tools, and the data analysis.

Two different study groups were chosen online from teachers working in the Samsun province during the 2022–2023 academic year in order to conduct exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) in the process of developing the scale of contemporary leadership behaviors of school principals. According to Erkuş (2012), the study group in scale development studies should be as diverse as feasible in terms of the trait being measured. It was requested that the convenience sampled data include teachers with a range of seniorities, school kinds, and levels. EFA was conducted using the data gathered from 253 teachers in the first stage, while CFA and measurement invariance analyses were conducted using the data gathered from 215 teachers in the second stage. Since one person's data was found to be a univariate outlier in the data gathered for CFA, 214 people were analyzed, and information about the research's study groups is included in Table 2.

When Table 2 is examined, it can be observed that the majority of the collected data for both EFA and CFA consists of female teachers, the number of teachers with 1-5 years of professional experience is smaller compared to other categories, the vast majority of teachers in both study groups work in middle schools, and they are graduates with an undergraduate's degree.

Table 2. *EFA and CFA study group.*

Data from the sample for EFA N_1 = 253			Data from the sample for CFA N_2 =214		
Gender	*f*	%	Gender	*f*	%
Female	167	66	Female	158	74
Male	86	34	Male	56	26
Professional Experience	*f*	%	Professional Experience	*f*	%
1 – 5 Year	17	7	1 – 5 Year	11	5
6 – 10 Year	69	27	6 – 10 Year	49	23
11 – 15 Year	72	28	11 – 15 Year	37	17
16 – 20 Year	37	15	16 – 20 Year	37	17
21+ Year	58	23	21+ Year	80	38
Institution of Duty	*f*	%	Institution of Duty	*f*	%
Preschool	8	3	Preschool	5	2
Primary School	58	23	Primary School	55	26
Middle School	123	49	Middle School	118	55
High School	45	18	High School	30	14
Other	19	7	Other	6	3
Education Status	*f*	%	Education Status	*f*	%
Undergraduate	215	85	Undergraduate	184	86
Master's Degree	38	15	Master's Degree	30	14

2.2. Scale Development Process

A literature research was done first in this scale development project, which was carried out to ascertain teachers' perceptions of school principals' levels of exhibiting current leadership behaviors. The scale's items were developed after research on leadership theories, professional standards for educational leaders updated by NPBEA in 2015, and measurement tools created for leadership (Bellibaş et al., 2016; Cerit et al., 2018; Doğan-Kılıç et al., 2011; Dursun et al., 2019; Turan & Ebiçoğlu, 2002; Yalçın & Atasoy, 2021). The prospective dimensions of leadership as well as the scale response categories were examined while studying the leadership literature. Although there are one-dimensional and multi-dimensional scales in the literature, as explained above, although the behaviors of the school principal are expressed in different ways, they complement each other and can be considered as a basic dimensional feature. Unidimensionality is defined as the presence of a dominant dimension in the presence of one or more small dimensions, and the estimations based on the dominant dimension being strong enough not to be affected by the presence of small dimensions (Stout, 1987). Thurstone (1931) put forward the idea that the most useful measurements are situations where only one thing is measured. Thurstone (1931, s.259) states that "The measurement of any object or entity describes only one characteristic of the measured object. This is a universal characteristic of measurement". This view was also supported by McNemar (1946) and Stout (1987) (as cited in Barış Pekmezci, 2022). Erkuş (2022) stated that most psychological variables are multidimensional/component in nature and it is difficult to obtain a pure one-dimensional structure due to other difficulties. However, the author emphasized that unidimensionality for the relevant feature is a goal that should be attempted to be established. The closer we get to unidimensionality, the more meaningful the total score will be and the more accurate, reliable and valid our measurements will be (Erkuş, 2022).

In line with the examinations and explanations made, an item pool was created by writing one-dimensional items to cover the professional standards of education leaders announced by NPBEA (2015). A pool of 69 items was developed considering these reviews. While creating the item pool, attention was paid to write as many items as possible in a way that would reflect the conceptual structure of the variable to be measured, but not exceed the conceptual framework, as stated by Erkuş (2012).

The 69-item draft form was sent via email in Excel format to five faculty members in the field of educational administration and three faculty members in the field of measurement and evaluation at the stage of seeking expert opinion to ensure the content validity of the scale. The experts were asked to evaluate the items in terms of suitability for the purpose, suitability in terms of language and expression, comprehensibility, suitability for the sub-dimension to be measured, and whether the items have similar meanings when the item evaluation Excel form was being created for them. The experts were asked to rate each item on a three-point scale as "appropriate," "should be improved," and "unnecessary." They were also asked to explain any reasons why an item was deemed unnecessary or should be improved, as well as to suggest any corrections that should be made. According to the experts' recommendations, the content validity ratio (CVR) for each item and the scale's content validity index (CVI) were computed using Excel and Lawshe's (1975) analysis approach. The acceptable critical value for an item to be included in the scale in this study was based on the CVR critical values from Ayre and Scally's (2014) study. According to the linked study, the CVR critical value for eight experts was 0.75 at a significance level of .05. 33 items that showed similarity-overlap with the expert opinion, were not appropriate for the structure, and had a CVR value below 0.75 were eliminated because of the analysis, and the remaining 36 items' CVI value was calculated to be 0.88.

The items that were decided to be included in the scale were examined for the last time by a faculty member who is an expert in the field of Turkish teaching in terms of item comprehensibility and compliance with Turkish grammar rules. At the end of these stages, the 36-item draft form was made ready for the pretest application. Teachers were asked to rate the extent to which the items in the scale reflect their school principals on a scale of 1-5, and the response categories of the items were formed as “1-Not at all”, “2-Reflects a little”, “3-Reflects moderately”, “4-Reflects a lot”, “5-Reflects completely”. A face-to-face pretest was conducted with 8 teachers to check whether the items were comprehensible, clear, and precise for the target group. The teachers found the trial form mostly clear and understandable. However, one participant stated that item 3 was difficult to understand and that he had to read it several times to understand it. This item was then transformed into a simpler version. After the pre-testing, the CFA was conducted by first collecting data from 253 teachers in December 2022 for the EFA and then from 215 teachers in March 2023 to test the accuracy of the construct obtained. The data were obtained through Google Forms, which provided the consent of the teachers.

2.3. Data Analysis

Firstly, EFA was conducted on the data collected from the first study group. For the suitability of the data for factor analysis, the assumptions of extreme value, missing value, normality, multicollinearity, and adequacy of sample size were reviewed. SPSS and Jamovi programs were used to test the assumptions. No missing values were found in the data set. To identify outliers, z scores of all individuals were calculated, and values ranging between -2.87 and +1.11 were obtained. No data was found to fall outside the -3 to +3 limits. The assumption of normality in each item score (univariate) was examined with skewness and kurtosis coefficients and a P-P graph. Tabachnick and Fidell (2009) state that the normality assumption is met when the kurtosis and skewness values are between -1.5 and +1.5. In the examinations, it was determined that the item scores met the normal distribution property. The collinearity problem was examined by Pearson Product Moment Correlation between the items; it was determined that there was a multicollinearity problem (r>0.90) between item 9-item 10, item 23-item 21, and item 23-item 25. These items were analyzed and it was decided to remove items 10 and 23 from the scale. To determine the multivariate outliers, the Mahalanobis distance was calculated for each subject and it was seen that the Mahalanobis value of 41 subjects exceeded the critical chi-square value at a .001 significance level. Although multivariate outliers are generally recommended to be excluded from the data set, it is also recommended to compare the results of the analyses without and with the exclusion of these values (Finch, 2012; Leys et al., 2018). For this reason, firstly, the analysis was performed without removing the multivariate outliers, and then the analysis was performed by removing the multivariate outliers. Since similar results were obtained in the analyses, the results were reported without excluding multivariate outliers. In addition, Kaiser-Meyer-Olkin (KMO) and Bartlett Sphericity Test were used for the suitability of the data for factor analysis and the suitability of the sample size. The fact that the KMO value is close to 1 and the Barlett Sphericity Test is significant indicates that the data are suitable for factor analysis. It is stated that if the multiple normality assumption is violated in the Likert scales, the Principal Axis Factors (PAF) calculation method should be preferred among the factor extraction methods. It is stated that the PAF method is a powerful enough method for factor extraction and is widely used in many cases (Costello & Osborne, 2005; Phakiti, Costa, Plonsky, & Starfield, 2018; as cited in Şencan & Fidan, 2020). Also, Grieder and Steiner (2022) listed various advantages of PAF in their articles comparing ML and PAF, which is a frequently used and recommended method. First, it has no distributional assumptions, whereas ML requires the data to follow a multivariate normal distribution (e.g., Fabrigar et al., 1999). Second, it is more robust in the case of unequal factor loadings, few indicators per factor, and small sample sizes (De Winter & Dodou, 2012; Briggs & MacCallum, 2003). Finally, it is better able to recover weak factors (Briggs & MacCallum, 2003; De Winter & Dodou, 2012). Since the multivariate normality assumption was not met in the data set, the PAF extraction technique was selected from the factor extraction methods. In deciding the number of factors of the scale, the parallel analysis method was taken as a basis, and the slope accumulation graph, eigenvalues, and explained variance ratios were taken into consideration. Since a single-factor structure was determined, no rotation technique was used.

To determine whether the one-factor structure of the scale determined as a result of EFA was confirmed or not on the data collected from 215 participants. As in EFA, assumptions were first tested to determine the suitability of the data for factor analysis. There were no missing values in the data set. The z scores of all individuals were calculated and it was determined that the z score of one individual was outside the range of -3 to +3 and that individual was excluded from the analysis. It was determined that the kurtosis and skewness values of the item scores were between -1.50 and +1.50 and the Pearson Product Moment Correlation calculated between the items was less than 0.90. Therefore, it can be stated that univariate outlier, normality, and multicollinearity assumptions are met in the data set. For multivariate outliers, Mahalanobis distance values of individual scores were examined and 18 multivariate outliers were found. As in the EFA, the analysis was first performed without removing the multivariate outliers, and then the analysis was repeated by removing the multivariate outliers. Since similar results were obtained in the analyses, the results were reported without excluding the multivariate outliers. As a result of the Henze-Zirkler multivariate normality test performed in R Shiny (Korkmaz et al., 2014), it was determined that this assumption was not met (*p*<.01). Therefore, Unweighted Least Squares (ULS), one of the estimation methods that does not require multivariate normality assumption, was used for parameter estimation of the CFA model. Following the CFA analysis, the item-total test correlations of the 34 items and the item discriminations of the 27% lower and upper groups were examined by t-test comparisons. A high item-total test score correlation indicates that the items measure a similar characteristic, that is, the internal consistency of the test is high.

After the scale structure was validated, a multiple-group confirmatory factor analysis (MGCFA) was conducted to determine whether the scale has measurement invariance in different groups. Measurement invariance of a scale in different groups means that the factor loadings, inter-factor correlations, and error variances of the items of the relevant scale are the same (Byrne, 1998; Jöreskog & Sörbom, 1993). In this study, four different models commonly used in the literature, namely configural invariance, metric invariance, scalar invariance, and strict invariance, were tested to test measurement invariance. It was determined that the distribution of individuals was not similar according to the variables of gender, level of education, and educational status. Therefore, the measurement invariance of the scale was tested in terms of the categorical variable of professional experience. The professional experience variable was analyzed by forming two groups above 15 years and below 15 years. Before analyzing the data for measurement invariance, assumptions were tested as in CFA, and ULS was used as a parameter estimation method since the multiple normality assumption was not met. For the reliability of the scale, Cronbach's Alpha and McDonald's Omega coefficients and the coefficients obtained from the Split-Half method were calculated. Since Cronbach's Alpha tends to give high values when there are many variables, it is also recommended to calculate composite reliability (CR) and average variance extracted (AVE) (Hair et al., 2010). Jamovi 2.3.21, IBM SPSS Statistic 22, and LISREL.8.51 package programs were used to analyze the data. The significance level was set as .05 in statistical analysis.

3. FINDINGS

In this section, the content validity findings, EFA, and CFA results conducted to test the construct validity, followed by reliability analyses and scale item statistics are presented respectively.

3.1. Exploratory Factor Analysis (EFA) Results

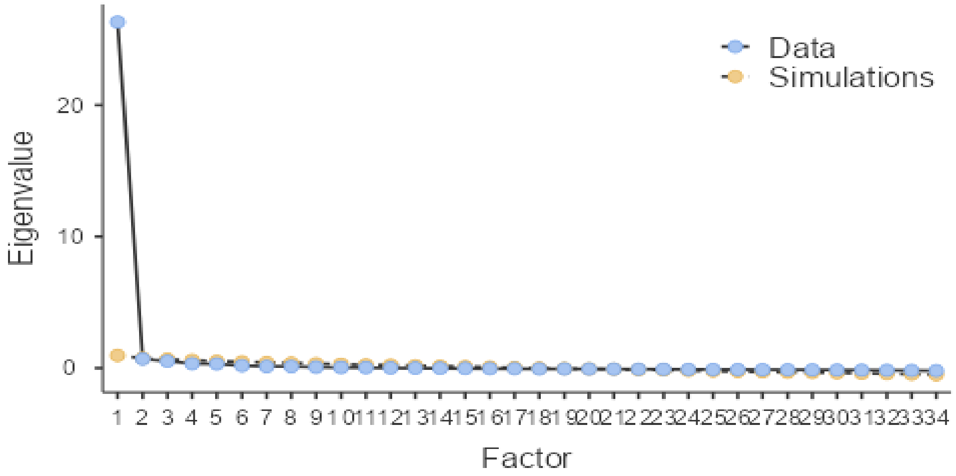
The results of the Barlett and Kaiser-Meyer-Olkin (KMO) analyses conducted to check the suitability of the data for factor analysis after it was seen that the assumptions required for conducting EFA were met are given in Table 3 below.

Table 3. *Kaiser-Meyer-Olkin (KMO) test and Bartlett's sphericity test results.*

Statistic		Value
Kaiser-Meyer-Olkin (KMO)		0.98
Bartlett's sphericity	χ^2	13535
	df	561
	p	<.001

When the suitability of the data for EFA was examined, it was determined that the KMO value was 0.98 and the Barlett Sphericity test result (χ^2= 13535, *df*=561, *p*<.001) was significant. Thus, the data were found to be suitable for factor analysis. To explore the factor structure of the scale, EFA was conducted without limiting the dimensions and it was seen that there was only one factor with an eigenvalue above 1. Oblique factor rotation was applied without any limitations and the eigenvalues were re-examined. As a result, a single factor structure was observed again. Tabachnick and Fidell (2001) stated that if the structure is very stable and consistent, the result will not change no matter which rotation method is used. The slope accumulation graph also indicates that the scale has a single factor. The slope accumulation graph obtained according to the parallel analysis method is given in Figure 1. The parallel analysis method also reveals that the scale shows a single-factor structure. The variance explained by the single factor is 74.4% of the total variance. Although Stevens (1996) suggested that the variance explained by the total scale should be 75%, there are also researchers who state that it is very difficult to meet this target in social sciences (Gorsuch, 1983; Henson & Roberts, 2006; as cited in Erkuş, 2012).

Figure 1. *Scree plot.*



After it was decided that the scale showed a single-factor structure, the factor loadings of the items were analyzed. Table 4 shows the factor loadings of the remaining 34 items in the scale after items 10 and 23 were removed from the scale due to the multicollinearity problem.

Table 4. *Factor loadings of the items.*

Item No	Item	Factor Loadings
M1	Involves the school community in the vision-mission development	0.753
M2	Open to new ideas for the development of the school.	0.832
M3	Implements its mission by transforming it into strategic goals.	0.861
M4	Updates the vision-mission according to changing needs.	0.892
M5	Acts under ethical principles concerning the school community.	0.848
M6	Communicates effectively with school stakeholders.	0.890
M7	Encourages ethical behavior among school stakeholders.	0.853
M8	Considering the benefit of the students in every practice in the school.	0.839
M9	Encourages fair treatment of all students.	0.872
M11	Strives to change prejudices in the school community.	0.893
M12	Supports inclusive education practices.	0.907
M13	Supports the use of technology in education.	0.890
M14	Encourages the effective implementation of curricula.	0.908
M15	Provides feedback to teachers on teaching practices.	0.918
M16	Encourages increased academic achievement.	0.922
M17	Takes measures to create a safe school environment.	0.854
M18	Supports the effective implementation of extracurricular activities.	0.864
M19	Encourages students to participate in in-school group activities.	0.873
M20	Supports students' relations with non-governmental organizations.	0.893
M21	Supports the professional development of teachers.	0.922
M22	Takes care to protect the work-life balance of teachers.	0.891
M24	Plans in-service training for teachers' professional development.	0.896
M25	Creates a culture of professional cooperation among teachers.	0.923

M26	Treats families and other visitors to the school in a hospitable.	0.819
M27	Maintains open two-way communication with families to increase	0.909
M28	Supports the use of school resources for the benefit of the	0.876
M29	Organizes parent education programs for the school environment.	0.848
M30	Cooperates with various organizations for the development of the	0.890
M31	Takes into account everyone's area of expertise in the distribution of	0.890
M32	Takes necessary measures to ensure that teaching is not interrupted.	0.889
M33	Utilizes technology to increase efficiency and quality.	0.925
M34	Follows good practices in other schools.	0.884
M35	Manages conflicts in the school effectively.	0.875
M36	Ensures effective use of school resources.	0.892

Table 4 shows that the factor loadings of the items vary between .753 - .925. Comrey and Lee (1992) state that factor loadings of .71 and above are excellent. The factor loadings of the items in Table 4 were examined and no item was removed from the scale. A high factor loading means that the item shows a high level of relationship with its factor. Therefore, high factor loadings are desirable.

3.2. Confirmatory Factor Analysis (CFA) Results

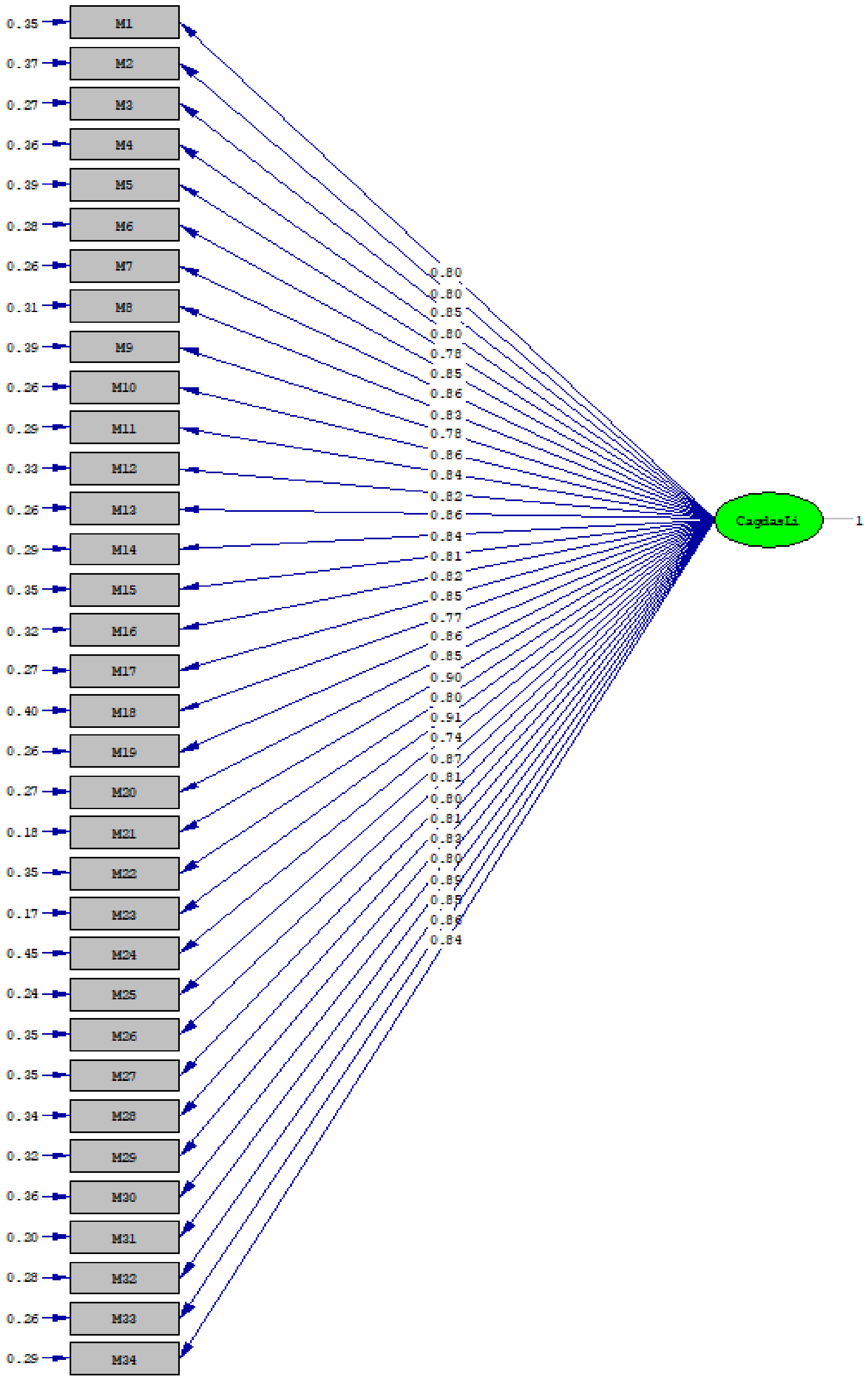
The findings of the CFA conducted to confirm the structure of the single-factor scale that emerged as a result of EFA are presented in Table 5 and Figure 2 below.

Table 5. *Standardized factor loadings and SH of the items.*

Item No	Standardized Factor Loadings	SH	Item No	Standardized Factor Loadings	SH
M1	0.80	0.35	M18	0.77	0.40
M2	0.80	0.37	M19	0.86	0.26
M3	0.85	0.27	M20	0.85	0.27
M4	0.80	0.36	M21	0.90	0.18
M5	0.78	0.39	M22	0.80	0.35
M6	0.85	0.28	M23	0.91	0.17
M7	0.86	0.26	M24	0.74	0.45
M8	0.83	0.31	M25	0.87	0.24
M9	0.78	0.39	M26	0.81	0.35
M10	0.86	0.26	M27	0.80	0.35
M11	0.84	0.29	M28	0.81	0.34
M12	0.82	0.33	M29	0.83	0.32
M13	0.86	0.26	M30	0.80	0.36
M14	0.84	0.29	M31	0.89	0.20
M15	0.81	0.35	M32	0.85	0.28
M16	0.82	0.32	M33	0.86	0.26
M17	0.85	0.27	M34	0.84	0.29

Table 5 and Figure 2 show the standardized factor loadings of the items on the relevant factor and the error variances of the items. As a result of the analysis, the significance of the factor loading values of the items should be checked first. It was determined that the *t* values of all items were greater than 2.56, that is, they were significant at a .01 significance level. It is seen that the standardized factor loading values of all items are between 0.77 and 0.91 and the error variances are considerably smaller than 0.90.

Figure 2. *Factor loadings of the items revealed by CFA results.*



After examining the coefficients obtained as a result of CFA, the goodness-of-fit indices produced to evaluate the model as a whole were examined. Goodness-of-fit index values for model-data fit are given in Table 6.

Table 6. *Goodness of fit ındex values for the model.*

χ^2	*sd*	χ^2/sd	AGFI	GFI	CFI	NFI	NNFI	PNFI	PGFI
1939.20	527	3.67	1.00	1.00	1.00	0.99	1.00	0.94	0.88
			RMSEA	SRMR	RMR				
			0.11	0.039	0.044				

When Table 6 is examined, a value between 3 and 5 obtained by dividing the χ^2 value by the degrees of freedom indicates a moderate fit (Kline, 2011). In confirmatory factor analysis, it is recommended that the evaluation of the model should be based on more than one fit index. When the fit indices related to the scale are examined, AGFI, GFI, CFI, NFI, and NNFI values above 0.95 are indicators of excellent fit. RMSEA and SRMR values between 0.05 and 0.08 indicate good fit, and values between 0.80 and 0.10 indicate acceptable fit. It is seen that the RMSEA value obtained is close to 0.10 acceptable fit and the SRMR value is below 0.05. When all the analysis results and goodness of fit values obtained with CFA are evaluated together, it can be said that the single-factor structure of the scale consisting of 34 items generally fit the data well and the scale structure is confirmed.

When the goodness of fit indices in Table 6 are examined, it is noteworthy that the one-factor structure overfitting with the data set. These results may not be replicated in different samples from the same population (in other studies). In scale development studies, researchers expect not only the structure that is suitable for their own data set, but also the resulting structure to be similar in different samples from the same universe (Osborne &Fitzpatrick, 2012). Because when researchers choose an improved scale, they will need to obtain a similar structure in the sample they will work with. Osborne and Fitzpatrick (2012) emphasized that the reproducibility studies of EFA will provide important information for researchers who will use the scale. In order to examine the reproducibility in this study, EFA was also performed on the second data set of 214 people collected for CFA, and the results were compared with the first EFA results obtained from a sample of 253 people. The results obtained are given in the table in Appendix 1. It is expected that the difference between the factor loading values of each item obtained from the two applications will be small. If the absolute value of the difference between the two factor loads is greater than 0.20, it can be said that the item is not stable, and if it is around 0.10, it can be said to be acceptable (Osborne & Fitzpatrick, 2012). In the table in Appendix 1, it is seen that the difference between the item factor loading values obtained from the two applications is below 0.11 for all items. Accordingly, it can be stated that the items are stable, and that a similar structure can occur in different samples from the same universe.

3.3. Item Analysis and Validity Analysis Based on Group Differences

To determine the discrimination levels of the items in the scale, the total scores obtained from the scale were determined and 27% lower-upper group (Nalt:59 and Nüst:58) comparisons were made. Pearson Product Moment Correlation Coefficient was used to calculate the corrected item-total test correlation, and an unrelated sample t-test was used for 27% lower-upper group comparisons. The findings obtained as a result of the analysis are given in Table 7.

Table 7. *Item analysis results.*

Item No	Corrected Item-Total Correlation	Upper and lower 27% *t* value	Item No	Corrected Item-Total Correlation	Upper and lower 27% *t* value
M1	0.79	-17.20	M18	0.76	-16.99
M2	0.78	-18.53	M19	0.85	-17.94
M3	0.84	-18.47	M20	0.83	-16.93
M4	0.79	-15.89	M21	0.90	-20.20
M5	0.77	-15.10	M22	0.79	-17.41
M6	0.83	-19.48	M23	0.90	-21.56
M7	0.85	-17.63	M24	0.74	-12.86
M8	0.82	-15.30	M25	0.87	-19.21
M9	0.77	-13.93	M26	0.81	-16.32
M10	0.85	-18.46	M27	0.79	-15.99
M11	0.83	-17.61	M28	0.80	-16.08
M12	0.82	-15.33	M29	0.82	-17.29
M13	0.86	-16.57	M30	0.79	-14.79
M14	0.84	-16.84	M31	0.88	-20.87
M15	0.80	-15.88	M32	0.84	-19.38
M16	0.81	-16.44	M33	0.85	-21.14
M17	0.85	-15.99	M34	0.83	-16.65

According to Table 7, the corrected item-total test correlation values ranged between 0.74 and 0.90. When the difference between the item-total mean scores of the lower and upper groups of 27% was examined, it was determined that the difference between the lower and upper groups was significant at the 0.01 level for all items. Accordingly, all of the items in the scale significantly distinguish between individuals who have the measured trait and individuals who do not.

3.4. Measurement Invariance

Before testing the models related to measurement invariance, it is necessary to examine the fit of the model with the data in each group separately. For the purpose of the study, firstly, fit indices were obtained separately in two different groups determined according to the professional experience variable (Brown, 2006). The findings obtained as a result of the analyzes are presented in Table 8.

Table 8. *Measurement invariance fit indexes.*

Models	χ^2	df	χ^2/df	SRMR	NNFI	CFI	RMSEA	$\Delta\chi^2$	$\Delta\chi^2/\Delta df$	ΔCFI	ΔRMSEA
15 years and less	1313.40	527	2.49	0.046	1.00	1.00	0.12				
More than 15 years	1350.72	527	2.56	0.046	1.00	1.00	0.11				
Configural invariance	2790.00	1122	2.49	0.10	0.99	0.99	0.12	-	-	-	-
Metric invariance	2761.01	1088	2.54	0.047	1.00	1.00	0.12	28.99	0.85	-0.01	0
Scalar invariance	2664.12	1054	2.53	0.046	1.00	1.00	0.12	96.89	2.85	0	0
Strict invariance	3974.63	1088	3.65	0.11	0.99	0.99	0.16	-1.310	38.53	0.01	-0.04

When the fit indices of the groups with less than 15 years and more than 15 years of experience are examined in Table 8, it can be stated that the model was confirmed separately in both groups when the fit indices obtained from both groups are evaluated together. When the findings regarding the structural invariance of the measurement model of the scale developed to determine the contemporary leadership levels of school principals are examined, it is seen that the χ^2/df value is below 3, the NNFI and CFI values are very close to 1, and the RMSEA value is outside the acceptable limits (acceptable value 0.05<RMSEA≤0.08). When all values are taken together, it shows that the model meets structural invariance. Since the factor loadings, inter-factor correlations, and error variances parameters related to the model are released in subgroups in structural invariance, it can be said that the structure of the measurement model is similar in subgroups. After determining that structural invariance was achieved, the metric invariance model was tested. In metric invariance, factor loadings are restricted; if the values resulting from this restriction do not show a worse fit than the first model, it is concluded that metric invariance is achieved. Otherwise, it is concluded that metric invariance cannot be achieved and the analysis cannot proceed to the next stage. To test metric invariance, the difference between CFI and RMSEA values obtained in the structural invariance and metric invariance stages was examined. Since the χ^2 value is affected by the sample size, the results are interpreted by considering ΔCFI and ΔRMSEA values. When the ΔCFI and ΔRMSEA values between the two models are in the range of +0.01 and -0.01, it is interpreted that the restriction does not cause a significant change in the model and that measurement invariance is achieved at the relevant stage (Cheung & Resvold, 2002; Wu et al., 2007). For metric invariance, ΔCFI and ΔRMSEA values were found to be within acceptable limits (ΔCFI ≤0.01; ΔRMSEA ≤0.01). In other words, it can be said that the factor loadings of the groups are similar. Since the metric invariance stage was achieved, the next scale invariance stage was started. At the scale invariance stage, the fit indices were within acceptable limits, and scale invariance was achieved (ΔCFI ≤ 0.01; ΔRMSEA ≤ 0.01). It was confirmed that the constants in the regression equations for the items were invariant in their subgroups. Based on this finding, it can be said that there is no bias based on items. After the scale invariance stage was achieved, the strict invariance stage was started. It can be stated that the values obtained at the strict invariance stage were out of the acceptance limits and therefore, strict invariance was not achieved (ΔCFI ≤ 0.01; ΔRMSEA<0.01).

3.5. Reliability Analysis Results

The Cronbach's Alpha and McDonald's Omega coefficients calculated for the reliability of the contemporary leadership scale were both 0.987. The internal consistency reliability of the single-factor 34-item scale was also calculated with the Split-Half method. The Cronbach's Alpha coefficient of 17 items in the first half was 0.974 and the Cronbach's Alpha coefficient of 17 items in the second half was 0.975. It can be said that the internal consistency coefficient values of the two groups formed by the Split-Half method are close to each other and very good. With this method, Guttman and Spearman-Brown coefficients were found to be 0.977. In addition to these values, the CR value of the one-factor scale was calculated as 0.98, and the AVE value as 0.69. The fact that the CR value is greater than 0.70 and the AVE value is greater than 0.50 indicates that the scale as a whole has a high level of reliability in terms of internal consistency and that convergent validity is provided (Hair et al., 2010).

4. DISCUSSION and CONCLUSION

The type and caliber of the work performed by principals or other educational leaders are outlined in the Professional Standards for Educational Leaders, which were updated by the NPBEA in 2015. To assist guarantee that every student is well-educated and ready for the 21st century, these standards lay forth the fundamentals of leadership (NPBEA, 2015). As a result of global advancements, organizations must be managed more effectively (Bhattacharyya, 2018; Froiland, 2019), and the manager concept is giving way to the leader concept. Studies have revealed that, despite the perception that school administrators are less concerned with students' learning and development, this is not the case (Gülbahar, 2014; Leithwood et al., 2022; Murphy, 2005; NAESP, 2001). According to the updated standards, it was deemed crucial to develop an inclusive scale with high validity and reliability to assess the extent to which school principals exhibit contemporary leadership behaviors based on teachers' perceptions (Blase & Blase, 2000; Liu et al., 2016; Taylor et al., 2014; Zaccaro & Klimonski, 2001).

First, an item pool was developed, the items were subjected to expert review, and a preliminary test of the 36-item draft form was carried out during the scale development phase. The scale items were subjected to exploratory factor analysis in the second step. The analysis produced a single-factor structure comprising 34 items and the exclusion of 2 items from the analysis. 74.4 percent of the total assumption is explained by the 34-item single-factor structure. The high overall score on the scale reveals that teachers have positive impressions of how well school principals exhibit modern leadership qualities. In the third stage, the unidimensional 34-item scale was reapplied to a different group for confirmatory factor analysis, and good fit values were estimated as a result of the analyses, and thus construct validity was ensured.

In addition to all these, it is not enough to state that the validity of the scale is high only by statistical analysis. Items should be related to the factors on which they are loaded with meaning and concept. When the items that make up the factor are examined, it should be understood that they measure the semantically similar feature. The information obtained as a result of factor analysis during the scale development process can provide a clue about the measured construct. The important thing is to understand what this information and values mean conceptually. Erkuş (2012) stated that when eigenvalues, explained variance, factor loads, item-total scale correlation, and internal consistency coefficient were examined in component type structures, the structure was predominantly single factored.

Following the EFA and CFA, the discrimination levels of the items were examined with the 27% sub-super group method and item-total test correlation, and the discrimination levels of all items were found to be high. Accordingly, all of the items in the scale significantly discriminate principals who have the measured trait from principals who do not.

For the reliability of the scale, Cronbach's alpha, McDonald's Omega coefficients, Guttman and Spearman-Brown coefficients were calculated by the Split-Half method, and the coefficients were found to be high. In addition to these values, the CR value and AVE value of the one-factor scale were calculated. The high coefficients obtained indicate that the scale has reliability in terms of internal consistency and convergent validity is provided.

Finally, to indicate whether the scale measures the same construct between the groups, CGFA was conducted according to the professional experience variable. Considering the changes in CFI and RMSEA, the scale met the structural, metric, and scalar invariance conditions for the professional experience variable. The fact that the first three invariance conditions were met shows that the scale can measure the same construct between groups that differ in terms of this variable. In this sense, it can be said that the scale can be used to compare teachers' perceptions of contemporary leadership behaviors of school principals among different groups.

The scores obtained from the devised scale were found to be valid and reliable in identifying the current leadership levels of school principals based on teachers' perspectives. This scale is believed to give researchers interested in school principal leadership, school growth, and school administration a thorough view on the modern leadership of school principals. The SCLBSP can be used to perform research on the links between modern leadership and other variables. The scale can be used to identify school principals who exhibit poor current leadership, and applications can be made to improve their contemporary leadership. However, as the proposed measurement instrument bases its conclusions on teachers' perceptions, there may be subjectivity in the outcomes.

One of the disadvantages of this study is that it only collected information from teachers in one city. The validity and reliability studies can be repeated by doing the study with instructors in other cities in different areas of Türkiye to boost generalizability and external validity. Additionally, the scale's criterion-referenced validity was not examined in this study. In a subsequent investigation, criterion-referenced validity evidence may also be attained.

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:** Ondokuz Mayıs University, 2022-1154.

Authorship Contribution Statement

İbrahim Gül: Theoretical foundations of the research, related literature, discussion and conclusion **Selda Örs-Özdil**: Statistical analyses, literature, findings and interpretation

Orcid

İbrahim Gül https://orcid.org/0000-0002-0501-8221
Selda Örs-Özdil https://orcid.org/0000-0002-7134-5896

REFERENCES

Adams, D. (1998). Eğitimde kalitenin tanımlanması [Defining quality in education] (çev. N. Cemaloğlu). *Kuram ve Uygulamada Eğitim Yönetimi*, *14*(14), 233-248. https://dergipark.org.tr/tr/pub/kuey/issue/10382/127041

Akyürek, M.İ., & Karabay, E. (2022). Developing innovative school leadership scale and teachers' views on innovative school leadership. *Journal of Educational Leadership and Policy Studies*, *6*(1), 1-20.

Antonijević, R. (2018). Efficiency and effectiveness of education as pedagogical and economic categories: problems of evaluation and measuring. *Open Journal for Research in Economics*, *1*(2), 37–44. https://doi.org/10.32591/coas.ojre.0102.02037a

Ayre, C., & Scally, A.J. (2014). Critical values for lawshe’s content validity ratio: Revisiting the original methods of calculation. *Measurement and Evaluation in Counseling and Development*, *47*(1), 79-86. https://doi.org/10.1177/0748175613513808

Bakan, İ., Büyükbeşe, T., Erşahan, B., & Kefe, İ. (2013). Kadın çalışanların yöneticilere ilişkin algıları: bir alan çalışması [Female employees' perceptions of managers: a field study]. *Çankırı Karatekin Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, *3*(2), 71-84. https://dergipark.org.tr/tr/pub/ckuiibfd/issue/ 32901/365488

Barış Pekmezci, F. (2022). The framework of dimensionality from one-dimensionality to multidimensionality. *Uludağ University Journal of Education, (35)*3, 516-531. https://dx.doi.org/10.19171/uefad.1143164

Bass, B.M. (1990). *Handbook of leadership*. Free Press

Bellibaş, M.Ş., Bulut, O., Hallinger, P., & Wang, W.C. (2016). Developing a validated instructional leadership profile of Turkish primary school principals. *International Journal of Educational Research*, 75, 115–133. https://doi.org/10.14527/kuey.2018.015

Bennett, M., & Bennett, J. (2004). Developing intercultural sensitivity: An integrative approach to global and domestic diversity. In D. Landis, J. Bennett, M. Bennett (Eds) *The Handbook of İntercultural Training*, Third Edition. Sage.

Bennis, W.G. (2009). *On becoming a leader*. Basic Books.

Bhattacharyya, S.S. (2018). Development of a conceptual framework on real options theory for strategic human resource management. *Industrial and Commercial Training, 50*, 272–284. https://doi.org/10.1108/ICT-07-2017-0061

Bowen, S.A. (2016). Clarifying ethics terms in public relations from A to V, authenticity to virtue. BledCom special issue of PR Review sleeping (with the) media: Media relations. *Public Relations Review, 42*(4), 564–572. https://doi.org/10.1016/j.pubrev.2016. 03.012

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

Burns, J.M. (1978). *Leadership*. Harper & Row Publishers.

Byrne, B.M. (1998). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (1st ed.). Routledge. https://doi.org/10.4324/9780203805534

Campbell, A., & Yeung, S. (1991). Creating a sense of mission. *Long Range Planning, 24*(4), 10–20. https://doi.org/10.1016/0024-6301(91)90002-6

Catano, N., & Stronge, J. (2007). What do we expect of school principals? Congruence between principal evaluation and performance standards. *International Journal of Leadership in Education, 10*(4), 379–399. https://doi.org/10.1080/1360312070 1381782

Center for Promise, (2019). *Building systems of integrated student support a policy brief for local and state leaders.* https://files.eric.ed.gov/fulltext/ED602237.pdf

Cerit, Y., Kadıoğlu Ateş, H., & Kadıoğlu, S. (2018). Etkili okul müdürlerinin liderlik nitelikleri ile öğretmenlerin değişime açık olma düzeyleri arasındaki ilişki [The relationship between effective school principals' leadership qualities and teachers' openness to change]. *Kalem Uluslararası Eğitim ve İnsan Bilimleri Dergisi, 8*(1), 105-129. https://doi.org/10.23863/kalem.2018.95

Campbell, A., & Yeiung, S. (1991). Creating a sense of mission. *Long Range Planning, 24*(4),10-20.

Chen, G.M., & Starosta, W.J. (2000). The development and validation of the intercultural sensitivity scale. *Human Communication, 3*(1), 1–15.

Cheng, C.K. (2009) Cultivating communities of practice via learning study for enhancing teacher learning. *KEDI Journal of Educational Policy, 6*(1), 81-104.

Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255.

Chopra, D., & Sehgal, K. (2019). 7 characteristics of an effective leader. https://www.cnbc.com/2019/01/29/be-a-better-boss-7-traits-of-highly-effective-leaders.html

Ciulla, J.B. (2005). *Integrating leadership with ethics: Is good leadership contrary to human nature?.* In P. J.

Conger, J.A. (1999). Charismatic and transformational leadership in organization: an insider's perspective on these developing streams of research. *Leaderhip Quarterly, 10*(2), 145-179.

Comrey, A.L., & Lee, H.B. (1992). *A first course in factor analysis* (2nd ed.). Erlbaum.

Cornelissen, J. (2004). *Corporate communications: Theory and practice.* Sage.

Council of Chief State School Officers [CCSSO], (1996). *Interstate School Leaders Licensure Consortium (ISLLC): Standards for School Leaders*. Washington, D.C.

Council of Chief State School Officers, (1996). *Interstate school leaders licensure consortium standards for school leaders*. Washington, D.C., Author.

Day, D.V., & Antonakis, J. (2012). Leadership: past, present, and future. In: Day, D.V., & Antonakis, J. (Eds). *The Nature of Leadership*. (p.3-25). Sage Publications.

DBE [Department of Basic Education (South Africa)], (2019). *Annual Performance Plan*. Pretoria: Department of Basic Education.

DDE, (1998). *Delaware department of education, the delaware administrator standarts*. Delaware: Administrator Standards Advisory Committee.

Doğan Kılıç, E., Üstün, A., & Önen, Ö. (2011). Öğrenen örgütlerde etkili liderlik: Burdur örneği [Effective leadership in learning organizations: Burdur case]. *Educational Policy Analysis and Strategic Research, 6*(1), 5-22. https://dergipark.org.tr/en/pub/epasr/issue/17481/182981

Drucker, P.F. (2000): *21. Yüzyıl için yönetim tartışmaları* [Management debates for the 21st century]. (2. Basım), (Ç. Bahçıvangil ve Gorbon). Epsilon Yayıncılık.

Durnalı, M. (2018). *Öğretmenlere göre okul müdürlerinin teknolojik liderlik davranışları ve bilgi yönetimini gerçekleştirme düzeyleri. [The relationship between effective school principals' leadership qualities and teachers' openness to change]* [Unpublished doctoral dissertation]. Hacettepe Üniversitesi.

Dursun, M., Günay, M., & Yenel, İ.F. (2019). Çok yönlü liderlik yönelimleri ölçeği (ÇYLYÖ): Geçerlik ve güvenirlik çalışması [Multifaceted leadership orientations scale (MLEAS): validity and reliability study]. *Uluslararası Yönetim Akademisi Dergisi, 2*(2), 333–347. https://doi.org/10.33712/mana.596370

Elma, C. (2004), *Öğrenen örgütlerde takım çalışması*. Öğrenen Örgütler [Learning Organizations]. Edit. K. Demir & C. Elma. Sandal Yayınları.

Erer, M., & Demirel, E. (2018). Modern liderlik yaklaşımlarına genel bir bakış [An overview of modern leadership approaches]. *Journal of Institute of Economic Development and Social Researches, 4*(13), 647-656. https://doi.org/10.31623/iksad.109

Erkuş, A. (2012). *Psikolojide ölçme ve ölçek geliştirme I: Temel kavramlar ve işlemler* [*Measurement and scale development in psychology I: Basic concepts and procedures*]. Pegem Akademi.

Erkuş, A. (2022). Ölçek geliştirmeye hazırlık. In M. Acar Güvendir & Y. Özer Özkan (Eds.), *Tüm yönleriyle ölçek geliştirme süreci* (1st ed., pp. 1-25). Pegem Academy.

Fairholm, M. (2002). *Defining leadership: A review of past, present, and future ıdeas*. The George Washington University Center for Excellence in Municipal Management, Washington, DC 20052.

Finch, W.H. (2012). Distribution of variables by method of outlier detection. *Frontiers in Psychology, 3,* 1-12. https://doi.org/10.3389/fpsyg.2012.00211

Freeman, R.E., & Stewart, L. (2006). *Developing ethical leadership*. Bridge Paperso.

Froiland, J.M. (2019). *Employee retention*. Great Neck Publishing.

Fry, L.W. (2003). Toward a theory of spiritual leadership, *The Leadership Quarterly, 14*(6), 693–727. https://doi.org/10.1016/j.leaqua.2003.09.001

Grieder, S., & Steiner, M.D. (2022). Algorithmic jingle jungle: A comparison of implementations of principal axis factoring and promax rotation in R and SPSS. *Behavior Research Methods, 54*(1), 54-74. https://doi.org/10.3758/s13428-021-01581-x

Gronn, P. (2002). Distributed Leadership as a unit of analysis. *The Leadership Quarterly, 13*(4), 423-451. https://doi.org/10.1016/S1048-9843(02)00120-0

Gülbahar, B. (2014). Okul yöneticilerinin öğretim programlarının uygulanmasındaki öğretim liderliği rollerini belirlemeye yönelik bir alanyazın tarama çalışması [A literature review study to determine the instructional leadership roles of school administrators in the implementation of curricula]. *Milli Eğitim, 201*, 83-108. https://dergipark.org.tr/tr/pub/milliegitim/issue/36164/406521

Hair, J.F., Anderson, R.E., Tatham, R.L., & Black, W.C. (2010). *Multivariate Data Analysis* (7th ed.). Prentice Hall, Inc. https://doi.org/10.1016/j.iheduc.2013.06.002

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

Hargreaves, A., & Fink, D. (2006). *Sustainable leadership*. Jossey-Bass cop.

Heck, R.H., & Hallinger, P. (2010). Testing a longitudinal model of distributed leadership effects on school improvement. *The Leadership Quarterly*, *21*, 867-885. https://doi.org/10.1016/J.LEAQUA.2010.07.013

Hemphill, J.K., & Coons, A.E. (1957). Development of the leader behavior description questionnaire. In Stogdill RM, Coons AE (Eds.), *Leader behavior: Its description and measurement*, 6–38. Ohio State University.

Hess, F.M., & Kelly, A.P. (2005). *Learning to lead? What gets taught in principal preparation programs*. Harvard University, Kennedy School of Government.

Hırlak, B., & Taşlıyan, M. (2018). Otantik liderliğin demografik özellikler açısından incelenmesi [Examining authentic leadership in terms of demographic characteristics]. *Uluslararası Toplumsal Araştırmaları Dergisi, 8*(15), 1080–1085. https://doi.org/10.26466/opus.439692

Hodgkinson, C. (2008). *Yönetim felsefesi, örgütsel yaşamda değerler ve motivasyon [Management philosophy, values and motivation in organizational life*]. Çev: İ.Anıl ve B. Doğan. Beta Yayım.

Hoy, W.K., & Miskel, C.G. (2012). Eğitim yönetimi, teori, araştırma ve uygulama [Educational administration, theory, research and practice]. (Ç. S. turan). Nobel.

Ireland, R.D., & Hitt, M.A. (2005) Achieving and maintaining strategic competitiveness in the 21st century: The role of strategic leadership. *Academy of Management Executive, 19*, 63-77. http://dx.doi.org/10.5465/AME.2005.19417908

ISLLC (2008). *Educational leadership policy standards: ISLLC 2008*, Council of Chief State School Officers, Washington, DC.

İlğan, A., & Ekiz, M. (2020). Okul müdürleri etik beklenti ölçeği ile etik liderlik ölçeğinin geliştirilmesi: geçerlik ve güvenirlik çalışması [Development of school principals' ethical expectation scale and ethical leadership scale: validity and reliability study]. *İş Ahlakı Dergisi, 13*(2), 38-81 https://doi.org/10.12711/tjbe.2020.13.2.0159

Jacob, N.E., & Shari, B. (2013). Organizational effectiveness in educational ınstitutions. *EDUCARE International Journal for Educational Studies, 6*(1), 17-26.

John, P., & Cole, A. (1999). Political leadership in the new urban governance: Britain and France compared. *Local Government Studies, 25*(4), 98-115. https://doi.org/10.1080/03003939908433969

Jones, M., & Harris, A. (2014), Principals leading successful organisational change: building social capital through disciplined professional collaboration. *Journal of Organizational Change Management, 27*(3), 474–485. https://doi.org/10.1108/JOCM-07-2013-0116

Jöreskog, K.G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Scientific Software International: Skokie, IL, USA. Available online: https://psycnet.apa.org/record/1993-97878-000

Judge, T.A., & Bono, J.E. (2000). Five-factor model of personality and transformational leadership. *Journal of Applied Psychology, 85*(5), 751–765.

Kabacoff, R. (1998). *Leadership effectiveness analysis: technical considerations. (Available from the Management Research Group).* http://www.eric.edu.gov

Komives, S., & Dugan, I.P. (2010). *Contemporary leadershıp theorıes, Couto, R. A. (Ed.). Political and Civic Leadership: A Reference Handbook (Vol. 1.)*. Sage.

Korkmaz, S., Göksuluk, D., & Zararsız, G. (2014). MVN: An R package for assessing multivariate normality. *The R Journal*, *6*(2), 151-162.

Kline, R.B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). The Guilford Press.

Klingborg, D.J., Moore, D.A., & Varea Hammond, S. (2006). What is the leadership? *Journal of Veterinary Medical Education*, *33*(2), 279-283. https://doi.org/10.3138/jvme.33.2.280

Kurt, T. (2005). MEB tarafından başlatılan eğitime destek projesinin değerlendirilmesi [Evaluation of the education support project initiated by MEB]. *Çağdaş Eğitim Dergisi, 31*(329), 23-29.

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

Leithwood, K., Steinbach, R., & Jantzi, D. (2002). School leadership and teachers' motivation to ımplement accountability policies. *Educational Administration Quarterly, 38*(1), 94-119. https://doi.org/10.1177/0013161X02381005

Leithwood, K., & McCullough, C. (2017), *Strong districts and their leadership project: final report of research strand.* Final Report of Research to the Council of Ontario Directors of Education, Toronto.

Leithwood, K., Harris, A., & Hopkins, D. (2020). Seven strong claims about successful school leadership revisited. *School Leadership and Management, 40*(1), 5-22. https://doi.org/10.1080/13632434.2019.1596077

Leithwood, K., Louis, K.S., Anderson, K., & Wahlstrom, K. (2022). *How leadership influences student learning*, Wallace, www.learningfromleadership.umn.edu

Leys, C., Klein, O., Dominicy, Y., & Ley, C. (2018). Detecting multivariate outliers: Use a robust variantof the Mahalanobis distance. *Journal of Experimental Social Psychology, 74*, 150–156. https://doi.org/10.1016/j.jesp.2017.09.011

Lissack, M., & Roos, J. (2001). Be coherent, Not visionary. *Long Range Planning*, 34, 53-70.

Liu, S., Hallinger, P., & Feng, D. (2016). Supporting the professional learning of teachers in China: Does principal leadership make a difference? *Teaching and Teacher Education,* 59, 79-91. https://doi.org/10.1016/j.tate.2016.05.023

Mantiri, O. (2011). Professional ethics in teaching. *SSRN Electronic Journal*, *17*, 1-12. http://dx.doi.org/10.2139/ssrn.2566122

Maier, A., Daniel, J., Oakes, J., & Lam, L. (2017). *Community schools as an effective School Improvement Strategy: A review of the evidence*. Learning Policy Institute.

Manzoor, S.R., Ullah, H., Hussain, M., & Ahmad, Z.M. (2011). Effect of teamwork on employee performance. *International Journal of Learning & Development, 1*(1), 110–126. http://dx.doi.org/10.5296/ijld.v1i1.1110

McDonald, R.P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, *34*(1), 100-117. https://doi.org/10.1111/j.2044-8317.1981.tb00621.x

McDonald, M., Kazemi, E., & Kavanagh, S.S. (2013). Core practices and pedagogies of teacher education: A call for a common language and collective activity. *Journal of Teacher Education*, *64*(5), 378–386. https://doi.org/10.1177/0022487113493807

Ministry of Education, (2007). *Planlı okul gelişim modeli, okulda stratejik yönetim [Planned school development model, strategic management at school]*. Milli Eğitim Bakanlığı Yayınları.

Menbarrow, Z. (2021). The importance and necessity of professional ethics in the organization and the role of managers. *Psychology and Behavioral Science, 18*(1), 1-6. https://doi.org/10.19080/PBSIJ. 2021.18.555979

Merriam-Webster (2023). *Dictionary online*. Retrieved; March 10, 2023.

Moran, R.T., Harris, P.R., & Moran, S. (2007). *Managing cultural differences*. Routledge.

Murphy, J. (2005). *Connecting teacher leadership and school ımprovement*. Corwin.

NAESP. (2001). *Leading learning commmunities: Standards for what principals should know and be able to do*. National Association of Elementary School Principals.

Northouse, P.G. (2007). *Leadership theory & practice* (8th ed.). Sage. [Google Scholar]

NPBEA. (2015). *Professional standards for educational leaders 2015.* Reston, VA: Author. http://www.npbea.org

Okutan, M. (2003). Okul müdürlerinin liderlik davranışları [Leadership behaviors of school principals]. *Milli Eğitim Dergisi*, 157. http://dhgm.meb.gov.tr/yayimlar/dergiler/Milli_Egitim_Dergisi/157/okutan.htm

Osborne, J.W., & Fitzpatrick, D.C. (2012). Replication analysis in exploratory factor analysis: What it is and why it makes your analysis better. *Practical Assessment, Research and Evaluation, 17*(15), 1-8. https://doi.org/10.7275/h0bd-4d11

Penlington, C., Kington, A., & Day, C. (2008). Leadership in improving schools: a qualitative perspective. *School Leadership and Management, 28*(1), 65-82. https://doi.org/10.1080/13632430701800086

Pont, B., Nusche, D., & Moorman, H. (2008). Improving school leadership, Volume 1: *Polıcy and practice*. OECD Publishing.

Rengi, Ö. (2014). *Sınıf öğretmenlerinin kültürel farklılık algıları ve kültürlerarası duyarlılıkları [Classroom teachers' perceptions of cultural differences and intercultural sensitivity]*. [Unpublished master thesis]. Kocaeli Üniversitesi.

Sabuncuoğlu, Z., & Tüz, M. (2001). *Örgütsel psikoloji [Organizational psychology]*. Ezgi Kitabevi.

Sezer, Ş. (2018). Okul müdürlerinin eğitimsel liderlik standartlarını karşılama düzeyi: Ölçek geliştirme ve uygulama çalışması [School principals' level of meeting educational leadership standards: Scale development and application study]. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi, 19*(1), 818-836. https://dergipark.org.tr/tr/pub/kefad/issue/59094/850540

Sharma, M.K., & Shilpa, J. (2013). Leadership management: Principles, models and theories, *Global Journal of Management and Business Studies, 3*(3), 309-318.

Sılver, S.R. (1990). *Perception of empowerment in engineer workgroups: The linkage to transformational leadership and performance* [Unpublished doctoral thesis]. George Washington University.

Spillane, J., Halverson, R., & Diamond, J. (2007). Toward a theory of school leadership practice: Implications of a distributed perspective. *Journal of Curriculum Studies, 36*(1), 3–34. https://doi.org/10.1080/0022027032000106726

Şencan, H., & Fidan, Y. (2020). Likert verilerinin kullanıldığı keşfedici faktör analizlerinde normallik varsayımı ve faktör çıkarma üzerindeki etkisinin SPSS, factor ve prelis yazılımlarıyla sınanması. [Normality assumption in the exploratory factor analysis with likert scale data and testing its effect on factor extraction]. *Business & Management Studies: An International Journal, 8*(1), 640-687. http://dx.doi.org/10.15295/bmij.v8i1.1395

Şişman, M. (2009). Öğretmen yeterlikleri: Modern bir söylem ve retorik [Teacher competencies: A modern discourse and rhetoric]. *İnönü Üniversitesi Eğitim Fakültesi Dergisi, 10*(3), 63-82.

Stodd, J. (2022). *Articles on social leadership*. http://julianstodd.wordpress.com/?s=social+leadership

Stogdill, R.M. (1948). Personal factors associated with leadership: a survey of the literature. *Journal of Psychol*, *25*, 35–71. https://doi.org/10.1080/00223980.1948.9917362

Stogdill, R.M. (1963). *Manual for the leader behavior description questionnaire: Form 12 Columbus*. Ohio State University Bureau of Business Research, College of Commerce and Administration.

Stone, A.G., & Patterson, K. (2005). *The history of lieadership focus, servant leadership routdtable*. Regent University.

Sun, J., & Leithwood, K. (2012) Transformational school leadership effects on student achievement. *Leadership and Policy in Schools, 11*(4), 418-451. https://doi.org/10.1080/5700763.2012.681001

Sutherland, I., & Gosling, J. (2010). Cultural leadership: Mobilizing culture from affordances to dwelling. *The Journal of Arts Management, Law, and Society*, *40*, 6-26.

Tabachnick, B.G., & Fidell, S.L. (2001). *Using multivariate statistics.* Allyn & Bacon Publishing.

Tabachnick, B.G., & Fidell, S.L. (2009). *Using multivariate statistics.* Allyn & Bacon Publishing.

Taylor, C.M., Cornelius, C.J., & Colvin, K. (2014). Visionary leadership and its relationship to organizational effectiveness. *Leadership & Organization Development Journal*, *35*(6), 566-583. https://doi.org/10.1108/LODJ-10-2012-0130

Taymaz, H. (2021). *İlköğretim ve ortaöğretim okul müdürleri için okul yönetimi [School management for primary and secondary school principals]*. (12. Baskı), Pegem Akademi.

Turan, S., & Ebiçoğlu, N. (2002). Okul müdürlerinin liderlik özelliklerinin cinsiyet açısından değerlendirilmesi [Evaluation of school principals' leadership characteristics in terms of gender]. *Eğitim Yönetimi Dergisi*, *3*, 444-458. https://dergipark.org.tr/tr/pub/kuey/issue/10366/126886

Verbiest, E., Ansems, E., Bakx, A., Grootswagers, A., Heijmen Versteegen, L., & Jongen, T. (2005). *Collective learning in schools described: building collective learning capacity*. The ICSEI Conference, Barcelona, Spain.

Wart, M.V. (2013). Lessons from leadership theory and the contemporary challenges of leaders. *Public Administration Review*, *73*(4), 553-565. https://doi.org/10.1111/puar. 12069

Werner, İ. (1993). *Liderlik ve yönetim [Leadership and management*]. (Çev: Vedat Üner). Rota Yayınları.

Wildavsky, A. (2006), *Cultural analysis*, (ed. B. Swedlow, D. Coyle, R. Ellis, R. Kagan & A. Ranney). Translation.

Wu, D.A., Li, Z., & Zumbo, B.D. (2007). Decoding the meaning of factorial invariances and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment Research and Evaluation*, *12*(3), 1-26.

Yalçın, M.T., & Atasoy, R. (2021). Okul liderliği ölçeği: Geçerlik ve güvenirlik çalışması [School leadership scale: Validity and reliability study]. *Karamanoğlu Mehmetbey. Uluslararası Eğitim Araştırmaları Dergisi*, *23*(2), 103-112, http://dergipark.gov.tr/ukmead

Yılmaz, E. (2006). *Okullardaki örgütsel güven düzeyinin okul yöneticilerinin etik liderlik özellikleri ve bazı değişkenler açısından incelenmesi [Examining the level of organizational trust in schools in terms of school administrators' ethical leadership characteristics and some variables]* [Unpublished doctoral thesis]. Selçuk Üniversitesi.

Yukl, G. (2010). *Leadership in organizations* (7th edition). Pearson: Prentice Hall, Upper Saddle River.

Zaccaro, S.J., & Klimoski, R.J. (2001). The nature of organizational leadership: An introduction. In S. J. Zaccaro & R.J. Klimoski (Eds.), *The nature of organizational leadership: Understanding the performance imperatives confronting today's leaders* (pp. 3–41). Jossey-Bass/Wiley.

Zorlu, K., & Korkmaz, F. (2020). Sürdürülebilir liderlik ölçeğinin Türkçe'ye uyarlanması: geçerlik ve güvenirlik çalışması [Adaptation of the sustainable leadership scale into Turkish: validity and reliability study]. *Gazi Akademik Bakış, 13*(26), 199–213. https://dergipark.org.tr/tr/pub/gav/issue/54827/750462

APPENDIX

Appendix 1. *Repeatability result table of EFA.*

Item No	Factor loading values in the first sample	Factor loading values in the second sample	Absolute value of difference
M1	0.753	0.797	0.044
M2	0.832	0.788	0.044
M3	0.861	0.844	0.017
M4	0.892	0.791	0.101
M5	0.848	0.774	0.074
M6	0.890	0.839	0.051
M7	0.853	0.855	0.002
M8	0.839	0.829	0.010
M9	0.872	0.783	0.089
M11	0.893	0.856	0.037
M12	0.907	0.839	0.068
M13	0.890	0.827	0.063
M14	0.908	0.863	0.045
M15	0.918	0.842	0.076
M16	0.922	0.811	0.111
M17	0.854	0.819	0.035
M18	0.864	0.852	0.012
M19	0.873	0.764	0.109
M20	0.893	0.859	0.034
M21	0.922	0.838	0.084
M22	0.891	0.903	0.012
M24	0.896	0.796	0.100
M25	0.923	0.904	0.019
M26	0.819	0.749	0.070
M27	0.909	0.879	0.030
M28	0.876	0.815	0.061
M29	0.848	0.801	0.047
M30	0.890	0.808	0.082
M31	0.890	0.830	0.060
M32	0.889	0.801	0.088
M33	0.925	0.891	0.034
M34	0.884	0.846	0.038
M35	0.875	0.852	0.023
M36	0.892	0.840	0.052