e-ISSN: 2148-7456 | a peer-reviewed online journal | hosted by DergiPark

International Journal of Assessment Tools in Education

Volume: 9 Issue: 2 June 2022

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

e-ISSN 2148-7456

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

Volume 9 Issue 2 2022

Editor : Dr. Hakan KOGAR
Address : Akdeniz University, Education Faculty,
Dumlupinar Bulvari 07058 Kampus Antalya / TÜRKİYE
Phone : +90 242 227 4400 Extension: 6079
E-mail : ijate.editor@gmail.com; hakankogar@akdeniz.edu.tr

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

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

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

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

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

International Journal of Assessment Tools in Education

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

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

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

Indexes and Platforms:

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

CONTENTS

Research Articles

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 2, 267–289
https://doi.org/10.21449/ijate.888936
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Problematic Technology Use Scale for Young Children (PTUS-YC): Validity and Reliability Study

Ahmet Sami Konca[1,*], **Onder Baltaci**[2], **Omer Faruk Akbulut**[3]

[1]Erciyes University, Faculty of Education, Department of Early Childhood Education, Kayseri, Türkiye
[2]Kirsehir Ahi Evran University, Faculty of Education, Department of Educational Sciences, Kirsehir, Türkiye
[3]Necmettin Erbakan University, Institute of Educational Sciences, Department of Educational Sciences, Konya, Türkiye

ARTICLE HISTORY

Received: Mar. 01, 2021
Revised: Jan. 23, 2022
Accepted: Mar. 05, 2022

Keywords:
Problematic technology use,
Preschool children,
Early childhood,
Scale development,
Parents.

Abstract: This study aimed to develop a measurement tool to identify preschool children's problematic technology use levels and contribute to Turkish literature. The study group included in the exploratory factor analysis was composed of 357 voluntary children in the preschool period. The study group included in the confirmatory factor analysis, proximal validity analysis, item discrimination analysis, and reliability analysis was composed of a total of 402 parents. Exploratory factor analysis (EFA) was conducted to present the factor structure of PTUS-YC. Confirmatory factor analysis (CFA) was performed to test the model obtained with EFA. Item discrimination values obtained as a result of the independent sample t-test was investigated to determine the internal validity of the measurement tool. Furthermore, the correlation values between PTUS-YC items and the relevant factors and the complete measurement tool were calculated, and total item correlation was used to test whether each item served a common purpose. In addition, CR and AVE values were examined in the proximal validity analysis conducted for PTUS-YC. Reliability analysis for PTUS-YC was performed using Cronbach alpha internal consistency coefficient and McDonald's Omega coefficient methods. As a result of the exploratory factor analysis, a 4-factor structure that explained 60.392% of the total variance was obtained: continuity of use, resistance to control, effects on development and deprivation-escape. Based on the results, PTUS-YC was a valid and reliable measurement tool that can be used to determine preschool children's problematic technology use levels.

1. INTRODUCTION

The rapid advances in technology in the 20th century have generated many changes in individuals' lives. Today, technological tools perform many different activities such as communication, shopping, following interactive content, benefiting from educational services and gaming. Technological tools can facilitate the lives of individuals with various access opportunities they offer. However, they can also generate problems since technology is used more and more at younger ages, and technological tools may be used in a manner that may negatively affect the development of younger users. A study conducted on the internet use of children ages 3 to 18 in the United States demonstrated that children's rate of internet use increased rapidly between 2010 and 2017 (National Center for Education Statistics, 2019).

*CONTACT: Ahmet Sami Konca samikonca@erciyes.edu.tr Erciyes University, Faculty of Education, Department of Early Childhood Education, Kayseri, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

These rates may have increased even higher from 2017 to 2021, mainly due to the rapid technological tools and opportunities they offer. With the help of a program developed to control technology use, a study exploring the content and the time spent by children in technological environments in 2020 concluded that 39.11% of children spent extensive amounts of time watching videos and listening to music, followed by 24.16% spending time on internet communication tools and 15.98% spending time on games (Securelist, 2020). Considering the purposes of internet use by children, it can be argued that children spend a long time with many activities that may risk their development. A study conducted with parents showed that parents under the age of 12 had concerns about the time their children spent in front of the screen. In addition, 71% of the parents thought that smartphones would do more harm than good to their children's development. In the study, 80% of the parents stated that their 5-11-year-old children used tablets or interacted with them, while this rate was 48% for parents with children under five years old. 66% of the parents who participated in the study stated that parenting was more complex than in the past due to children's widespread use of technological tools. Finally, the research investigated children's screen time for various technological tools. 88% of children aged 0-11 spent time on television, 67% on tablets, 60% on smartphones, 44% on computers and 44% on gaming devices (Pew Research Center, 2020). When these studies conducted on large samples are considered in general, it can be argued that today, children often spend time with technological tools and technology development. This may pose a risk for the development of children who have not yet completed their physical, psychological and social development (Rideout et al., 2011). In this context, the American Academy of Pediatrics (2016) emphasized that newborn children should not interact with technology until 18 months of age and that technology use by children older than 18 months should be limited to 1 hour with parent-controlled content. Similarly, the Turkish Green Crescent Society (2021) draws attention to these situations in its statements. Due to a lack of knowledge and skills, children and their parents may have difficulty using technological tools in a healthy way most of the time. This points to the use of technology that may cause problems in children's lives and can be considered as "problematic technology use".

Having difficulty controlling the use of the smartphone and causing deterioration in the daily functions of the individual are considered as problematic smartphone use while experiencing these problems in digital games reveals challenging gameplay. These sub-headings are handled within the framework of problematic technology use. Based on different concepts of problematic technology use in the literature such as problematic internet use, problematic gaming disorder, problematic social media use, problematic smartphone use, problematic technology use can be defined as a concept expressing the difficulties experienced by persons in controlling their use of technological tools, the problems experienced by persons when they stay away from technological tools, and the negative effect of technology use on their physical, psychological and social lives (Caplan, 2010; Young, 2011). Children's physical, psychological and social lives can be negatively affected due to problematic technology use (Avşaroğlu & Akbulut, 2020; Mustafaoğlu et al., 2018). Relevant studies in the literature present that children's problematic technology use negatively affect their social interaction with peers (Savcı & Aysan, 2016; Yavuzer, 2019; Zorbaz & Tuzgöl Dost, 2014), their family relationships (Gunuc & Dogan, 2013; Lam et al., 2009; Lee & Chae, 2007; Wu et al., 2016), their language development (Chonchaiya & Pruksananonda, 2008; McCarrick & Li, 2007), their psychological states (Akboğa & Gürgan, 2019; Derin & Bilge, 2016; Orben & Przybylski, 2019; Plowman et al., 2010) and their academic lives (Anlayışlı & Bulut Serin, 2019; Chi et al., 2020; Geng et al., 2018; Peng et al., 2019; Soldatova & Teslavskaia, 2017; Zhang et al., 2018). In addition, problematic use of technology is associated with problems such as attention problems (Christakis et al., 2004; Kawabe et al., 2019; Soldatova & Teslavskaia, 2017), behavioural problems (Alonso & Romero, 2017; Ybarra et al., 2011), sleep problems (Bruni et al., 2015;

Calamaro et al., 2012; Cespedes et al., 2014; Fuller et al., 2017; Johansson et al., 2016; Mei et al., 2018), nutritional problems (De Jong et al., 2013; Mitchell et al., 2013; Rosen et al., 2014) as well as reductions in physical activity (Cox et al., 2012; Rosen et al., 2014) and developmental problems (Howie et al., 2017; Pagani et al., 2010).

Although problematic use of technology is a problem in all children, this study addressed the problematic use in preschool children. The preschool period includes various critical developmental tasks in regard to the physical, psycho-motor, psychological and social development of children. Sustaining this period in a healthy way is of great importance for the development of children. Technology use in preschool children is on the rise with the widespread use of technology in younger age groups in recent years, and it may pose risks for the development of preschool children.

A study conducted by Genc (2014) on parents with children in the preschool period found that children spent an average of 2-3 hours a day on technological tools. It can be argued that this time spent by preschool children by using technological tools can be defined as a risk-based on the healthy use criteria identified by international organizations (American Academy of Pediatrics, 2016). In addition, according to the study, parents stated that their children's use of technology could cause problems because of radiation, the inappropriateness of the tools regarding their developmental characteristics, possible adverse effects on their development, health hazards and social withdrawal (Genc, 2014). Many reports and studies in the literature show that the problematic use of technology in the preschool period can cause various damages in regards to the physical, psychological and social development of children (Cox et al., 2012; National Association for the Education of Young Children, 2012; Turkish Green Crescent Society, 2021). Preschool children's problematic technology use is associated with problems such as obesity, sleep disorders, behavioural problems, poor academic performance, social and language development problems (Anderson & Pempek, 2005; Christakis et al., 2004; DeLoache et al., 2010; Rogow, 2007; White House Task Force on Childhood Obesity Report to the President, 2010). In this context, it is of great importance to evaluate children's problematic technology use and ensure that parents are provided with skills regarding the healthy use of technology.

In Turkey, several measurement tools are used to identify children’s and adolescents’ problematic technology use levels (Kabadayı, 2020). The target group is mostly adolescents in these measurement tools which were developed or adapted into Turkish (Anlı & Taş, 2018; Arıcak et al., 2019; Bayraktar, 2001; Çakıroğlu & Soylu, 2019; Canan et al., 2010; Canoğulları-Ayazseven & Cenkseven-Önder, 2019; Ceyhan & Ceyhan, 2014; Eşgi, 2014; Fidan, 2016; Fırat & Balcı-Çelik, 2017; Günüç & Kayri, 2010; Güzeller & Coşguner, 2012; Ilgaz, 2015; Irmak & Erdoğan, 2015; Kaya, 2013; Kutlu et al., 2016; Ögel et al., 2015; Şar et al., 2015; Taş, 2017, 2019). In addition to these measurement tools, there are a limited number of measurement tools used to determine problematic use of various technological tools for secondary school students (Balantekin, 2009; Hazar & Hazar, 2017; Mıhçı & Çakmak, 2017; Yılmaz et al., 2017), primary school students (Horzum et al., 2008) and preschool students (Ünsal & Ulutaş, 2019). These measurement tools used in Turkey only focus on identifying children's and adolescents' problematic use of a specific technological tool such as gaming, computer, Internet and smartphone. In addition, the measurement tools to identify preschool children's problematic technology use in Turkey are somewhat limited. For instance, the measurement tool adapted into Turkish by Ünsal and Ulutaş (2019) solely aims to identify preschool children’ computer game addiction levels. Another measurement tool adapted to Turkish by Furuncu and Öztürk (2020) examines preschool children's problematic media use levels. The items in this measure were developed by considering the criteria for Internet Gaming Disorder found in DSM-V. This measurement tool, adapted to Turkish, has a 27-item long-form and a 9-item short form with

one dimension in both forms. One-dimension structure, especially in the long-form in this measurement tool, can be regarded as a limitation because the problematic technology use is a complex concept with various dimensions. In the international literature, measurement tools are developed to measure the problematic technology use of individuals in different age groups (Ding et al., 2018; Foerster et al., 2015; Pancani et al., 2019). It has been observed that these measurement tools are generally developed for secondary school, high school, university students and adults (Harris et al., 2020). In addition, it is seen that the measurement tools that can measure the problematic technology use levels of young children are limited. The measurement tool developed by Domoff et al. (2019) to measure the problematic media use levels of children aged 4-11 is one. Furuncu and Öztürk (2020) adapted this measurement tool into Turkish.

Investigating the relationship between children and technology and describing their use of technological tools can be crucial in preventive and protective services in the preschool period, which includes many critical development tasks for children's development. In addition, the lack of a measurement tool with a broader framework to identify the preschool children's problematic technology use in Turkey and the world can be regarded as a shortcoming. In this context, this study aimed to develop a measurement tool to determine the level of problematic technology use of children in the preschool period and contribute to Turkish literature.

2. METHOD

This research was designed as a scale development study since the main aim was to develop a tool to measure preschool children' problematic technology use.

2.1. Study Group

Two different study groups were used in conducting the validity and reliability studies in this research. The study group included in the exploratory factor analysis of the research was composed of 357 voluntary parents (334 mothers and 23 fathers) with children in the preschool period. Among these voluntary participants, 101 parents were primary or secondary school graduates, 102 were high school graduates, and 154 were university graduates. The mean age of this parent group was 33.2. The study group included the confirmatory factor analysis, proximal validity analysis, item discrimination analysis, and reliability analysis composed of 402 voluntary parents (372 mothers and 30 fathers) with children in the preschool period. Among these voluntary participants, 103 parents were primary or secondary school graduates, 108 were high school graduates, and 191 were university graduates. The mean age of this parent group was 33.4. According to Bryman and Cramer (2011), in determining the number of participants in factor analysis, it may be sufficient to reach five or ten times the number of items in the measurement tool. Hence the study group can be argued to be large enough for both validity and reliability analyses.

2.2. Ethical Statement

The principles of scientific research and publication ethics were adhered to during the planning and implementation of this research. Approval was obtained from the Social and Human Sciences Ethics Committee of Erciyes University (Document No: 2020/176) at the beginning of the research. While collecting the data, the participants were informed about the study, an informed consent form was collected from each participant, and the data were collected based on voluntary participation. The data obtained from the research were not shared with any person or institution and were used only within the scope of this research.

2.3. Development Process of the Measurement Tool

The process of developing a psychological measurement tool can be defined as the process and procedure of creating the expressions that will stimulate the relevant characteristics of

individuals that are intended to be measured and developing the appropriate response categories for these expressions (Erkuş, 2012). First of all, conducting a detailed theoretical and conceptual review of the relevant field is required so that the feature desired to be measured in individuals can be transformed into items in an appropriate manner. In this context, first of all, a detailed literature review was conducted on the topics included in problematic technology use, such as technology addiction, gaming addiction, internet addiction, problematic internet use and smartphone addiction to develop the item pool for Problematic Technology Use Scale for Young Children (PTUS-YC). In addition, the diagnostic criteria for internet addiction identified by various researchers were also taken into consideration. Also, previous studies on children's technology use in the preschool period were examined since the measurement tool developed in this study aimed to identify children's problematic technology use levels in the preschool period. Furthermore, the content developed by the American Academy of Pediatrics (2016), Turkish Green Crescent Society (Türkiye Yeşilay Cemiyeti, 2021) and the National Association for the Education of Young Children (2012), which explored the use of technology by developmental stages, were examined in detail in regards to the development of preschool children. In this framework, the risks that can be addressed in terms of preschool children's use of technology were added to the item pool by considering problematic use. An item pool was created by the researchers based on the literature review. While creating the item pool, structures related to the dimensions of problematic technology use such as duration, deterioration in functionality, tolerance, conflict and lack of control were considered. In order to ensure the content and face validity of the item pool, expert opinions were obtained from two field experts with doctoral dissertations in the field of problematic technology use and one assessment-evaluation specialist. Additionally, an expert from the field of Turkish education was consulted to ensure that the items in the item pool had proper language usage and grammar and were clear to the reader. Appropriate corrections were made within the framework of the feedback from field experts, and the first form of the measurement tool was developed as a 5-point Likert type scale with 31 items. "My child spends more than 1.5 hours a day with technological tools on average.", "My child's use of technological tools causes him to be inactive." and "When my child feels sad, he relaxes by spending time with technological tools." statements are some examples among these items. In order to understand whether these items were understandable by the respondents, six parents were asked to fill in the measurement tool within the scope of the pilot study and whether the relevant items were clear and understandable. Some expressions have been simplified within the framework of these feedbacks. Later, the measurement tool was implemented online to 357 parents with children in the preschool period to perform the exploratory factor analysis and to 402 parents with children in the preschool period to perform the confirmatory factor analysis, proximal validity analysis, and item discrimination and reliability analysis.

2.4. Data Analysis

SPSS 25.00 and AMOS 24.00 programs were used to analyze the data obtained from two different study groups to conduct this research's validity and reliability studies. An exploratory factor analysis (EFA) was conducted on the data obtained from the study group of 357 people in the first stage of the study to reveal the factor structure of PTUS-YC. First, KMO and Barlett test analyzes were conducted to determine whether the data set was suitable for factor analysis. When it was confirmed that the data obtained from the analysis were suitable for factor analysis, exploratory factor analysis was conducted. In addition, a normality test was conducted for the whole measurement tool. The principal axis factoring method was used when performing the exploratory factor analysis. Factor loadings were calculated using the Varimax rotation technique (Balcı, 2009). The process was continued by removing the items with factor loads below .30 and items with a difference between the two-factor loadings of less than .10 based on the principal axis factoring analysis (Can, 2019; Eroğlu, 2008; Kline, 2011). In the field of

behavioural sciences, it was considered to be sufficient in the process of developing or adapting a measurement tool when the factor loads of the items were higher than .30, and at least 40% of the total variance could be explained with the items in the measurement tool (Kline, 2011). The factor loads of the items were taken into consideration while evaluating the factor analysis results (Balcı, 2020). In addition, common factor variance value is also known to be important in multi-factor structures. In this context, it was checked whether there was any item with a common factor variance below .20 (Çokluk et al., 2018). In the second stage of the study, the data obtained from a study group of 402 people were tested by performing a confirmatory factor analysis (CFA) to test the model obtained due to EFA. As a result of the analysis, the results were addressed based on the model fit indices' acceptable and perfect fit ranges (Schumacker & Lomax, 2004). The item discrimination values obtained as a result of the independent sample t-test were examined to determine the internal validity of the measurement tool and construct validity. Also, CR and AVE values were examined in the proximal validity analysis of PTUS-YC. Furthermore, the correlation values between PTUS-YC items and the relevant factors and the complete measurement tool were calculated, and total item correlation was used to test whether each item served a common purpose.

The reliability analysis of PTUS-YC was performed using the Cronbach alpha internal consistency coefficient and McDonald's Omega coefficient methods. Reliability was interpreted by considering that the reliability coefficients should be .70 and above for measurement tools to be reliable (Can, 2019).

3. RESULT

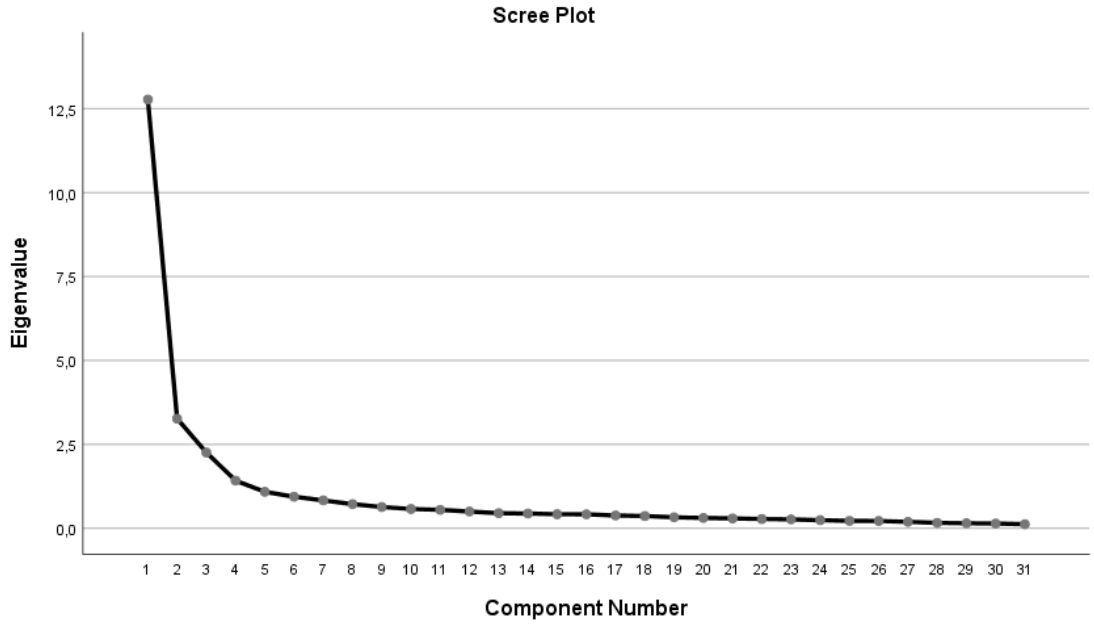
This section presents the results of the analyses conducted in line with the purpose of the research.

3.1. Construct Validity

3.1.1. *Exploratory factor analysis*

Within the scope of the research, exploratory factor analysis (EFA) was performed to examine the factor structure of PTUS-YC. KMO coefficient and Bartlett test coefficient were found to be .926 and 6328.143 (p<.001) respectively during the KMO and Barlett tests performed to determine the suitability of the data obtained from the first study group factor analysis. The finding that KMO was higher than .60 and Barlett test was significant shows that the data were suitable for factor analysis (Pallant, 2017). In addition, the skewness and kurtosis values were examined in order to comment on normality. Skewness and kurtosis values between -2 and +2 demonstrate normality of distribution (Karagöz, 2016). The skewness and kurtosis values examined for the whole measurement tool were found to be between -2 and +2, and based on these values, it was confirmed that the measurement tool had a normal distribution. When the anti-image correlation matrix was examined in order to decide whether each item was included in the factor analysis, it was observed that the values at the intersection point were over 0.5 (Can, 2019). Afterwards, Principal Axis Factoring and Varimax rotation method were used to determine whether the measurement tool was unidimensional or multidimensional. When determining the items to be included in the measurement tool in the exploratory factor analysis, attention was paid to ensure that the eigenvalue of items was at least 1.00, that the item factor load value was at least .30, that the items were included in a single factor and that there was at least .10 difference between the items with sufficient factor loading (Seçer, 2018). When the total variance values of the measurement tool and the line graph were examined, it was observed that the tool could have a 4-factor structure. A line graph was used to determine the factor number of the measurement tool, as shown in Figure 1.

Figure 1. *Eigenvalue factor plot for PTUS-YC.*



Then, using the Varimax technique, the factor load values of the items were examined by rotation, and it was found that items 11, 18, 22 and 27 were overlapping. As a result of excluding these items from the analysis, a four-factor structure that explained 60.392% of the total variance was obtained. Table 1 presents the findings obtained from the exploratory factor analysis. The result of exploratory factor analysis demonstrated that the measurement tool's items' factor load values were between .374 and .826. Considering that the factor load values of the items should be at least .30 in the measurement tool development process, it can be argued that the item factor load values of the items in the measurement tool were sufficient (Hair et al., 2006). In addition, considering that the common factor variances should not be less than .20, it can be claimed that the factor common variance values of the items in the measurement tool were sufficient (Çokluk et al., 2018).

Based on the exploratory factor analysis result, a four-factor structure that explained 60.392% of the total variance was obtained. These factors were named based on their content (see Appendix). The first of these factors, the "continuity of use" sub-dimension, was composed of 8 items. The second factor, the "resistance to control" sub-dimension, included six items. The third factor, the "effect on development" sub-dimension, had five items. The fourth factor, the "deprivation-escape" sub-dimension, consisted of 7 items.

Table 1. *Item factor loads of PTUS-YC, variances explained by sub-scales and item analysis.*

Item	Continuity of Use	Resistance to Control	Effect on Development	Deprivation-Escape	Factor Common Variance
2	.801				.690
1	.764				.589
4	.725				.667
3	.685				.643
26	.591				.716
25	.560				.619
23	.540				.546
17	.524				.503
31		.690			.611
28		.675			.628
15		.672			.475
30		.620			.673
29		.618			.616
12		.575			.656
20			.826		.702
19			.806		.704
13			.762		.607
24			.759		.599
21			.730		.662
7				.769	.683
6				.726	.731
9				.683	.646
8				.576	.613
10				.568	.615
5				.476	.566
14				.374	.389
	%17.718	%14.523	%14.603	%13.549	
Total Variance			%60.392		

3.1.2. *Confirmatory factor analysis*

The model fit of the four-factor structure obtained due to the exploratory factor analysis of PTUS-YC was examined by single-level confirmatory factor analysis. Figure 2 presents the findings obtained from the single level confirmatory factor analysis. Figure 2 shows that the factor loads of the measurement tool ranged from .59 to .85 for the continuity of use dimension, .62 to .80 for the resistance to control dimension, .77 to .85 for effect on development dimension, and .48 to .87 for the deprivation-escape dimension. In addition, when the t values were examined, the absence of red arrows indicated that all items were significant at the .05 level (Seçer, 2013). Table 2 presents the fit index values obtained from the single level confirmatory factor analysis and the acceptable fit criteria for the examined fit index values.

Table 2. *Fit values and fit index values obtained from CFA.*

Examined Fit Indices	Perfect Fit	Acceptable Fit	Confirmatory Factor Analysis Fit Indices
χ^2/sd	$0 \leq \chi^2/sd \leq 2.00$	$2.00 \leq \chi^2/sd < 5.00$	3.289
RMS	$0 \leq$ RMSEA ≤ 0.05	$0.05 \leq$ RMSEA ≤ 0.08	0.076
NFI	$0.95 \leq$ NFI ≤ 1.00	$0.90 \leq$ NFI ≤ 0.95	0.871
TLI	$0.95 \leq$ TLI ≤ 1.00	$0.90 \leq$ TLI ≤ 0.95	0.890
CFI	$0.97 \leq$ CFI ≤ 1.00	$0.95 \leq$ CFI ≤ 0.97	0.906
IF	$0.95 \leq$ IFI ≤ 1.00	$0.90 \leq$ IFI ≤ 0.95	0.907
RFI	$0.95 \leq$ RFI ≤ 1.00	$0.90 \leq$ RFI ≤ 0.95	0.850
GFI	$0.95 \leq$ GFI ≤ 1.00	$0.90 \leq$ GFI ≤ 0.95	0.852
AGFI	$0.95 \leq$ AGFI ≤ 1.00	$0.85 \leq$ AGFI ≤ 0.95	0.813

Figure 2. *Eigenvalue factor plot for PTUS-YC.*

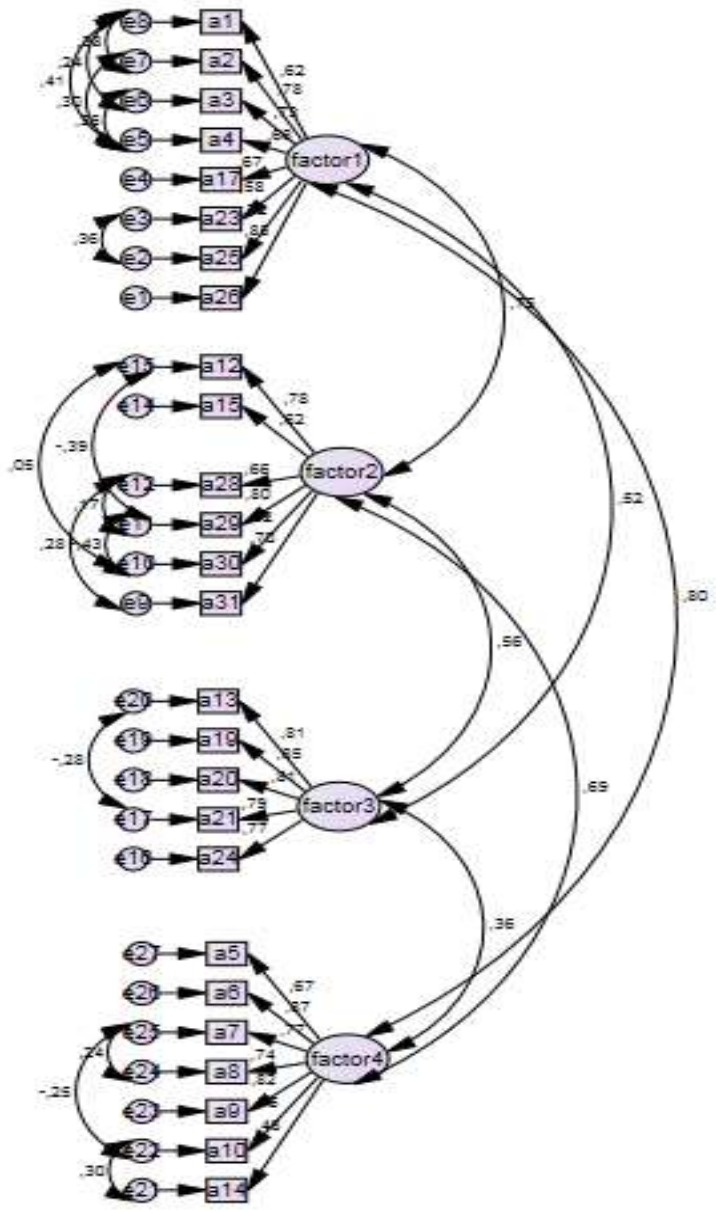


Table 2 demonstrates that the four-factor structure of PTUS-YC, with 26 items and four sub-factors, produced acceptable fit values (χ^2/sd=3.289; p=.00). The fit index values were found as RMSEA=.076, NFI=.871, CFI=.906, IFI=.907, SRMR=.071, GFI=.852 and AGFI=.813. Based on the acceptability of fit indices, the RMSA value should be less than .08; CFI, NFI, and IFI values should be greater than .90; SRMR value should be less than .09 and GFI, and AGFI values should be greater than .85 (Hu & Bentler, 1999; Kline, 2011). However, NFI and AGFI values are greatly affected by sample size (Bentler, 1990; Tabachnick & Fidell, 2007; Yadama & Pandey, 1995). Kline (2015) recommends 20 observations for each estimated parameter in the model. Therefore, smaller sample size of this study could influence NFI and AGFI values of the model. For this reason, it was decided that the relevant indices also indicated fit. Thus, it can be argued that the fit indices of this structural model created in the single-level CFA analysis were at an acceptable level. In addition, in the single-level CFA analysis, modifications were done in accordance with the suggestions between items 1 and 2, 1 and 3, 1 and 4, 2 and 4, 3 and 4, 23 and 25, 7 and 8, 7 and 10, 10 and 14, 13 and 21, 12 and 29, 12 and 30, 28 and 29, 28 and 31 and 29 and 30 and it was observed that after the modifications the model had a better fit.

3.1.3. *Correlation between dimensions*

Table 3 presents the correlation coefficient values obtained as a result of the Pearson correlation analysis performed to determine the relationship between the factors of PTUS-YC and the relationship between the factors and the total factor.

Table 3. *Correlation coefficients between factors.*

Factor	Continuity of Use	Resistance to Control	Effect on Development	Deprivation-Escape	PTUS-YC
Continuity of Use	1	.604**	.458**	.683**	.868**
Resistance to Control		1	.572**	.646**	.832**
Effect on Development			1	.387**	.733**
Deprivation-Escape				1	.829**
PTUS-YC					1

**p<.01

Examination of the correlation coefficient values between the factors demonstrated a significant relationship between the continuity of use, resistance to control, effects on development and deprivation-escape factors and the measurement tool. In addition, examination of the correlation coefficients for the sub-dimensions and the whole measurement tool pointed to a high level of significant relationship (Büyüköztürk, 2018).

3.1.4. *Item-total statistics*

The correlation values between the items of PTUS-YC and the factors in which they were included and the measurement tool were calculated, and total item correlation was used to test whether each item served a common purpose. An independent sample t-test was employed to calculate the difference between the lower 27% group and the upper 27% group to calculate the item discrimination index of the items of PTUS-YC. Table 4 demonstrates the item-total statistics values.

Table 4 shows that item-subscale correlation values varied between .688 and .853 for the continuity of use dimension; item-subscale correlation values varied between .707 and .817 for the resistance to control dimension; item-subscale correlation values f varied between .820 and .873 for the effect on development dimension the item-subscale correlation values varied between .622 and .860 for the deprivation-escape dimension. In addition, item-test correlation

values for the continuity of use dimension were found to be between .574 and .813, item-test correlations for the dimension of resistance to control were found to be between.519 and .711, item-test correlations values for effect on development dimension were found to be between .487 and .656, and the item-test correlation values for the deprivation-escape dimension were found to be between .547 and .728. Furthermore, a significant positive relationship was observed among each item, the factor it belonged to and the whole measurement tool. According to this finding, it can be argued that each item was related to the factor to which it belonged and to the complete the measurement tool and functioned for the same purpose.

Table 4. *Item-total statictics.*

Item No	Factor	Item-Subscale Correlation	Item-Subscale Correlation	Upper/Lower 27% *t*
1	Continuity of Use	.783**	.574**	-97.502**
2	Continuity of Use	.853**	.718**	-97.502**
3	Continuity of Use	.787**	.673**	-75.184**
4	Continuity of Use	.792**	.651**	-87.093**
17	Continuity of Use	.698**	.658**	-76.112**
23	Continuity of Use	.688**	.597**	-97.502**
25	Continuity of Use	.761**	.730**	-53.368**
26	Continuity of Use	.814**	.813**	-112.158**
12	Resistance to Control	.783**	.696**	-22.166**
15	Resistance to Control	.707**	.519**	-11.167**
28	Resistance to Control	.780**	.602**	-18.268**
29	Resistance to Control	.779**	.663**	-24.712**
30	Resistance to Control	.817**	.711**	-39.434**
31	Resistance to Control	.806**	.597**	-20.049**
13	Effect on Development	.834**	.581**	-140.405**
19	Effect on Development	.873**	.656**	-99.464**
20	Effect on Development	.862**	.487**	-93.995**
21	Effect on Development	.820**	.582**	-119.190**
24	Effect on Development	.835**	.571**	-479.000**
5	Deprivation-Escape	.722**	.673**	-41.657**
6	Deprivation-Escape	.860**	.728**	-41.334**
7	Deprivation-Escape	.812**	.618**	-67.127**
8	Deprivation-Escape	.783**	.631**	-47.881**
9	Deprivation-Escape	.840**	.688**	-44.095**
10	Deprivation-Escape	.780**	.677**	-26.107**
14	Deprivation-Escape	.622**	.547**	-36.044**

**$p<.01$

As a result of the independent sample t-test performed between the lower 27% group and the upper 27% group of the items in PTUS-YC, the t-test values were found to be between -53.368 and -112.158 for the continuity of use dimension; between -11.167 and -39.434 for the resistance to control dimension; between -93.995 and -479.000 for effect on development dimension and between -26,107 and -67,127 for the deprivation-escape dimension.

In addition, when the significant differentiation between the groups was examined, a significant difference was observed between the group with everyday problematic technology use and the group with high problematic technology use in both sub-dimensions and in the total scores (p <.001). This finding shows that all PTUS-YC items can distinguish between children with low or high problematic technology use, and this finding demonstrates that the scale has internal validity.

3.1.5. *Proximal validity*

Proximal validity shows that the expressions associated with the variables are related to each other and the factor they form. CR and AVE values were analyzed in order to make evaluations regarding the proximal validity of PTUS-YC. Table 5 provides the CR and AVE values obtained from this analysis.

Table 5. *AVE and CR values of PTUS-YC.*

Dimensions	AVE	CR
Continuity of Use	.43	.86
Resistance to Control	.41	.81
Effect on Development	.60	.88
Deprivation-Escape	.43	.80

According to Table 5, the composite reliability (CR) values were higher than the average variance extracted (AVE) values in all dimensions of PTUS-YC, but the AVE values were not more significant than .50 except for the effect on the development dimension. It is expected that the AVE value should be higher than 0.50, the CR value should be higher than 0.60, and the condition of "CR> AVE> 0.50" should be met to ensure that the tool has convergent validity (Hair et al., 2006). When CR and AVE values were examined in terms of PTUS-YC dimensions, it was observed that while the CR values were within the acceptable range, AVE values were not within the acceptable range except for effect on the development dimension. However, in line with the views of Fornell and Larcker (1981) and Lam (2012) that convergent validity is provided when the CR value is higher than .60, it was seen that PTUS-YC had convergent validity in this study.

3.2. Reliability

Cronbach Alpha and McDonald Omega values were calculated to examine the reliability of PTUS-YC within the scope of this research. Table 6 provides the findings obtained as a result of the reliability analysis.

Table 6. *Internal consistency and split-half reliability analysis coefficients of PTUS-YC.*

Dimensions	Cronbach Alpha	McDonald's Omega
Continuity of Use	.903	.903
Resistance to Control	.876	.880
Effect on Development	.902	.902
Deprivation-Escape	.882	.886
Total	.938	.939

It was observed that all reliability coefficients obtained from Cronbach Alpha and McDonald's Omega analyses for the reliability analysis of PTUS-YC were at a reasonable level for the total scale and sub-dimensions. It can be argued that the broad-scale and its sub-dimensions met the reliability criteria. Because measurement tools are considered reliable when the reliability coefficients are .70 or above, it can be argued that the internal consistency and split-half reliability coefficients of PTUS-YC were sufficient (Can, 2019).

4. DISCUSSION and CONCLUSION

This research developed a measurement tool to determine preschool children's problematic technology use levels. There is always a need for individual diagnostic tools such as scales, tests and inventories for educational research and therapy studies conducted with children. Measurement and evaluation processes are crucial both in obtaining accurate results in scientific

research and in making accurate decisions in the services to be offered to individuals. Therefore, scale development and adaptation studies are valuable. Unless psychological tests are proven to be independent of culture, they must be authentic and developed separately for each country (Tezbaşaran, 1996). Unfortunately, the number of authentically developed measurement tools is limited in Turkey compared to adaptation studies. Considering all these and focusing on the fact that the technology is becoming more and more common among very young users, it was planned to develop a measurement tool to identify preschool children's level of problematic technology use. Based on the results of various analyses, a scale was developed with 26 items and four factors. The scale was titled "Problematic Technology Use Scale for Children" and included the following sub-dimensions: continuity of use, resistance to control, effects on development and deprivation-escape.

Internal consistency and McDonald's Omega coefficients were investigated to determine the reliability of PTUS-YC. While the internal consistency coefficient for the whole PTUS-YC was found as .938, the internal consistency coefficients for the sub-dimensions were as follows: .903 for the continuity of use dimension, .876 for the resistance to control dimension, .902 for effect on development dimension and .882 for the deprivation-escape dimension. While McDonald's Omega coefficient was found to be .939 for the whole PTUS-YC, McDonald's Omega coefficients for the sub-dimensions were as follows: .903 for the continuity of use dimension, .880 for the continuity of use the resistance to control dimension, .902 for effect on development dimension and .886 for the deprivation-escape dimension. It can be argued that the measurement tool is reliable based on the reliability coefficients obtained in the analyses performed to determine the reliability of the PTUS-YC (Can, 2019).

Exploratory factor analysis, confirmatory factor analysis, convergent validity analysis, item discrimination indices and item factor correlation values were examined to determine the validity of PTUS-YC. In the exploratory factor analysis, which was performed first, the measurement tool was collected under four factors named continuity of use, resistance to control, effects on development and deprivation-escape. These four factors were found to explain 60.392% of the total variance. The confirmatory factor analysis performed later to verify this 4-factor structure showed that the model produced good fit values. The independent sample t-test conducted to determine the item discrimination index value of PTUS-YC demonstrated a significant difference between the 27%group with the highest score and the 27% group with the lowest score in all items of the measurement tool. This finding proves that the scale can distinguish children with low problematic technology use from children with high problematic technology use. In addition, when CR and AVE values were examined for the proximal validity of PTUS-YC, it was observed that the CR values were within the acceptable range while AVE values were not within the acceptable range except for effect on the development sub-dimension. However, based on the views of Fornell and Larcker (1981) and Lam (2012) that convergent validity is provided when the CR value is higher than .60, it was seen that the convergent validity criteria were met for PTUS-YC. Finally, the correlation values between the items of PTUS-YC, the factors they were involved in and the whole measurement tool were calculated, and total item correlation was used to test whether each item served a common purpose. Item-subscale correlation values were between .688 and .853 for the continuity of use dimension, item-subscale correlation values were between .707 and .817 for the dimension of resistance to control, item-subscale correlation values were between .820 and .873 for effect on development dimension, and item-subscale correlation values were between .622 and .860 for the deprivation-escape dimension. In addition, item-test correlation values for the continuity of use dimension were found to be between .574 and .813, item-test correlations for the dimension of resistance to control were found to be between.519 and .711, item-test correlations values for effect on development dimension were found to be between .487 and .656, and the item-test correlation values for the deprivation-escape

dimension were found to be between .547 and .728. Furthermore, a significant positive relationship was observed among each item, the factor it belonged to and the whole measurement tool. According to this finding, it can be argued that each item was related to the factor to which it belonged and to the complete the measurement tool and functioned for the same purpose.

The validity and reliability findings obtained from the study (Furuncu & Öztürk, 2020), which was developed to measure the problematic media use of preschool children in Turkey, also show similarities and differences in some respects. In the related study, the internal consistency coefficients of both the long-form and the short form were over .90. This finding shows that the measurement tools in the related research are also reliable, just like our measurement tool. However, while a four-factor structure emerged in our study, a one-dimensional structure emerged in the related study. Considering that problematic technology use is a comprehensive concept, it can be said that the structure in our research could be stronger. In addition, while the one-dimensional structure explains 57.6% of the total variance in the related study, the four-factor structure explains 60.3% of the total variance in our study. When considered in terms of the explained variance, the findings constitute the strength of the measurement tool developed in the research. Finally, concordant validity and incremental validity analyses were performed for the related research's validity study. Evidence regarding the measurement tool's validity related to these analyses has been presented. The fact that these validity analyzes are included in the relevant research, unlike our research, constitutes a different aspect of our research.

Based on the results of this study, PTUS-YC is a valid and reliable measurement tool that can be used with the parents of preschool children to determine their children's problematic technology use level. It can be argued that PTUS-YC is a measurement tool that can meet the need for identifying preschool children's problematic technology use in fields such as preschool education, psychological counseling and guidance and child development, and it can be used in relevant research as well. It is believed that retesting the validity and reliability of the scale in future studies with different sample groups (preschool, primary school, etc.) and implementing and interpreting the scale on groups with different characteristics will significantly contribute to the power of the measurement tool.

Acknowledgments

This study has been supported by Erciyes University Scientific Research Projects Coordination Unit under grant number SBA-2021-10739.

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: Erciyes University/Social and Humanities Ethics Committee, 2020-176.

Authorship Contribution Statement

All authors have equally contributed to all sections of this study.

Orcid

Ahmet Sami Konca https://orcid.org/0000-0002-6423-6608
Önder Baltacı https://orcid.org/0000-0002-9974-8507
Ömer Faruk Akbulut https://orcid.org/0000-0001-5152-8102

REFERENCES

Akboğa, Ö. F., & Gürgan, U. (2019). Gençlerde internet bağımlılığı ile sosyal kaygı bozukluğu arasındaki ilişkinin yaşam doyumu ve bazı değişkenler açısından incelenmesi [The

relationship between internet addiction and social anxiety disorder in young people life satisfaction and some variables in terms of examination]. *Avrasya Sosyal ve Ekonomi Araştırmaları Dergisi*, *6*(6), 443–464. https://dergipark.org.tr/en/download/article-file/768703

Alonso, C., & Romero, E. (2017). Problematic technology use in a clinical sample of children and adolescents. Personality and behavioral problems associated. *Actas Espanolas de Psiquiatria*, *45*(2), 62–70. https://pubmed.ncbi.nlm.nih.gov/28353291/

American Academy of Pediatrics. (2016). *American Academy of Pediatrics Announces New Recommendations for Children's Media Use*. https://www.pathwaypeds.com/american-academy-of-pediatrics-announces-new-recommendations-for-childrens-media-use/

Anderson, D.R., & Pempek, T.A. (2005). Television and very young children. *American Behavioral Scientist*, *48*(5), 505–522. https://doi.org/10.1177%2F0002764204271506

Anlayışlı, C., & Bulut Serin, N. (2019). Lise öğrencilerinde internet bağımlılığı ve depresyonun cinsiyet, akademik başarı ve internete giriş süreleri açısından incelenmesi [A study on internet addiction and depression among high school students due to gender, academic success and internet usage duration]. *Folklor/Edebiyat*, *25*(97), 753–767. https://doi.org/10.22559/folklor.977

Anlı, G., & Taş, İ. (2018). Ergenler için oyun bağımlılığı ölçeği kısa formunun geçerlik ve güvenirlik çalışması [The validity and reliability of the game addiction scale for adolescents-short form]. *Electronic Turkish Studies*, *13*(11), 189203. http://dx.doi.org/10.7827/TurkishStudies.

Arıcak, O.T., Dinç, M., Yay, M., & Griffiths, M. D. (2019). İnternet oyun oynama bozukluğu ölçeği kısa formunun (İOOBÖ9-KF) Türkçeye uyarlanması: Geçerlik ve güvenirlik çalışması [Adapting the short form of the internet gaming disorder scale into Turkish: Validity and reliability]. *Addicta: The Turkish Journal on Addictions*, *6*(1), 615-636. http://dx.doi.org/10.15805/addicta.2018.5.4.0027

Avşaroğlu, S., & Akbulut, Ö.F. (2020). Sağlıklı aile yapısı açısından bir risk faktörü: İnternet bağımlılığı [A risk factor in terms of healthy family structure: Internet addiction]. *International Social Sciences Studies Journal*, *6*(65), 2879-2902. http://dx.doi.org/10.26449/sssj.2456

Balantekin, Y. (2009). *10-14 yaş arası çocuklarda televizyon bağımlılığı üzerine bir araştırma [A research ontelevision dependency of children between ages of 10-14]* [Unpublished master's thesis]. Uludag University.

Balcı, A. (2020). *Sosyal bilimlerde araştırma: Yöntem, teknik ve ilkeler [Research in social sciences: Methods, techniques and principles]*. Pegem Akademi.

Bayraktar, F. (2001). *İnternet kullanımının ergen gelişimindeki rolü [The Role of internet usage in the development of adolescents]* [Unpublished master's thesis]. Ege University.

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238-246. https://psycnet.apa.org/doi/10.1037/0033-2909.107.2.238

Bruni, O., Sette, S., Fontanesi, L., Baiocco, R., Laghi, F., & Baumgartner, E. (2015). Technology use and sleep quality in preadolescence and adolescence. *Journal of Clinical Sleep Medicine*, *11*(12), 1433–1441. https://doi.org/10.5664/jcsm.5282

Bryman, A., & Cramer, D. (2011). *Quantitative data analysis with IBM SPSS 17, 18 and 19.* Routledge.

Büyüköztürk, Ş. (2018). *Sosyal bilimler için veri analizi el kitabı [Manual of data analysis for social sciences]*. Pegem Akademi.

Çakıroğlu, S., & Soylu, N. (2019). Adaptation of internet gaming disorder questionnaire to Turkish: Reliability and validity study. *Turkish Journal of Psychiatr*, *30*(2), 130–136. https://doi.org/10.5080/u23537

Calamaro, C.J., Yang, K., Ratcliffe, S., & Chasens, E.R. (2012). Wired at a young age: the effect of caffeine and technology on sleep duration and body mass index in school-aged children. *Journal of Pediatric Health Care*, *26*(4), 276-282. https://doi.org/10.1016/j.pedhc.2010.12.002

Can, A. (2019). *SPSS file bilimsel araştırma sürecinde nicel veri analizi [Quantitative data analysis in the scientific research process with SPSS]*. Pegem Akademi.

Canan, F., Ataoglu, A., Nichols, L.A., Yildirim, T., & Ozturk, O. (2010). Evaluation of psychometric properties of the internet addiction scale in a sample of Turkish high school students. *Cyberpsychology, Behavior, and Social Networking*, *13*(3), 317–320. https://doi.org/10.1089/cyber.2009.0160

Canoğulları-Ayazseven, Ö., & Cenkseven-Önder, F. (2019). Genelleştirilmiş problemli internet kullanım ölçeği 2'nin Türkçe'ye uyarlama çalışması [Turkish adaptation study of the generalized problematic internet use scale 2]. *OPUS Uluslararası Toplum Araştırmaları Dergisi*, *11*(18), 1540–1565. https://doi.org/10.26466/opus.529016

Caplan, S.E. (2010). Theory and measurement of generalized problematic Internet use: A two-step approach. *Computers in Human Behavior*, *26*(5), 1089-1097. https://doi.org/10.1016/j.chb.2010.03.012

Cespedes, E.M., Gillman, M.W., Kleinman, K., Rifas-Shiman, S.L., Redline, S., & Taveras, E. M. (2014). Television viewing, bedroom television, and sleep duration from infancy to mid childhood. *Journal Ofthe American Academy of Pediatrics*, *133*(5), 1163–1171. https://doi.org/10.1542/peds.2013-3998

Ceyhan, A.A., & Ceyhan, E. (2014). The validity and reliability study of problematic internet use scale for adolescents. *Bağımlılık Dergisi*, *15*(2), 56-64. https://www.researchgate.net/publication/234628682_The_Validity_and_Reliability_of_the_Problematic_Internet_Usage_Scale

Chi, X., Hong, X., & Chen, X. (2020). Profiles and sociodemographic correlates of internet addiction in early adolescents in Southern China. *Addictive Behaviors*, *106*, 1–7. https://doi.org/10.1016/j.addbeh.2020.106385

Chonchaiya, W., & Pruksananonda, C. (2008). Television viewing associates with delayed language development. *ActaPaediatrica*, *97*(7), 977–982. https://doi.org/10.1111/j.1651-2227.2008.00831.x

Christakis, D.A., Zimmerman, F.J., DiGiuseppe, D.L., & McCarty, C.A. (2004). Early television exposure and subsequent attentional problems in children. *Pediatrics*, *113*(4), 708–713. https://doi.org/10.1542/peds.113.4.708

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2018). *Sosyal bilimler için çok değişkenli istatistik SPSS ve LISREL uygulamaları [Multivariate statistics SPSS and LISREL applications for social sciences]*. Pegem Akademi.

Cox, R., Skouteris, H., Rutherford, L., Fuller-Tyszkiewicz, M., & Hardy, L.L. (2012). Television viewing, television content, food intake, physical activity and body mass index: A cross-sectional study of preschool children aged 2-6 years. *Health Promotion Journal of Australia*, *23*(1), 58–62. https://doi.org/10.1071/HE12058

De Jong, E., Visscher, T., HiraSing, R., Heymans, M., Seidell, J., & Renders, C. (2013). Association between TV viewing, computer use and overweight, determinants and competing activities of screen time in 4-to 13-year-old children. *International Journal of Obesity*, *37*(1), 47–53. https://doi.org/10.1038/ijo.2011.244

DeLoache, J.S., Chiong, C., Sherman, K., Islam, N., Vanderborght, M., Troseth, G. L., ..., & O'Doherty, K. (2010). Do babies learn from baby media?. *Psychological Science*, *21*(11), 1570–1574. https://doi.org/10.1177%2F0956797610384145

Derin, S., & Bilge, F. (2016). Ergenlerde internet bağımlılığı ve öznel iyi oluş düzeyi [Internet addiction and the level of subjective well-being in adolescents]. *Türk Psikolojik Danışma ve Rehberlik Dergisi*, *6*(46), 35–51.

Ding, J. E., Liu, W., Wang, X., Lan, Y., Hu, D., Xu, Y., ... & Fu, H. (2019). Development of a smartphone overuse classification scale. *Addiction Research & Theory, 27*(2), 150-155. https://doi.org/10.1080/16066359.2018.1474204

Domoff, S. E., Harrison, K., Gearhardt, A. N., Gentile, D. A., Lumeng, J. C., & Miller, A. L. (2019). Development and validation of the Problematic Media Use Measure: A parent-report measure of screen media "addiction" in children. *Psychology of Popular Media Culture, 8*(1), 2–11. https://doi.org/10.1037/ppm0000163

Erkuş, A. (2012). *Psikolojide ölçme ve ölçek geliştirme-1 [Measurement and scale development in psychology-1]*. Pegem Akademi.

Eroğlu, A. (2008). Faktör analizi [Factor analysis]. In Ş. Kalaycı (Ed.), *SPSS uygulamalı çok değişkenli istatistik teknikleri [SPSS applied multivariate statistical techniques]*. Asil Yayınevi.

Eşgi, N. (2014). Aile-çocuk internet bağımlılık ölçeğinin Türkçeye uyarlanması: Geçerlik ve güvenirlik çalışması [The adaptation of parent-child internet addiction scale into Turkish: The study of validity and reliability]. *Kastamonu Eğitim Dergisi*, *22*(2), 807–839. https://dergipark.org.tr/en/download/article-file/209938

Fidan, H. (2016). Mobil bağımlılık ölçeğinin geliştirilmesi ve geçerliliği: Bileşenler modeli yaklaşımı [Development and validation of the mobile addiction scale: The components model approach]. *Addicta: The Turkish Journal on Addictions*, *3*, 452-469. http://dx.doi.org/10.15805/addicta.2016.3.0118

Fırat, N., & Balcı-Çelik, S. (2017). The adaptation of mobile phone addiction scale into Turkish: Validity and reliability study. *Journal of Human Sciences*, *14*(3), 2875–2887. http://dx.doi.org/10.14687/jhs.v14i3.1592

Foerster, M., Roser, K., Schoeni, A., & Röösli, M. (2015). Problematic mobile phone use in adolescents: derivation of a short scale MPPUS-10. *International Journal of Public Health, 60*(2), 277-286. https://doi.org/10.1007/s00038-015-0660-4

Fornell, C., & Larcker, D.F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of marketing research*, *18*(1), 39-50. https://doi.org/10.1177/002224378101800104

Fuller, C., Lehman, E., Hicks, S., & Novick, M.B. (2017). Bedtime use of technology and associated sleep problems in children. *Global Pediatric Health*, *4*, 1-8. https://doi.org/10.1177/2333794X17736972

Furuncu, C., & Öztürk, E. (2020). Problemli medya kullanım ölçeği Türkçe formunun geçerlik güvenirlik çalışması: Çocuklarda ekran bağımlılığı ölçeği ebeveyn formu [Validity and reliability study of Turkish version of problematic media use measure: A parent report measure of screen addiction in children]. *Erken Çocukluk Çalışmaları Dergisi*, *4*(3), 535–566. https://doi.org/10.24130/eccd-jecs.1967202043237

Genc, Z. (2014). Parents' perceptions about the mobile technology use of preschool-aged children. *Procedia-Social and Behavioral Sciences*, *146*, 55-60. https://doi.org/10.1016/j.sbspro.2014.08.086

Geng, J., Han, L., Gao, F., Jou, M., & Huang, C-C. (2018). Internet addiction and procrastination among Chinese young adults: A moderated mediation model. *Computers in Human Behavior*, *84*, 320–333. https://doi.org/10.1016/j.chb.2018.03.013

Gunuc, S., & Dogan, A. (2013). The relationships between Turkish adolescents' Internet addiction, their perceived social support and family activities. *Computers in Human Behavior*, *29*(6), 2197–2207. https://doi.org/10.1016/j.chb.2013.04.011

Günüç, S., & Kayri, M. (2010). The profile of internet dependency in Turkey and development of internet addiction scale: Study of validity and reliability. *Hacettepe University Journal of Education*, *39*, 220–232.

Güzeller, C.O., & Coşguner, T. (2012). Development of a Problematic Mobile Phone Use Scale for Turkish adolescents. *Cyberpsychology, Behavior, and Social Networking*, *15*(4), 205–211. https://doi.org/10.1089/cyber.2011.0210

Hair, J., Black, W., Babin, B., Anderson, R., & Tatham, R. (2006). *Multivariate Data Analysis*. Prentice-Hall.

Harris, B., Regan, T., Schueler, J., & Fields, S.A. (2020). Problematic mobile phone and smartphone use scales: A systematic review. *Frontiers in Psychology, 11*, 672. https://doi.org/10.3389/fpsyg.2020.00672

Hazar, Z., & Hazar, M. (2017). Digital game addiction scale for children. *Journal of Human Sciences*, *14*(1), 203–216. https://www.guvenliweb.org.tr/dosya/jm4Ki.pdf

Horzum, M.B., Ayas, T., & Çakır-Balta, Ö. (2008). Çocuklar için bilgisayar oyun bağımlılığı ölçeği [Computer game addiction scale for children]. *Türk Psikolojik Danışma ve Rehberlik Dergisi*, *3*(30), 76-88. https://app.trdizin.gov.tr/publication/paper/detail/T0RVMU5UazU

Howie, E.K., Coenen, P., Campbell, A.C., Ranelli, S., & Straker, L.M. (2017). Head, trunk and arm posture amplitude and variation, muscle activity, sedentariness and physical activity of 3 to 5 year-old children during tablet computer use compared to television watching and toy play. *Applied Ergonomics*, *65*, 41-50. https://doi.org/10.1016/j.apergo.2017.05.011

Hu, L.T., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Ilgaz, H. (2015). Ergenler için oyun bağımlılığı ölçeğinin Türkçeye uyarlama çalışması [Adaptation of game addiction scale for adolescents into Turkish] *İlköğretim Online*, *14*(3), 874–884. http://dx.doi.org/10.17051/io.2015.75608

Irmak, A.Y., & Erdoğan, S. (2015). Validity and reliability of the Turkish version of the digital game addiction scale. *Anatolian Journal of Psychiatry*, *16*, 10-18. http://doi.org/10.5455/apd.170337

Johansson, A.E., Petrisko, M.A., & Chasens, E.R. (2016). Adolescent sleep and the impact of technology use before sleep on daytime function. *Journal of Pediatric Nursing*, *31*(5), 498–504. https://doi.org/10.1016/j.pedn.2016.04.004

Kabadayı, F. (2020). Psychometric properties of Turkish cyberpsychology scales. *Turkish Psychological Counseling and Guidance Journal*, *10*(58), 385–411.

Karagöz, Y. (2016). *SPSS ve AMOS 23 uygulamalı istatistiksel analizler [Applied statistical analyses with SPSS and AMOS 23]*. Nobel Yayıncılık.

Kawabe, K., Horiuchi, F., Miyama, T., Jogamoto, T., Aibara, K., Ishii, E., & Ueno, S. (2019). Internet addiction and attention-deficit / hyperactivity disorder symptoms in adolescents with autism spectrum disorder. *Research in Developmental Disabilities*, *89*, 22–28. https://doi.org/10.1016/j.ridd.2019.03.002

Kaya, A.B. (2013). *Çevrimiçi oyun bağımlılığı ölçeğinin geliştirilmesi: Geçerlik ve güvenirlik çalışması [Development of online game addiction scale: A scale validity and reliability study]* [Unpublished master's thesis]. Tokat Gaziosmanpasa University.

Kline, R.B. (2011). *An easy guide to factor analysis*. Guilford Publications.

Kline, R.B. (2015). *Principles and practice of structural equation modeling*. Guilford Publications.

Kutlu, M., Savcı, M., Demir, Y., & Aysan, F. (2016). Young internet bağımlılığı testi kısa formunun Türkçe uyarlaması: Üniversite öğrencileri ve ergenlerde geçerlilik ve

güvenilirlik çalışması [Turkish adaptation of Young's Internet Addiction Test-Short Form: a reliability and validity study on university students andadolescents]. *Anadolu Psikiyatri Dergisi*, *17*(1), 69–76. https://doi.org/10.5455/apd.190501

Lam, L.W. (2012). Impact of competitiveness on salespeople's commitment and performance. *Journal of Business Research*, *65*(9), 1328-1334. https://doi.org/10.1016/j.jbusres.2011.10.026

Lam, L.T., Peng, Z. W., Mai, J.C., & Jing, J. (2009). Factors associated with Internet addiction among adolescents. *CyberPsychology & Behavior*, *12*(5), 551-555. https://doi.org/10.1089/cpb.2009.0036

Lee, S.J., & Chae, Y.G. (2007). Children's internet use in a family context: Influence on family relationships and parental mediation. *CyberPsychology & Behavior*, *10*(5), 640–644. https://doi.org/10.1089/cpb.2007.9975

McCarrick, K., & Li, X. (2007). Buried treasure: The impact of computer use on young children's social, cognitive, language development and motivation. *AACE Journal*, *15*(1), 73–95. https://www.learntechlib.org/p/19982/

Mei, X., Zhou, Q., Li, X., Jin, P., Wang, X., & Hu, Z. (2018). Sleep problems in excessive technology use among adolescent: a systemic review and meta-analysis. *Sleep Science and Practice*, *2*(9), 1–10. https://doi.org/10.1186/s41606-018-0028-9

Mitchell, J., Pate, R., Beets, M., & Nader, P. (2013). Time spent in sedentary behavior and changes in childhood BMI: A longitudinal study from ages 9 to 15 years. *International Journal of Obesity*, *37*(1), 54–60. https://doi.org/10.1038/ijo.2012.41

Mıhçı, P., & Çakmak, E.K. (2017). Öğrenci siber sağlık ölçekleri geliştirme çalışması [A study on student cyberwellness scales development]. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi*, *37*(2), 457–491.

Mustafaoğlu, R., Zirek, E., Yasacı, Z., & Özdinçler, A.R. (2018). Dijital teknoloji kullanımının çocukların gelişimi ve sağlığı üzerine olumsuz etkileri [The negative effects of digital technology usage on children's development and health]. *Addicta: The Turkish Journal on Addictions*, *5*(2), 1–21. http://dx.doi.org/10.15805/addicta.2018.5.2.0051

National Association for the Education of Young Children. (2012). *Technology and Interactive Media as Tools in Early Childhood Programs Serving Children from Birth through Age 8*. https://www.naeyc.org/sites/default/files/globally-shared/downloads/PDFs/resources/topics/PS_technology_WEB.pdf

National Center for Education Statistics. (2019). *Student Access to Digital Learning Resources Outside of the Classroom (NCES 2017-098)*. https://nces.ed.gov/fastfacts/display.asp?id=46#:~:text=Overall

Ögel, K., Karadağ, F., Satgan, D., & Koç, C. (2015). Bağımlılık profil indeksi internet bağımlılığı formunun (BAPİNT) geliştirilmesi: Geçerlik ve güvenilirliği [Development of the addiction profile index internet addiction form (APIINT): Validity and reliability. *Düşünen Adam*, *28*, 337–343. http://dx.doi.org/10.5350/DAJPN2015280405

Orben, A., & Przybylski, A.K. (2019). The association between adolescent well-being and digital technology use. *Nature Human Behaviour*, *3*(2), 173-182. https://doi.org/10.1038/s41562-018-0506-1

Pagani, L.S., Fitzpatrick, C., Barnett, T.A., & Dubow, E. (2010). Prospective associations between early childhood television exposure and academic, psychosocial, and physical well-being by middle childhood. *Archives of Pediatrics & Adolescent Medicine*, *164*(5), 425–431. https://doi.org/10.1001/archpediatrics.2010.50

Pallant, J. (2017). *Spss kullanma kılavuzu: Spss ile adım adım veri analizi [Spss user guide: step-by-step data analysis with Spss]* (S. Balcı & B. Ahi (eds.)). Anı Yayıncılık.

Pancani, L., Preti, E., & Riva, P. (2020). The psychology of smartphone: The development of the smartphone impact scale (SIS). *Assessment, 27*(6), 1176-1197. https://doi.org/10.1177%2F1073191119831788

Peng, W., Li, D., Li, D., Jia, J., Wang, Y., & Sun, W. (2019). School disconnectedness and Adolescent Internet Addiction: Mediation by self-esteem and moderation by emotional intelligence. *Computers in Human Behavior*, *98*, 111-121. https://doi.org/10.1016/j.chb.2019.04.011

Pew Research Center. (2020). *Parenting children in the age of screens*. https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2020/07/PI_2020.07.28_kids-and-screens_FINAL.pdf

Plowman, L., McPake, J., & Stephen, C. (2010). The technologization of childhood? Young children and technology in the home. *Children & Society*, *24*(1), 63–74. https://doi.org/10.1111/j.1099-0860.2008.00180.x

Rideout, V., Saphir, M., Tsang, V., & Bozdech, B. (2011). *Zero to eight children's media use in America*. Common Sense Media.

Rogow, F. (2007). *Two View or Not Two View: A Review of the Research Literature on the Advisability of Television Viewing for Infants and Toddlers*. https://a.s.kqed.net/pdf/education/earlylearning/media-symposium/tv-under-two-rogow.pdf?trackurl=true

Rosen, L.D., Lim, A.F., Felt, J., Carrier, L.M., Cheever, N.A., Lara-Ruiz, J.M., ..., & Rokkum, J. (2014). Media and technology use predicts ill-being among children, preteens and teenagers independent of the negative health impacts of exercise and eating habits. *Computers in Human Behavior*, *35*, 364–375. https://doi.org/10.1016/j.chb.2014.01.036

Şar, A.H., Ayas, T., & Horzum, M.B. (2015). Akıllı telefon bağımlılığı ölçeği geliştirme: Geçerlik ve güvenirlik çalışması [Developing the smart phone addiction scale and its validity and reliability study]. *Online Journal Of Technology Addiction & Cyberbullying*, *2*(1), 1–17. https://dergipark.org.tr/en/download/article-file/290287

Savcı, M., & Aysan, F. (2016). Bağlanma stilleri, akran ilişkileri ve duyguların internet bağımlılığını yordamadaki katkıları [The role of attachment styles, peer relations, and affections in predicting internet addiction]. *Addicta: The Turkish Journal on Addictions*, *3*(3), 401–432. http://dx.doi.org/10.15805/addicta.2016.3.0028

Schumacker, R.E., & Lomax, R.G. (2004). *A beginner's guide to structural equation modeling*. Lawrence Erlbaum Associates Publishers.

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

Securelist. (2020). *Kids on the Web in 2020*. https://securelist.com/children-report-2020/97191/

Soldatova, G.U., & Teslavskaia, O.I. (2017). Videogames, academic performance and attention problems: practices and results of foreign empirical studies of children and adolescents. *Journal of Modern Foreign Psychology*, *6*(4), 21-28. https://doi.org/10.17759/jmfp.2017060402

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

Taş, İ. (2017). The social media addiction scale (SF) for adolescents: A study of validity and reliability. *Online Journal of Technology Addiction and Cyberbullying*, *4*(1), 27–40. https://dergipark.org.tr/tr/download/article-file/320071

Taş, İ. (2019). Internet addiction scale for adolescents: Validity and reliability study. *Journal of Kırşehir Education Faculty*, *20*(2), 875-905. https://doi.org/10.29299/kefad.2019.20.02.011

Tezbaşaran, A.A. (1996). *Likert tipi ölçek geliştirme kılavuzu [Likert type scale development guide]*. TPD Yayınları.

Türkiye Yeşilay Cemiyeti (2021). *Türkiye bağımlılıkla mücadele eğitim programı [Turkey struggles with addiction training program]*. Retrieved January 22, 2021, from https://www.tbm.org.tr/

Ünsal, A., & Ulutaş, İ. (2019). Bilgisayar oyun bağımlılığı ölçeğinin okul öncesi dönem çocuklarına uyarlanması [Psychometric properties of computer game addiction scale for preschool children]. *Aksaray Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *3*(2), 324–342. http://aseddergi.aksaray.edu.tr/en/download/article-file/907660

White House Task Force on Childhood Obesity Report to the President. (2010). *Solving the Problem of Childhood Obesity Within a Generation*. https://letsmove.obamawhitehouse.archives.gov/sites/letsmove.gov/files/TaskForce_on_Childhood_Obesity_May2010_FullReport.pdf

Wu, C.S.T., Wong, H.T., Yu, K.F., Fok, K.W., Yeung, S.M., Lam, C.H., & Liu, K.M. (2016). Parenting approaches, family functionality, and internet addiction among Hong Kong adolescents. *BMC Pediatrics*, *16*(1), 1–10. https://doi.org/10.1186/s12887-016-0666-y

Yadama, G.N., & Pandey, S. (1995). Effect of sample size on goodness-fit of-fit indices in structural equation models. *Journal of Social Service Research*, *20*(3-4), 49-70. https://doi.org/10.1300/J079v20n03_03

Yavuzer, H. (2019). *Çocuk psikolojisi [Child psychology]*. Remzi Kitabevi

Ybarra, M.L., Mitchell, K.J., & Korchmaros, J.D. (2011). National trends in exposure to and experiences of violence on the Internet among children. *Pediatrics*, *128*(6), e1376–e1386. https://doi.org/10.1542/peds.2011-0118

Yılmaz, E., Griffiths, M.D., & Kan, A. (2017). Development and validation of videogame addiction scale for children (VASC). *International Journal of Mental Health and Addiction*, *15*(4), 869–882. https://doi.org/10.1007/s11469-017-9766-7

Young, K.S. (2011). CBT-IA: The first treatment model for internet addiction. *Journal of Cognitive Psychotherapy*, *25*(4), 304–312. https://doi.org/10.1891/0889-8391.25.4.304

Zhang, Y., Qin, X., & Ren, P. (2018). Adolescents' academic engagement mediates the association between Internet addiction and academic achievement: The moderating effect of classroom achievement norm. *Computers in Human Behavior*, *89*, 299–307. https://doi.org/10.1016/j.chb.2018.08.018

Zorbaz, O., & Tuzgöl Dost, M. (2014). Lise öğrencilerinin problemli internet kullanımının cinsiyet, sosyal kaygı ve akran ilişkileri açısından incelenmesi [Examination of problematic internet use of high school student in terms of gender, social anxiety and peer relations]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *29*(1), 298–310. https://dergipark.org.tr/en/download/article-file/87094

APPENDIX

Problematic Technology Use Scale for Young Children (PTUS-YC)

Item No	Please rank the following items considering your preschool-age child's frequency of using technological tools (computer, tablet, smartphone, television, etc.). Please check only one option for each item.	Completely Disagree	Somewhat Disagree	Undecided	Somewhat Agree	Completely Agree
1	My child spends an average of more than 1.5 hours a day with technological tools.	(1)	(2)	(3)	(4)	(5)
2	My child exceeds the time limit we have set for technology use.	(1)	(2)	(3)	(4)	(5)
3	My child often expresses a desire to spend time with technological tools.	(1)	(2)	(3)	(4)	(5)
4	There is a significant time increase between my child's early and present technology use.	(1)	(2)	(3)	(4)	(5)
5	My child starts to spend time with technological tools without fulfilling daily responsibilities.	(1)	(2)	(3)	(4)	(5)
6	My child experiences negative emotions when he/she is not spending time with technological tools.	(1)	(2)	(3)	(4)	(5)
7	My child experiences positive emotions when he/she starts to spend time with technological tools.	(1)	(2)	(3)	(4)	(5)
8	My child relaxes by spending time with technological tools when he/she feels sad.	(1)	(2)	(3)	(4)	(5)
9	My child thinks about technological tools even when he/she is not spending time with them.	(1)	(2)	(3)	(4)	(5)
10	My child prefers to spend time with technological tools instead of spending time with us or with his/her friends.	(1)	(2)	(3)	(4)	(5)
11	My child spends time on technological tools by playing games or watching movies that are not suitable for his/her age.	(1)	(2)	(3)	(4)	(5)
12	My child's technology use makes him lonely.	(1)	(2)	(3)	(4)	(5)
13	My child prefers playing games on technological tools to playing games in real life.	(1)	(2)	(3)	(4)	(5)
14	My child does not want to go to school because he/she wants to spend his/her time with technological tools.	(1)	(2)	(3)	(4)	(5)
15	My child spends time alone with technological tools	(1)	(2)	(3)	(4)	(5)
16	My child's use of technological tools negatively affects his/her interaction with his/her environment.	(1)	(2)	(3)	(4)	(5)
17	My child's use of technological tools causes problems in his/her language development.	(1)	(2)	(3)	(4)	(5)
18	My child's use of technological tools has decreased the duration of his/her sleep.	(1)	(2)	(3)	(4)	(5)
19	My child eats/wants to eat while spending time on technological tools.	(1)	(2)	(3)	(4)	(5)
20	My child's use of technological tools makes him/her sedentary.	(1)	(2)	(3)	(4)	(5)
21	My child spends time with technological tools just before going to sleep.	(1)	(2)	(3)	(4)	(5)
22	I have disagreements with my child about the duration of his/her technology use.	(1)	(2)	(3)	(4)	(5)
23	My child does not tell us or lies about what he/she is doing while using technological tools.	(1)	(2)	(3)	(4)	(5)
24	My child tries to use technological tools secretly, although we limit his/her use of technology.	(1)	(2)	(3)	(4)	(5)
25	My child is annoyed when we try to communicate with him/her while spending time with technological tools.	(1)	(2)	(3)	(4)	(5)
26	My child does not allow us to track his/her technology use.	(1)	(2)	(3)	(4)	(5)

Continuity of Use: 1-2-3-4-15-19-21-22
Resistance to Control: 11-14-23-24-25-26
Effect on Development: 12-16-17-18-20
Deprivation-Escape: 5-6-7-8-9-10-13

Turkish version of the scale:

Çocuklar İçin Problemli Teknoloji Kullanımı Ölçeği

Madde No	Aşağıda yer alan ölçekteki maddelerin okul öncesi düzeyinde eğitim gören çocuğunuzun teknolojik araçları (bilgisayar, tablet, akıllı telefon, televizyon vb.) kullanım sıklıklarını düşünerek değerlendiriniz. Lütfen her madde için yalnızca bir seçeneği işaretleyiniz.	Hiç Katılmıyorum	Kısmen Katılmıyorum	Kararsızım	Kısmen Katılıyorum	Tamamen Katılıyorum
1	Çocuğum gün içerisinde ortalama 1,5 saatten fazla teknolojik araçlarla vakit geçirir.	(1)	(2)	(3)	(4)	(5)
2	Çocuğum teknoloji kullanımı konusunda belirlediğimiz süre sınırını aşar.	(1)	(2)	(3)	(4)	(5)
3	Çocuğum teknolojik araçlarla vakit geçirme isteğini sıklıkla dile getirir.	(1)	(2)	(3)	(4)	(5)
4	Çocuğumun teknolojiyi ilk zamanlardaki kullanım süresi ile şimdiki kullanım süresi arasında kayda değer bir artış vardır.	(1)	(2)	(3)	(4)	(5)
5	Çocuğum gündelik sorumluluklarını yerine getirmeden teknolojik araçlarla vakit geçirmeye başlar.	(1)	(2)	(3)	(4)	(5)
6	Çocuğum teknolojik araçlarla vakit geçirmediği zamanlarda olumsuz duygular hisseder.	(1)	(2)	(3)	(4)	(5)
7	Çocuğum teknolojik araçlarla vakit geçirmeye başladığında olumlu duygular hisseder.	(1)	(2)	(3)	(4)	(5)
8	Çocuğum kendisini üzgün hissettiğinde teknolojik araçlarla vakit geçirerek rahatlar.	(1)	(2)	(3)	(4)	(5)
9	Çocuğum vaktini teknolojik araçlarla geçirmediği zamanlarda dahi teknolojik araçları düşünmektedir.	(1)	(2)	(3)	(4)	(5)
10	Çocuğum bizimle veya arkadaşlarıyla beraber vakit geçirmek yerine teknolojik araçlarla vakit geçirmeyi tercih eder.	(1)	(2)	(3)	(4)	(5)
11	Çocuğum teknolojik araçlar üzerinden yaşına uygun olmayan oyunlar/filmler ile vakit geçirir.	(1)	(2)	(3)	(4)	(5)
12	Çocuğumun teknoloji kullanımı onun yalnızlaşmasına neden olur.	(1)	(2)	(3)	(4)	(5)
13	Çocuğum teknolojik araçlar üzerinden oyun oynamayı gerçek yaşamda oyun oynamaya tercih eder.	(1)	(2)	(3)	(4)	(5)
14	Çocuğum vaktini teknolojik araçlarla geçirmek istediği için okula gitmek istemez.	(1)	(2)	(3)	(4)	(5)
15	Çocuğum teknolojik araçlarla tek başına vakit geçirir.	(1)	(2)	(3)	(4)	(5)
16	Çocuğumun teknolojik araçları kullanması, çevresiyle iletişimini olumsuz olarak etkiler.	(1)	(2)	(3)	(4)	(5)
17	Çocuğumun teknolojik araçları kullanması, dil gelişiminde problemlere neden olur.	(1)	(2)	(3)	(4)	(5)
18	Çocuğumun teknolojik araçları kullanması uyku süresinin azalmasına neden olur.	(1)	(2)	(3)	(4)	(5)
19	Çocuğum yemeklerini teknolojik araçlarla vakit geçirirken yer/yemek ister.	(1)	(2)	(3)	(4)	(5)
20	Çocuğumun teknolojik araçları kullanması hareketsiz kalmasına neden olur.	(1)	(2)	(3)	(4)	(5)
21	Çocuğum uyumadan hemen önce teknolojik araçlarla vakit geçirir.	(1)	(2)	(3)	(4)	(5)
22	Çocuğum ile teknoloji kullanımı süresi konusunda anlaşmazlıklar yaşarım.	(1)	(2)	(3)	(4)	(5)
23	Çocuğum teknolojik araçlar kullanırken yaptıkları hakkında bize bilgi vermez veya yalan söyler.	(1)	(2)	(3)	(4)	(5)
24	Çocuğumun teknoloji kullanımını sınırlandırdığımızda bile gizlice kullanmaya çalışır.	(1)	(2)	(3)	(4)	(5)
25	Çocuğum teknolojik araçlarla vakit geçirirken kendisiyle iletişim kurulmasından rahatsız olur.	(1)	(2)	(3)	(4)	(5)
26	Çocuğum kendisinin teknoloji kullanımının takip edilmesine izin vermez.	(1)	(2)	(3)	(4)	(5)

Kullanım Sürekliliği: 1-2-3-4-15-19-21-22
Kontrole Karşı Direnç: 11-14-23-24-25-26
Gelişime Etki: 12-16-17-18-20
Yoksunluk-Kaçış: 5-6-7-8-9-10-13

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 2, 290–318
https://doi.org/10.21449/ijate.946743
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

The Methodological Quality of Experimental STEM Education Articles Published in Scholarly Journals from 2014 to 2020

Ramazan Avcu[1,*], **Seher Avcu**[2]

[1,2]Aksaray University, Faculty of Education, Department of Mathematics and Science Education, Aksaray, Türkiye

ARTICLE HISTORY
Received: June 02, 2021
Revised: Feb. 08, 2022
Accepted: Mar. 15, 2022

Keywords:
STEM education,
Methodological quality,
Experimental research,
Scholarly journal articles.

Abstract: Experimental studies have a considerable impact on the educational policies and practices of many countries. In Turkey, policymakers are planning to initiate a STEM education reform in K-12 schools based on experimental studies. However, the methodological flaws in these studies may lead to biased outcomes and may mislead the STEM education community. Despite the importance of methodological quality, to the best of our knowledge, there are no studies that investigate the methodological quality of experimental STEM education articles published in scholarly journals. Therefore, in this study, we conducted a methodological review to examine the methodological quality of experimental STEM education articles published in refereed Turkish journals from 2014 to 2020. During the targeted period, we located 68 articles. We analyzed these articles by developing a coding framework. We found that the selected articles suffer seriously from various methodological flaws. We discuss the findings in light of the literature on methodological quality and suggest ways to improve the rigor of the experimental designs used. Ultimately, we discuss some implications for authors, journals editors, policymakers, and curriculum developers.

1. INTRODUCTION

Experimental research findings have a considerable impact on the decisions taken by the policymakers about the educational practices that should be adopted in their own countries (Borman et al., 2005; Slavin, 2008). For instance, in the USA, Finn and Achilles (1999) conducted an experiment to investigate the effect of class size on elementary students' academic achievement and found out that the students in small classes (13–17 students) had superior performances compared to the students in regular classes (22–26 students). Finn and Achilles' (1999) findings initiated the educational reform entitled class-size reduction and led many states to reduce the number of students in the classrooms to improve student learning. In another experimental study, Schweinhart et al. (1993) examined the benefits of pre-school programs to children who live in poverty and who are at the risk of failing at school. They revealed that the pre-school students had significantly higher achievement scores, high school graduation rates, and earnings, while they had less crime rates and welfare use compared to the non-preschool students as of age 27. Their findings urged the legislators to deliver publicly funded programs in many states and localities in the USA.

Given the impact of experimental studies on educational policies and practices (Borman et al., 2005; Slavin, 2008), using rigorous methodological designs and techniques is of crucial

*CONTACT: Ramazan AVCU ramazanavcu@aksaray.edu.tr Aksaray University, Faculty of Education, Department of Mathematics and Science Education, Aksaray, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

importance in establishing that the observed effects in an experiment are caused by the treatment (e.g., a specific teaching technique, a newly developed curriculum, or an instructional program) but not by the extraneous variables. On the other hand, the flaws or errors in an experimental study may lead to false reports in the literature; other researchers may build theories or conduct other experiments by using these spoiled findings; and as a result, a great deal of time, money, effort, and other resources may be wasted (Gravetter et al., 2021). For instance, the United States Department of Education (2020) announced that almost 1.5 billion dollars were invested between the years 2018 and 2020 to support high-quality STEM (Science, Technology, Engineering, and Mathematics) education for students. In return for this, it expects researchers to conduct experiments with random assignments (i.e., randomized trials) and prioritizes researchers whose grant applications involve such rigorous methodological designs (Hedges & Schauer, 2018) because only in such designs it can be ascertained that the observed effects on important student outcomes such as academic achievement are caused by STEM education practices but not by other extraneous variables.

In Turkey, traditional educational practices being implemented in schools are not considered sufficient for students in solving real-world problems and gaining the knowledge and skills that are compulsory for maintaining their future careers (Akgündüz et al., 2015). For this reason, the Ministry of National Education (2016) is planning to initiate a STEM education reform in the near future to help students gain the technical knowledge and skills needed in the contemporary workplace and consequently to better prepare them for real life. Unfortunately, an action plan for implementing STEM education in Turkey has not been prepared yet (Ministry of National Education, 2018). However, the Ministry of National Education (2016) advocates the conduction of research studies on STEM education as a first step in developing this action plan. Thus, examining the methodological quality of experimental STEM education articles may help educational policymakers and curriculum developers determine a clear STEM education road map for students in all educational stages. With this idea in mind, in this study, we aimed to conduct a methodological review of experimental STEM education articles published in refereed Turkish journals to reveal whether current research practices in these journals are in agreement with the canons of educational research as described in commonly used methodology textbooks such as Creswell and Creswell (2018), Frankel et al. (2012), Gall et al. (2007), Cohen et al. (2018), and Johnson and Christensen (2020).

1.1. Significance and Research Questions

Although plenty of researchers conducted content analysis studies to determine the trends in STEM education research (e.g., Aydın Günbatar & Tabar, 2019; Brown, 2012; Çavaş et al., 2020; Çevik, 2018; Daşdemir et al., 2018; Elmalı & Balkan Kıyıcı, 2017; Kaya & Ayar, 2020; Li et al., 2020; Mizell & Brown, 2016), there is a dearth of studies that explore the methodological quality of educational research articles published in refereed journals (e.g., Horton et al., 1993; Shaver & Norton, 1980; Sung et al., 2019; Wallen & Fraenkel, 1988). What is more, to the best of our knowledge, there are no studies that investigate the methodological quality of experimental research articles on STEM education. Thus, this study attempts to fill this gap by analyzing the experimental STEM education articles published in refereed Turkish journals with respect to the following categories: formulating purpose statements, research questions, and hypotheses; clarifying contribution to the literature; describing the type of experimental design; describing the sample, sampling strategy, and the population; establishing instrument validity and reliability and describing their types; fulfilling the basic assumptions of the parametric tests used; attending to minimum sample size in experimental and control groups; and reporting effect sizes and statistical powers for the parametric tests used.

Since a methodological review of experimental research on STEM education in refereed Turkish journals had not been undertaken before, the findings of the study first provide the

researchers and other stakeholders with a snapshot of prevailing research reporting practices in Turkish journals. Second, they inform the STEM education community about the state-of-the-art and the soundness of experimental research practices in these journals. More importantly, since "a periodic review of common research practices in a scholarly discipline aids in improving those practices" (Horton et al., 1993, p. 858), our findings may enhance the quality of experimental research articles that will be published in these journals. Based on our findings, the editors and editorial board members of these journals may increase their article publication standards by ensuring that information about the above categories is provided by the authors who intend to publish their manuscripts in these journals. As authors pay increased attention to ensuring methodological rigor in their research manuscripts, deficiencies in their research reporting practices may diminish, and this may pave the way for more meaningful and consistent research on STEM education in Turkish journals.

In the transition from traditional education to STEM education, the Ministry of National Education (2016) deems it very significant to prepare and implement a good action plan considering the common sense of all stakeholders in the educational arena. Thus, the findings of our study may benefit but much to the Ministry of National Education. If publishing high-quality research articles becomes a standard practice for refereed journals in Turkey, the research findings about STEM education in these journals may point curriculum developers in the Ministry of National Education in the right direction. Frankly, the rigorous research findings accumulated from these journals may help curriculum developers design relevant STEM education materials and optimal STEM learning environments for students and help them integrate STEM education into Turkish school curricula in the best possible way.

Due to the above considerations, we conducted a methodological review to determine whether authors' experimental research reporting practices on STEM education in refereed Turkish journals are consistent with the commonly suggested research methods and procedures. Through this purpose, we sought to find answers to the following research questions:

1. Which research components do authors typically report in their articles?
    a. Do they report how they contribute to the scholarly literature?
    b. Do they report purpose statements?
    c. Do they report research questions?
    d. Do they report hypotheses?
    e. Do they report the type of experimental research design used?
    f. Do they describe the sample, the sampling strategy, and the population?
    g. Do they report instrument validity and reliability and describe their types?
    h. Do they report the basic assumptions that must be fulfilled for the parametric tests used?
    i. Do they attend to the minimum sample size required for experimental and control groups?
    j. Do they report effect sizes and statistical powers?

2. METHOD

2.1. Research Design

We conducted a *methodological review* to determine the methodological quality of experimental STEM education articles published in refereed Turkish journals. Methodological reviews describe the research designs, methods, and procedures used in scientific research and they foreground the strengths and weaknesses of methodological tools used in such research (Dochy, 2006). They are used in many fields to "improve research practice, inform debate, and identify islands of practice" (Randolph et al., 2013, p. 2). In these reviews, the focus is on identifying *how* research studies are conducted (i.e., the research methodologies used) rather than on identifying *which* research outcomes (i.e., the findings) are presented (Shukla, 2017).

Taking all these together, our methodological review helped us uncover authors' prevailing research reporting practices in the articles published in Turkish journals, determine the publishing standards of these journals, and suggest ways to improve the methodological quality of the articles published in these journals.

2.2. Data Sources

To locate the STEM articles published in refereed Turkish journals, we first formulated the following search terms (i.e., keywords or descriptors): STEM, STEM education, integrated STEM education, FeTeMM, FTMM (Turkish equivalents of STEM), FeTeMM eğitimi, FTMM eğitimi (Turkish equivalents of STEM education), entegre FeTeMM eğitimi, and entegre FTMM eğitimi (Turkish equivalents of integrated STEM education). Next, we typed these keywords in the following databases: TR Index (https://trdizin.gov.tr/), DergiPark (https://dergipark.org.tr/tr/), and Google Scholar (https://scholar.google.com.tr/). TR Index and DergiPark are national databases in Turkey, while Google Scholar is a search engine that is widely used all over the world. Our reason for using Google Scholar is that it "provides a simple way to do a broad search for scholarly literature, including peer-reviewed papers, theses, books, abstracts and articles" (Fraenkel et al., 2012, p. 55). More importantly, Google Scholar searches the entire internet. By this means, we were able to locate the articles that were not produced by the TR Index and DergiPark. In Google Scholar, we limited our search to "Turkish pages" to locate the articles published in Turkish journals and not to locate too many references. However, we used Google Scholar only as a supplement to TR Index and DergiPark and not as a substitute for them.

We delimited our search to the articles published till December 30, 2020. Our search elicited many studies with different research designs such as survey studies, correlational studies, theoretical papers, literature reviews, meta-analysis studies, scale development studies, and content analysis studies. However, to act in accordance with the purpose of the current study, we considered only the articles that used purely experimental research designs and the articles that combined experimental research designs with qualitative research designs (i.e., mixed methods studies). Thus, in the current study, 44 purely experimental research articles, 24 mixed methods articles, and in total 68 STEM education articles from 52 different journals underwent content analysis.

2.3. Coding Framework

To formulate the coding categories that are pertinent to our study, we first developed a tentative coding framework based on previous research on methodological quality (e.g., Horton et al., 1993; Shaver & Norton, 1980; Sung et al., 2019; Wallen & Fraenkel, 1988). Namely, we first used the predetermined categories developed by past researchers. As we coded the journal articles selected for our study, we had to make some changes to some of the codes or categories included in the tentative coding framework. Namely, we added some new codes or categories, deleted some of the codes or categories that were specified a priori, and refined some of these a priori codes or categories until the remaining codes and categories totally reflected the structure of our data. A final coding framework was developed when we were able to code all the journal articles exhaustively and explicitly with the codes and categories at hand. The categories and codes included in our final coding framework are explained below.

2.3.1. *Contribution to the literature*

Contribution to the literature refers to relating the intended study to previous studies in a planned way (Nelson & Shaver, 1985). In other words, it refers to situating the intended study in the context of the existing body of literature simply to "avoid reinventing the wheel" (Orne, 1981, p. 1). There are four ways to report how a piece of research contributes to the literature: *i*) filling a gap or void in the literature, *ii*) replicating past research, *iii*) extending past research,

and *iv*) developing new ideas in the scholarly literature (Brown & Dant, 2008). Filling a void refers to examining concepts or ideas not addressed in the existing literature (Creswell, 2015). Replication refers to repeating a past study using a different group of participants and under different conditions such as different places, abilities, and socioeconomic status (Fraenkel et al., 2012). Extending past research refers to broadening a published study to a new topic or field or simply carrying out the study more deeply and exhaustively, for example, by incorporating new variables into the study (Ary et al., 2014; Creswell, 2015). Developing new ideas in the scholarly literature means dealing with new problems (i.e., problems that have not been explored before) that concern researchers and practitioners (Brown & Dant, 2008).

2.3.2. *Purpose statement*

Creswell and Creswell (2018) defined a purpose statement as a passage that conveys the overall intent of a research study in one or more sentences. They emphasized that a good purpose statement must be clear, specific, and informative and proposed the following design features in writing a good purpose statement: (*i*) words such as *purpose*, *intent*, *aim*, and *objective* should be used to draw attention to the central topic of the study, (*ii*) the study should be narrowed to a single phenomenon, concept, or idea, (*iii*) action verbs such as *examine*, *explore*, *discover*, *develop*, *generate*, and *understand* should be used to convey how the topic of the study will be learnt, (*iv*) directional language should be avoided and instead, neutral words or phrases should be used, (*v*) working definitions should be formulated especially for the terms that are not normally known by a large number of people in the research community, (*vi*) words that specify the research design of the study should be used, (*vii*) the participants of the study should be mentioned, (*viii*) the research site should be identified, and (*ix*) the research participants and sites should be delimited. In the current study, the journal articles that did not consider any one or more of these design features in their purpose statements were categorized as articles having *unclear* purpose statements.

2.3.3. *Research question*

A research question is a statement that is used to narrow the purpose statement to specific questions that a researcher attempts to answer by carrying out a study (Plano Clark & Creswell, 2015). Research questions are "concrete questions, carefully composed in order to address the research objectives, to constitute a fair operationalization and embodiment of a valid set of indicators for addressing the research objectives, providing answers which address the research purposes with warranted data" (Cohen et al., 2018, p. 165). The journal articles analyzed in the present study were divided into two as articles containing research questions and articles not containing research questions.

2.3.4. *Hypothesis*

A hypothesis is a prediction of the anticipated findings from scientific research (Fraenkel et al., 2012; Gall et al., 2014). Stating a hypothesis in a research study helps to ponder more thoroughly and precisely on the findings anticipated from a study, build a body of knowledge, and notice whether relationships between different variables are, or are not, examined (Fraenkel et al., 2012). There are two opposing hypotheses as null hypothesis (H_0) and alternative hypothesis (H_1) and hypothesis testing works under the premise that the null hypothesis is true (Gravetter et al., 2021). More precisely, researchers start with the null hypothesis, cast their research in the form of a null hypothesis, and turn to the alternative hypothesis when their data do not support the null hypothesis (Cohen et al., 2018). Simply put, "the null hypothesis is the focal point in hypothesis testing because it is the null hypothesis, not the alternative hypothesis, that is tested directly" (Johnson & Christensen, 2020, p. 514). In the current study, the journal articles were categorized into two as articles reporting a hypothesis and articles not reporting a

hypothesis. Articles reporting a hypothesis were further categorized into two as articles reporting a null hypothesis and articles reporting an alternative hypothesis.

2.3.5. *Type of experimental research design*

In experimental research studies, the effect of a treatment or an intervention (an independent variable) on an outcome (a dependent variable) is tested by attempting to control for all other factors (extraneous variables) that may influence that outcome (Creswell & Creswell, 2018). Experimental research studies provide the best way to establish cause-effect relationships between different variables (Fraenkel et al., 2012). Namely, they produce the strongest evidence of causality (Johnson & Christensen, 2020). In the current study, we considered the typology proposed by Fraenkel et al. (2012) and Johnson and Christensen (2020) and divided the experimental research designs used in the articles selected for analysis into three types: true experimental designs, quasi-experimental designs, and weak experimental designs. In true experimental designs, there are both experimental and control groups and the study participants are randomly assigned to these groups. In quasi-experimental designs, there are again experimental and control groups, but the study participants are not randomly assigned to these groups. In weak experimental designs, there is either no control group (i.e., there is only an experimental group) or the study participants are not randomly assigned to the groups (i.e., the groups are pre-existing or intact/static groups). Thus, true experiments are the most rigorous experimental design types, while weak experiments are the least rigorous ones.

2.3.6. *Sampling strategy*

The strategy used while selecting a sample from a population is called the sampling strategy (Johnson & Christensen, 2020). Educational research textbooks describe two main types of sampling as random sampling (probability sampling) and nonrandom sampling (nonprobability sampling). In these textbooks, the most commonly reported random sampling strategies are simple random sampling, systematic sampling, stratified sampling, and cluster random sampling, while the most commonly reported nonrandom sampling strategies are convenience sampling, purposive sampling, and quota sampling (Ary et al., 2014; Cohen et al., 2018; Johnson & Christensen, 2020; Mills & Gay, 2016). In the present study, when coding the selected articles, we considered the abovementioned strategies. For the articles that did not report any specific strategy, we coded only the main sampling strategy used. For the articles that did not provide any information about the main or specific sampling strategy used, we used the code “not reported”.

2.3.7. *Sample description*

Sample description refers to the information given about societal, demographic, economic, and other characteristics of the subjects who take part in a research study (Erdoğan, Marcinkowski, & Ok, 2009). In the current methodological review study, the subjects recruited were pre-service teachers and K-12 students. The socio-demographic characteristics of the pre-service teachers were age, gender, university fund type (privately versus publicly funded university), department studied, year level, cumulative grade point average, and type of high school graduated (traditional high school, foreign language intensive high school, or vocational and technical high school). Location of the university (the region or city where the university is located) was also another characteristic reported in the articles that recruited pre-service teachers as participants.

The socio-demographic characteristics of the K-12 students included their age, gender, ethnicity, educational stage (preschool, elementary school, middle school, and high school), school fund type (privately versus publicly funded school), giftedness, achievement level, attitude level, English language proficiency, English as a second language, special education status, at-risk status, and high school type (traditional high school or inclusive STEM high

school). K-12 students' familial characteristics included socio-economic status (SES), economic status, and residence (state, region, city, district, or village). K-12 students' other characteristics as reported in the selected articles were class size and educational opportunity.

To operationalize the extent of sample description in each journal article, we used the following categorization including three levels: poor description, mediocre description, and rich description. Poor description refers to articles that describe at most three different characteristics of the sample. Mediocre description refers to articles that describe four, five, or six different characteristics of the sample. Rich description refers to articles describing more than six different characteristics of the sample.

2.3.8. *Population*

The notion of science rests entirely on the idea of generalization (Fraenkel et al., 2012). In quantitative studies, researchers obtain information from a small group of individuals and usually wish to generalize their findings to a larger group of individuals (Fraenkel et al., 2012). This larger group, which includes all possible members of a group of people, events, or objects, is called a population (Ary et al., 2014). Briefly, population refers to the "set of all the individuals of interest in a particular study" (Gravetter et al., 2021, p. 4). Defining a population helps researchers determine the extent of generalizability of their findings (Mills & Gay, 2016). In respect to this, Fraenkel et al. (2012) emphasized that researchers should avoid narrowly defined populations as much as possible because in such studies the usefulness of the obtained findings is severely restricted. They also remarked that it is not worth spending a considerable amount of time, energy, and money on studies that produce low applicable findings.

In the present study, the selected STEM articles were categorized into two as those that reported a population and those that did not report a population. Moreover, for those that reported a population, we also evaluated population sizes.

2.3.9. *Instrument validity and reliability*

Instrument validity refers to the "appropriateness, correctness, meaningfulness, and usefulness of the specific inferences researchers make based on the data they collect" (Fraenkel et al., 2012, p. 148) and instrument reliability refers to the "consistency of the scores obtained" (Fraenkel et al., 2012, p. 154). In the current study, we first categorized whether the authors of the selected journal articles developed their own instruments or used pre-existing instruments developed by others. Next, we coded the availability of validity and reliability information about the instruments used by the authors no matter who developed these instruments. In other words, we also paid particular attention to coding availability of validity and reliability information for the articles in which already developed instruments were administered. This is because even formerly developed instruments with perfect validity and reliability do not guarantee that they will function in the same way in the latter studies. Differences in participants and contexts may make earlier validity and reliability coefficients non-transferable to novel participants and contexts. Moreover, validity is always contingent upon the goals and interpretations of the researchers (Fraenkel et al., 2012).

We further categorized the journal articles with respect to the types of validity and reliability used in them. We delimited our analysis of instrument validity to the following three major types: content validity, criterion-related validity, and construct validity (Ary et al., 2014; Cohen et al., 2018; Fraenkel et al., 2012; Mills & Gay, 2019). Similarly, we considered the following commonly reported reliability types in the educational literature when categorizing the selected articles: internal consistency (*i*. Cronbach's alpha, *ii*. Kuder-Richardson, and *iii*. split-half), test re-test, equivalent-forms, and interrater agreement (Ary et al., 2014; Creswell, 2015; Fraenkel et al., 2012; Johnson & Christensen, 2020; Mills & Gay, 2019).

2.3.10. *Basic assumptions of parametric tests*

Experimental research studies involve comparing scores obtained from two or more groups or under different conditions (Gravetter et al., 2021). Parametric tests are a subcategory of inferential statistics tests and are usually used to compare differences between the groups or conditions (Pagano, 2013). However, they require the fulfillment of several assumptions about the population and nature of data (Pallant, 2016).

In the present study, to examine how well the selected journal articles fulfilled the basic assumptions required for conducting parametric tests, we used the judgment tree proposed by Sung et al. (2019). According to this tree, *t*-tests and between-groups ANOVAs must meet the basic assumptions of normality and homogeneity of variance. Thus, articles that examined these assumptions and that did not report any violations were categorized as fulfilling the assumptions. However, if the articles did not examine normality and homogeneity of variance for *t*-tests and between-groups ANOVAs, we considered the following two criteria: (1) Are the number of participants in each group or cell equal to or greater than 30? (2) Are there an equal number of participants in each cell or group? Articles that met these criteria were also categorized as fulling the assumptions because *t*-tests and ANOVAs are robust with respect to violations of the normality and homogeneity of variance assumptions (Howell, 2017; Pagano, 2013).

According to Sung et al.'s (2019) judgment tree, to conduct repeated-measures ANOVA and mixed-design ANOVA, homogeneity of regression slopes in addition to normality and homogeneity of variance must be met (Hair et al., 2019; Kirk, 2013). Thus, the journal articles that met the three assumptions for repeated-measures ANOVA and mixed-design ANOVA were categorized as fulfilling the basic assumptions; otherwise, they were categorized as not fulfilling the basic assumptions. Similarly, to conduct ANCOVA, researchers must satisfy the sphericity assumption in addition to normality and homogeneity of variance (Hair et al., 2019; Kirk, 2013). Thus, articles meeting these three basic assumptions for ANCOVA were also coded as fulling the basic assumptions.

2.3.11. *Sample size*

Sample size refers to the number of participants in a research study (Frey, 2018). It is important to note that by sample size we refer to the final sample size, not to the designated sample size (Shapiro, 2008), because the number of participants in a designated sample may be much fewer if a considerable number of individuals drop out of experimental research studies. Authors of commonly used educational research textbooks (e.g., Ary et al., 2014; Fraenkel et al., 2012; Mills & Gay, 2016) recommend a minimum of 30 participants in each cell or group (i.e., experimental and control groups). Including a minimum of 30 participants in each group is significant because a sample size less than 30 for each group may lead to low statistical power and this may, in turn, endanger the validity of experimental research (Cheung & Slavin, 2012). A summary of commonly used experimental research designs and the sample sizes needed to conduct the corresponding statistical tests are presented in Table 1.

Table 1. *The required sample size for statistical tests conducted under different experimental research designs when statistical power is 0.80, effect size is moderate, and* α = 0.05 (Sung et al., 2019, p. 18).

	Within-subject design		Mixed design	Between-subject design				
Experimental	Paired *t*-test	Multi-factor ANOVA (2*2 levels, interaction)	Multi-factor ANOVA (2*2 levels, interaction)	Independent *t*-test	ANOVA (2 levels)	ANOVA (3 levels)	ANCOVA (2 levels)	Multi-factor ANOVA (2*2 levels, interaction)
Pre-experiment	34							
Quasi-experiment								
Pre test				128	128	159		
Gain Score				128	128	159		
ANCOVA							128	
Counterbalance	34	30						
Multi-factor		30	34					179
True-experiment			34	128	128	159	128	179

Upon examining the selected articles, we used the following categories for the number of participants in each group or cell: 10–19, 20–29, 30–49, 50–99, and 100 and above.

2.3.12. *Effect size*

Effect size is a measure that quantifies the magnitude of difference between two groups (Coe, 2021). In experimental research, it refers to the treatment effect (Gravetter et al., 2021). It supplements statistical significance because statistical significance alone does not provide enough evidence for the importance of the findings (Warner, 2013). Besides, studies with large sample sizes can easily reach statistical significance even if the difference between the groups has little or no practical significance at all (Pallant, 2016). On the other hand, effect size is independent of sample size (Gravetter et al., 2021). Thus, it is not influenced by very small or large sample sizes. For this reason, reporting effect sizes in addition to statistical significance tests plays a crucial role in adopting a more rigorous approach to determining the effectiveness of experimental interventions and consequently encouraging a more scientific approach to the accumulation of research findings (Coe, 2021). As effect sizes are valuable means to report and interpret educational effectiveness, in this study, we categorized the selected articles into two as those that reported effect sizes and those that did not report effect sizes.

2.3.13. *Statistical power*

Statistical power refers to the "probability that the test will identify a treatment effect if one really exists" (Gravetter et al., 2021, p. 275). In experimental studies, it refers to the experiment's sensitivity to detect a treatment effect that really exists (Pagano, 2013). Statistical power is dependent upon three factors: sample size, effect size, and alpha level set by the researcher. It should be cautioned that in studies that are carried out with a quite small sample size (e.g., 20 participants), non-significant results may be obtained because of low statistical power. Thus, statistical power demonstrates how much confidence researchers should have in the results when they fail to reject the null hypothesis (Pallant, 2016). Besides, Cohen (1998) recommended that the power of a statistical test should be at least 0.80 (i.e., 80% probability of detecting an effect if there is actually one). Similarly, Sung et al. (2019) pointed out that if the power of a statistical test is less than 0.50, then obtaining a significant or non-significant result will be similar to guessing. Accordingly, given the importance of reporting statistical power, in this study, we categorized the selected journal articles into two as those that reported statistical power and those that did not report statistical power.

2.4. Data Analysis

Content analysis is used to analyze "written or visual materials for the purpose of identifying specified characteristics of the material" (Ary et al., 2014, p. 488). This research method helps to study human behavior indirectly usually through analysis of documents such as textbooks, essays, and magazine articles (Fraenkel et al., 2012). It is commonly used by educational researchers for several reasons such as revealing textbook biases, prejudices, and propaganda; analyzing error types in learners' writings; identifying prevailing practices; determining the difficulty level of a textbook content; and finding out the importance given to and the interest shown in some topics compared to the other ones (Ary et al., 2014). In the current study, we used this method to reveal the methodological quality of experimental STEM education articles published in refereed Turkish journals. When conducting our content analysis, we followed the steps recommended by Ary et al. (2014). These steps are explained below.

Specifying the phenomenon to be investigated: The phenomenon that we investigated in our content analysis was the *methodological quality* of publications on STEM education in the scholarly literature. More clearly, the phenomenon explored was Turkish educational researchers' prevailing research reporting practices. By examining this phenomenon, we aimed to determine the consistency between authors' research reporting practices and the commonly suggested research methods and procedures by well-known methodology textbooks.

Selecting the media from which the observations are to be made: The media selected for investigating the phenomenon of methodological quality were *peer-reviewed articles* published in Turkish journals. However, not all articles on STEM education were subjected to content analysis. That is, we only analyzed purely experimental STEM education articles and mixed methods articles on STEM education that used any type of experimental research design in their quantitative dimensions. There are several reasons for delimiting our analysis to peer-reviewed journal articles on STEM education in which experimental designs are used either completely or partially. First, journal articles provide the most recent research for the audience (Stebbins, 2006). Second, they are primary sources because the authors report their findings directly to the readers through them (Fraenkel et al., 2012). Third, they are expected to maintain higher standards to ensure quality (Creswell, 2015). Last and foremost, experimental research is "the only type of research that directly attempts to influence a particular variable", and consequently, is "the best way to establish cause-and-effect relationships among variables" (Fraenkel et al., 2012, p. 265). It is for this reason that policymakers and other stakeholders consider experimental research findings when making decisions about nationwide educational practices (Gall et al., 2007).

Formulating coding categories: Based on previously developed coding frameworks (e.g., Horton et al., 1993; Sung et al., 2019; Wallen & Fraenkel, 1988) and recommendations of well-known methodology textbook authors (e.g., Creswell, 2015; Frankel et al., 2012; Cohen et al., 2018; Johnson & Christensen, 2020) for conducting more rigorous research, we designed a comprehensive coding framework that comprises the following categories: formulating purpose statements, research questions, and hypotheses; clarifying contribution to the literature; describing the type of experimental design; describing the sample, sampling strategy, and the population; establishing instrument validity and reliability and describing their types; fulfilling the basic assumptions of the parametric tests used; attending to minimum sample size in experimental and control groups; and reporting effect sizes and statistical powers for the parametric tests used. In the previous section, these coding categories were defined and explained in some detail.

Deciding on the sampling plan to be used: STEM education research does not have a long history in Turkey. Several researchers (e.g., Aydın Günbatar & Tabar, 2019; Daşdemir et al., 2018; Elmalı & Balkan Kıyıcı, 2017) indicated that, in Turkey, STEM education research was

first initiated in 2014. Similarly, Özcan and Koca (2019) expressed that STEM education research in Turkey has gained momentum only over the past 5 years. Furthermore, our extensive review of literature also shows that there are a limited number of research studies on STEM education in Turkey. More importantly, we could locate a significantly fewer number of refereed journal articles on STEM education that employed an experimental research design. This extensive literature review helped us decide on the sampling plan to be used. Namely, in our content analysis, we attempted to locate the entire population of experimental STEM education articles published in refereed Turkish journals and thereby aimed to obtain an almost perfectly representative sample.

Training the coders: The first and second author of the current study coded the methodological quality of the journal articles. Before the actual coding, the two authors conducted sample coding for several experimental articles (different from the 68 articles selected for actual analysis) published in refereed Turkish journals. They independently coded the sample articles. Next, they held several sessions to discuss their independent coding and resolve the conflicting codes. These sessions also helped to clarify the meanings of the categories, make them more complete, and consequently revise and refine the coding framework. After the training session, the two authors separately coded all articles with respect to methodological quality by using the final form of the coding framework. In the first round of coding, the intercoder agreement (Miles et al., 2014) between the two coders was around 80%. Miles et al. (2014) recommended that "intercoder agreement should be within the 85% to 90% range, depending on the size and range of the coding scheme" (p. 85). Thus, the coders first identified the conflicting codes and re-examined the corresponding articles. In the second round of coding, the intercoder agreement reached 93%. The two authors discussed the rest of the conflicting codes periodically until they negotiated and arrived at a full consensus.

Analyzing the data: Once we have finalized coding our data, we counted the frequency of each code under each category. We also calculated percentages for these codes. Next, as recommended by Fraenkel et al. (2012), we assigned a label for each article (i.e., A1–A68) to facilitate data analysis. In these labels, the letter A stands for "Article" and the numbers ranging between 1 and 68 denote articles' IDs. The frequencies and percentages about each code under each category helped us summarize and interpret our research data. Namely, through frequencies and percentages, we were able to reveal refereed Turkish journals' trends in STEM education regarding methodological quality. More specifically, we could detect the decreasing and increasing trends in authors' use of research methods and procedures to improve the rigor of their articles.

2.5. Trustworthiness of the Study

To establish the trustworthiness of our study, we applied the following criteria proposed by Lincoln and Guba (1985): credibility, transferability, dependability, and confirmability. To ensure credibility, we explained our rationale for using a methodological review, thickly described the categories of our coding framework, examined previous research findings on the methodological quality of journal articles, and compared our findings with these previous research findings in the discussion section. Moreover, we used both data and investigator triangulation, spent prolonged time reading the full texts of the articles (over 3 months) to become familiar with the data and to obtain rich data for analysis, and finally, we used peer debriefing. That is, we had an associate professor review our data and examine our codes and categories. He has considerable experience in experimental research and methodological reviews. Through his review and examination, we received constructive feedback and thereby improved the quality of our findings.

To ensure transferability, we clearly described how we selected the articles for content analysis, explained our potential inclusion criteria, and described the main characteristics of the articles

so that other researchers who would like to examine the methodological quality of experimental studies may evaluate whether the findings drawn from the articles analyzed in the current study are applicable to other scientific documents such as books and proceedings. To establish dependability, first, we tried to maintain consistency across the entire study period including the starting point of research, data collection, and analysis. Meanwhile, we described each of the research steps transparently. Second, we explained our data analysis process as clearly as possible and used tables to report our findings to help other researchers evaluate the whole data coding process and replicate our study if desired. To achieve confirmability, we tried to do our best to control our biases and paid careful attention to shaping our findings solely by the data collected from the journal articles. More specifically, we used Ahern's (1999) ten tips to achieve reflexive bracketing.

3. FINDINGS

In this section, we report the findings related to the methodological quality of experimental STEM education articles published in refereed Turkish journals. In what follows, we present the number and percentage of articles with respect to each category of our coding framework.

3.1. Contributions to the Literature

Articles' contribution types to the STEM education literature are presented by year of publication in Table 2.

Table 2. *Articles' contribution types to the STEM education literature.*

Contribution to the literature	2014		2015		2016		2017		2018		2019		2020		Total	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Filling a gap or void in the literature	1	50	1	25	-	-	1	17	5	38	3	27	11	39	22	32
Replicating past research	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Extending past research	-	-	-	-	-	-	-	-	-	-	-	-	1	4	1	1
Developing new ideas in the scholarly literature	-	-	-	-	-	-	2	33	6	46	2	18	8	29	18	26
Not reported	1	50	3	75	4	100	3	50	2	15	6	55	8	29	27	40

Note. Numbers inside the parentheses are percentages and the 2018, 2020, and Total columns do not add up to 100% due to round-off errors.

As shown in Table 2, by and large, less than half of the articles (40%) did not report how they contributed to the STEM education literature. Besides, only one article (A14) was designed to extend the findings of past research. More strikingly, none of the articles attempted to replicate past research. When the articles are examined on a yearly basis, it can be seen that 50% or more of the articles did not report how they contributed to the literature in 2014, 2015, 2016, 2017, and 2019. Encouragingly, this percentage decreased drastically in 2018 (15%) and 2020 (29%). It also appears that filling a gap or void in the literature was a more standard reporting practice for the articles published from 2014 to 2020 because, in each year excluding 2016, at least one article used this contribution type. On the other hand, developing new ideas in the literature seems to be a more recent practice since it was only used in the articles published from 2017 to 2020.

3.2. Purpose Statements

The breakdown of articles with respect to the formulation of purpose statements and publication years are presented in Table 3. As indicated in Table 3, it is encouraging to find that, all told, most of the articles (81%) contained clearly formulated purpose statements. Besides, the pattern across the 7-year period indicates that each year at least 50% of the articles provided purpose statements and this percentage reached its peak in 2018 (85%), 2019 (91), and 2020 (82%). On

the other hand, eight out of 68 articles (A2, A7, A14, A15, A36, A46, A47, and A52; 12%) included unclear purpose statements.

Table 3. *Articles' formulation of purpose statements.*

Formulation of purpose statements	2014		2015		2016		2017		2018		2019		2020		Total	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Purpose statement is clear	1	50	3	75	3	75	4	67	11	85	10	91	23	82	55	81
Purpose statement is not clear	-	-	-	-	-	-	1	17	2	15	1	9	4	14	8	12
Purpose statement is not formulated	1	50	1	25	1	75	1	17	-	-	-	-	1	4	5	7

Note. Numbers inside the parentheses are percentages and the 2017 column does not add up to 100% due to round-off errors.

3.3. Research Questions

The distribution of articles with regards to the formulation of research questions and publication years are given in Table 4.

Table 4. *Articles' formulation of research questions.*

Formulation of research questions	2014		2015		2016		2017		2018		2019		2020		Total	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Research questions are formulated	-	-	3	75	2	50	6	100	13	100	7	64	21	75	52	76
Research questions are not formulated	2	100	1	25	2	50	-	-	-	-	4	36	7	25	16	24

Note. Numbers inside the parentheses are percentages.

As depicted in Table 4, overall, about three-quarters of the articles (76%) formulated their research questions. When the articles are examined on a yearly basis, it can be seen that none of the articles specified research questions in 2014, while at least half of them specified research questions from 2015 to 2020. Notably, all of the articles reported research questions in 2017 and 2018, while there was some decrease in articles' research question reporting percentages in 2019 (64%) and 2020 (75%).

3.4. Hypotheses

The distribution of articles with respect to the formulation of hypotheses and publication years are given in Table 5.

Table 5. *Articles' formulation of hypotheses.*

Formulation of hypotheses	2014		2015		2016		2017		2018		2019		2020		Total	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Hypotheses are formulated	-	-	-	-	1	25	2	33	-	-	-	-	-	-	3	4
Hypotheses are not formulated	2	100	4	100	3	75	4	66	13	100	11	100	28	100	65	96

Note. Numbers inside the parentheses are percentages and the 2017 column does not add up to 100% due to round-off errors.

It appears from Table 5 that formulation of hypotheses is not an accepted standard for experimental research articles published in refereed Turkish journals. All in all, only three out of 68 articles (A55, A58, and A62) formulated hypotheses. One of these articles (A62) was published in 2016 and the remaining two articles (A55 and A58) were published in 2017. Furthermore, the hypotheses formulated in these articles were all in null hypothesis (H_0) form.

3.5. Types of Experimental Research Designs

The distribution of articles by experimental design types and publication years is presented in Table 6.

Table 6. *Distribution of experimental design types.*

Experimental design types	2014		2015		2016		2017		2018		2019		2020		Total	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Weak experimental design	2	100	2	50	2	50	3	50	4	31	6	55	9	32	28	41
Quasi-experimental design	-	-	2	50	2	50	3	50	9	69	5	45	17	61	38	56
True experimental design	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Not specified	-	-	-	-	-	-	-	-	-	-	-	-	2	7	2	3

Note. Numbers inside the parentheses are percentages.

As given in Table 6, altogether, more than half of the STEM education articles (56%) used quasi-experimental designs and roughly 40% of the articles used weak experimental designs. However, none of the articles adopted true experimental designs. In two articles (A68 and A28) experimental research designs were used but their types were not specified. An examination of the trend over the targeted period reveals that the emphasis on weak experimental designs declined gradually from 2014 to 2020 and that quasi-experimental designs became more prevalent in the articles published in recent years, especially in 2020.

3.6. Sampling Strategies

The classification of articles by sampling strategy and year of publication is presented in Table 7.

Table 7. *Sampling strategies used in the articles.*

	Sampling strategies	2014		2015		2016		2017		2018		2019		2020		Total	
		n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
I	Stratified random sampling	1	50	1	25	-	-	-	-	-	-	-	-	-	-	2	3
	Not specified	-	-	-	-	-	-	1	17	-	-	-	-	4	14	5	7
II	Convenience sampling	-	-	1	25	1	25	1	17	2	15	6	55	11	39	22	32
	Purposive sampling	-	-	-	-	-	-	1	17	1	8	1	9	2	7	5	7
	Not specified	-	-	-	-	-	-	1	17	-	-	-	-	1	4	2	3
Not reported		1	50	2	50	3	75	2	33	10	77	4	36	10	36	32	47

Note. I represents random sampling strategies and II represents nonrandom sampling strategies. Numbers inside the parentheses are percentages and the 2017 and Total columns do not add up to 100% due to round-off errors.

As seen in Table 7, on the whole, nearly half of the articles (47%) did not report the sampling strategy used. Meanwhile, about one-third of them (32%) used samples of convenience, generally intact classrooms that are easily available to the STEM education researchers, and only 10% of them used random sampling strategies. From 2014 to 2020, each year at least one-third of the articles did not report their sampling strategies. Convenience sampling was used in all years excluding 2014 and the tendency to use this sampling strategy increased drastically in 2019 (55%) and 2020 (39%). A year-by-year examination also shows that, excluding 2020, each year either one or none of the articles used random sampling strategies. Purposive sampling was used in 2017 and thereafter. In 2017, 2018, and 2019 one article and in 2020 two articles reported the use of this sampling strategy.

3.7. Description of Samples

As mentioned previously, the selected STEM education articles recruited either K-12 students or pre-service teachers as their subjects. The level of description of K-12 students in the selected articles is presented in Table 8.

Table 8. *The level of description of K-12 students in the selected articles.*

Level of description	2014		2015		2016		2017		2018		2019		2020		Total	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Poor description	-	-	2	67	1	25	1	33	2	25	1	17	8	36	15	32
Mediocre description	1	100	-	-	3	75	2	67	6	75	5	83	14	64	31	66
Rich description	-	-	1	33	-	-	-	-	-	-	-	-	-	-	1	2

Note. Poor description refers to reporting at most three different sample characteristics, mediocre description refers to reporting four, five, or six different sample characteristics, and rich description refers to reporting more than six different sample characteristics. Numbers inside the parentheses are percentages.

As shown in Table 8, overall, 31 out 47 articles (66%) provided mediocre description, about 30% of the articles provided poor description, and only one article provided rich description for their samples. It also appears from this table that there was not a detectable pattern in terms of sample description across the 7-year period. The level of description of pre-service teachers in the selected articles is presented in Table 9.

Table 9. *The level of description of pre-service teachers in the selected articles.*

Level of description	2014		2015		2016		2017		2018		2019		2020		Total	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Poor description	-	-	-	-	-	-	1	33	1	20	-	-	1	17	3	14
Mediocre description	1	100	1	100	-	-	2	67	4	50	4	80	3	50	15	71
Rich description	-	-	-	-	-	-	-	-	-	-	1	20	2	33	3	14

Note. Numbers inside the parentheses are percentages and the Total column does not add up to 100% due to round-off errors.

As given in Table 9, altogether, 15 out of 21 articles (71%) provided a mediocre description and three articles provided a rich description for their samples, while the remaining three articles poorly described their samples. A year-by-year examination shows that there is an increasing trend towards mediocre description in 2017 and 2018 and towards rich description in 2019 and 2020.

3.8. Populations

The breakdown of articles with respect to defining a population and year of publication is presented in Table 10.

Table 10. *Articles' description of their populations.*

Description of populations	2014		2015		2016		2017		2018		2019		2020		Total	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Population is defined	1	50	1	25	-	-	2	33	-	-	1	9	3	11	8	12
Population is not defined	1	50	3	75	4	100	4	67	13	100	10	91	25	89	60	88

Note. Numbers inside the parentheses are percentages.

As demonstrated in Table 10, nearly 90% of the articles did not define their populations. Moreover, it appears that reporting populations became an almost overlooked research reporting practice, especially in the last three years. Of the eight articles that reported population, six described very narrow populations (i.e., A4, A6, A10, A39, A53, and A67). That is, these articles reported their populations as only the schools or faculties from which the samples were selected.

3.9. Validity and Reliability of Data Collection Instruments

The selected STEM education articles either used the instruments that were developed in previous studies (i.e., pre-existing instruments) or developed their own instruments to collect their own data. Eighty-nine pre-existing instruments (74%), 32 self-developed instruments (26%), and altogether 121 different instruments were used in these articles. Unfortunately, for the pre-existing instruments, only three articles (i.e., A30, A49, and A53) reported the validity

of data collected from their own samples, while the rest of them solely reported the validity of data obtained from the original studies. The validity types reported for the instruments used in the STEM education articles are shown in Table 11.

Table 11. *The validity types reported for the instruments used in the STEM education articles.*

Validity of the instruments	2014		2015		2016		2017		2018		2019		2020		Total	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Likert scale																
Construct validity	1	50	-	-	1	100	7	70	7	41	11	85	15	60	42	62
Content and construct validity	-	-	-	-	-	-	-		-	-	-	-	1	4	1	1
Not reported	1	50	-	-	-	-	3	30	10	59	2	15	9	36	25	37
Achievement test																
Content and construct validity	-	-	2	40	-	-	3	100	6	60	3	75	7	39	21	50
Content validity	-	-	-	-	-	-	-	-	-	-	-	-	2	11	2	5
Construct validity	1	100	-	-	-	-	-	-	-	-	-	-	2	11	3	7
Not reported	-	-	3	60	1	100	-	-	4	40	1	25	7	39	16	38
Questionnaire																
Content validity	-	-	-	-	-	-	1	100	-	-	-	-	-	-	1	17
Not reported	-	-	-	-	4	100	-	-	-	-	-	-	1	100	5	83
Performance test																
Content validity	-	-	-	-	-	-	-	-	-	-	-	-	2	100	2	100
Semantic differential scale																
Construct validity	-	-	-	-	1	100	-	-	-	-	-	-	-	-	1	100
Ability test																
Not reported	-	-	-	-	-	-	-	-	-	-	-	-	1	100	1	100

Note. Thirty-two articles used only one instrument, 24 articles used two different instruments, 10 articles used three different instruments, and 2 articles used 5 different instruments. Numbers inside the parentheses are percentages.

As presented in Table 11, the STEM education articles mainly used Likert scales and achievement tests as data collection instruments. Overall, for a large proportion of the Likert scales (62%), only construct validity was reported. Both content and construct validity were provided for only one Likert scale. On the other hand, for a considerable proportion of Likert scales (37%), validity information was not provided. On a yearly basis, construct validity again seems to be a more standard validity reporting practice for the articles that used Likert scales as data collection instruments. Altogether, for half of the achievement tests used (50%), both content and construct validity were reported. On a yearly basis, reporting content and construct validity was also a predominant practice for the achievement tests used in the articles published from 2014 to 2020.

Pre-existing instruments used in the articles were Likert scales, achievement tests, questionnaires, and semantic differential scales. Sixty-seven pre-existing Likert scales were used in the articles. For 29 of them (40%), reliability information from both the original studies and their own data were reported. For 27 of them (40%), only reliability information from the original studies was provided. For 7 of them (10%), only reliability information from their own data was reported. For 4 of them (6%) reliability information was not provided. Sixteen pre-existing achievement tests were used in the articles. For 11 of them (65%), only reliability information from the original studies was provided. For 4 of them (24%), reliability information from both the original studies and their own data was reported. For one of them (6%), only reliability information for its own data was provided. For the remaining one (6%), reliability information was not provided. Five pre-existing questionnaires were used in the articles. For 4 of them (80%), only reliability information for their own data was provided and for the remaining one (20%), reliability information was not reported. Finally, one pre-existing semantic differential scale was used in A61 and reliability information from both the original study and its own data was reported.

Self-developed instruments used in the articles were achievement tests, performance tests, ability tests, and Likert scales. Twenty-six self-developed achievement tests were used in the selected STEM education articles and for 22 of them (85%), reliability information was provided. Two self-developed performance tests were used in A15 and reliability information was provided for both of them. One self-developed Likert scale was used in A22 and one self-developed questionnaire was used in A57 and reliability information was reported for both instruments. Finally, one self-developed ability test was used in A11. However, reliability information was not reported for this instrument. The reliability types reported for the instruments used in the selected STEM education articles are given in Table 12.

Table 12. *The reliability types reported for the instruments used in the STEM education articles.*

Reliability of the instruments	2014		2015		2016		2017		2018		2019		2020		Total	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Likert scale																
Cronbach's alpha	1	50	-	-	1	100	8	80	15	88	13	100	23	92	61	90
Cronbach's alpha and test-retest	-	-	-	-	-	-	-	-	1	6	-	-	1	4	2	3
Internal consistency not specified	-	-	-	-	-	-	-	-	1	6	-	-	-	-	1	1
Not reported	1	50	-	-	-	-	2	20	-	-	-	-	1	4	4	6
Achievement test																
KR-20	1	100	1	20	-	-	1	33	9	90	2	50	12	67	26	62
Cronbach's alpha	-	-	-	-	-	-	-	-	1	10	2	50	3	17	6	14
Inter-rater agreement	-	-	-	-	-	-	-	-	-	-	-	-	2	11	2	5
KR-21	-	-	-	-	-	-	1	33	-	-	-	-	-	-	1	2
Internal consistency not specified	-	-	1	20	1	100	-	-	-	-	-	-	-	-	2	5
Not reported	-	-	3	60	-	-	1	33	-	-	-	-	1	6	5	12
Questionnaire																
Cronbach's alpha	-	-	-	-	3	75	-	-	-	-	-	-	-	-	3	50
Inter-rater agreement	-	-	-	-	-		1	100	-	-	-	-	1	100	2	33
Not reported	-	-	-	-	1	25	-	-	-	-	-	-	-	-	1	17
Performance test																
Inter-rater agreement	-	-	-	-	-	-	-	-	-	-	-	-	2	100	2	100
Semantic differential scale																
Cronbach's alpha	-	-	-	-	1	100	-	-	-	-	-	-	-	-	1	100
Ability test																
Cronbach's alpha	-	-	-	-	-	-	-	-	-	-	-	-	1	100	1	100

Note. Thirty-two articles used only one instrument, 24 articles used two different instruments, 10 articles used three different instruments, and 2 articles used 5 different instruments. Numbers inside the parentheses are percentages and the 2017 and 2020 columns for achievement tests do not add up to 100 due to round-off errors.

Table 12 indicates that the selected STEM education articles mainly used Likert scales and achievement tests as data collection instruments and that they mostly used internal consistency estimates when reporting reliability. More specifically, for most of the Likert scales (93%), Cronbach's alpha estimates were calculated. Similarly, for more than half of the achievement tests (64%), Kuder-Richardson formulas (KR-20 and KR-21) were used. When examined per year, it can be seen that the tendency to use Cronbach's alphas and KR-20s for reporting reliabilities of Likert scales and achievement tests is especially more evident in the last three years (i.e., between 2018 and 2020). On the other hand, other methods such as test-retest and inter-rater agreement were seldom used to report reliabilities of data collection instruments.

3.10. Basic Assumptions of Parametric Tests

Table 13 demonstrates the extent to which the basic assumptions of the parametric tests are fulfilled in the selected articles. As shown in Table 13, the STEM education articles mainly used paired-samples *t*-test (55%) and independent-samples *t*-test (28%) as parametric tests. However, only less than half of the articles using paired-samples *t*-test (46%) could fulfill the corresponding basic assumptions. Similarly, only 37% of the articles using independent-

samples *t*-test satisfied the basic assumptions related to this test. On an annual basis, there seems to appear a stable trend for the STEM education articles towards not fulling the basic assumptions of parametric tests. For the paired-samples *t*-test, every year, at least nearly half of the articles did not check the basic assumptions. For independent-samples *t*-test, this is far more manifest because, each year excluding 2019, articles not fulling the basic assumptions outnumbered the ones that fulfilled the basic assumptions.

Table 13. *The fulfillment of basic assumptions of parametric tests used in the STEM education articles.*

Fulfillment of basic assumptions	2014		2015		2016		2017		2018		2019		2020		Total	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Paired-samples t-test																
Fulfilled	-	-	2	67	-	-	2	50	1	14	4	57	8	57	17	46
Not fulfilled	2	100	1	33	-	-	2	50	6	86	3	43	6	43	20	54
Independent-samples t-test																
Fulfilled	-	-	-	-	-	-	-	-	1	17	3	100	3	38	7	37
Not fulfilled	-	-	1	100	-	-	1	100	5	83	-	-	5	63	12	63
One-way between-groups ANOVA																
Fulfilled	-	-	-	-	-	-	1	100	1	100	-	-	1	100	3	100
Not fulfilled	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
One-way repeated-measures ANOVA																
Fulfilled	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Not fulfilled	-	-	-	-	-	-	-	-	-	-	-	-	1	100	1	100
One-way ANCOVA																
Fulfilled	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Not fulfilled	-	-	-	-	-	-	-	-	1	100	-	-	-	-	1	100
Two-way ANCOVA																
Fulfilled	-	-	-	-	-	-	-	-	-	-	-	-	1	100	1	100
Not fulfilled													-	-	-	-
One-way mixed-design ANOVA																
Fulfilled	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Not fulfilled	-	-	-	-	-	-	-	-	-	-	-	-	2	100	2	100
Two-way mixed-design ANOVA																
Fulfilled	-	-	-	-	-	-	1	50	-	-	-	-	-	-	1	33
Not fulfilled	-	-	-	-	-	-	1	50	1	100	-	-	-	-	2	67

Note. Thirty-four articles used only one parametric test, 15 articles used two different parametric tests, one article used three different parametric tests, 16 articles used nonparametric tests, one article used hierarchical linear modeling, and the remaining one used the Wald test for multi-group analysis. Numbers inside the parentheses are percentages and the 2020 column for independent-samples *t*-test does not add up to 100 due to round-off errors.

3.11. Sample Sizes

The number of participants used in each group or cell of the selected articles is categorized in Table 14. As can be calculated from Table 14, overall, 60% of the articles had less than 30 participants in their experimental and/or control groups. In 38% of the articles, the number of participants in each group ranged between 20 and 29 (38%), while only three articles used 100 or more participants in each group. Table 14 also shows that there does not appear an increasing and deliberate attempt to use at least 30 participants in experimental and control groups from 2014 to 2020.

Table 14. *The number of participants used in each group or cell.*

Number of participants	2014		2015		2016		2017*		2018*		2019*		2020		Total	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
10–19 participants	-	-	-	-	-	-	-	-	3	19	4	33	9	32	16	22
20–29 participants	2	100	-	-	2	50	3	43	7	44	4	33	10	36	28	38
30–39 participants	-	-	1	25	-	-	1	14	6	38	-	-	5	18	13	18
40–49 participants	-	-	1	25	-	-	1	14	-	-	1	8	1	4	4	5
50–99 participants	-	-	1	25	-	-	2	29	-	-	3	25	3	11	9	12
100 or more participants	-	-	1	25	2	50	-	-	-	-	-	-	-	-	3	4

Note. In columns marked with *, A38 (2019), A40 (2018), A45 (2018), A46 (2018), and A53 (2017) were counted twice due to unequal number of participants in the experimental (EG) and control groups (CG). In A38, EG = 20 and CG = 13. In A40, EG1 = 28, EG2 = 33, and CG = 26. In A45, EG = 28 and CG = 30. In A46, EG = 34 and CG = 22. In A53, EG1 = 30, EG2 = 26, and CG = 22. Numbers inside the parentheses are percentages and the 2017, 2018, 2019, and Total columns do not add up to 100% due to round-off errors.

3.12. Effect Sizes

The STEM education articles' reporting of effect sizes for the parametric tests they used are presented in Table 15. As depicted in Table 15, overall, more than half of the articles did not report effect sizes for the paired-samples *t*-tests (65%), independent-samples *t*-tests (63%), and two-way mixed-design ANOVAs (67%) they used. For one-way between-groups ANOVA, one-way repeated-measures ANOVA, and one-way ANCOVA, none of the articles reported effect sizes. All of the articles that used two-way ANCOVA (i.e., A27) and one-way mixed-design ANOVA (i.e., A15 and A24) reported effect sizes. When examined on a yearly basis, it can be seen that, each year excluding 2018, more than half of the articles did not report effect sizes for the paired-samples *t*-tests they conducted. Moreover, on a yearly basis, more than 60% of the articles that used independent-samples *t*-test did not report effect sizes.

Table 15. *Articles' reporting of effect sizes for the parametric tests used.*

Effect sizes reported	2014		2015		2016		2017		2018		2019		2020		Total	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Paired-samples *t*-test																
Reported	-	-	1	33	-		1	25	4	57	3	43	4	29	13	35
Not reported	2	100	2	67	-		3	75	3	43	4	57	10	71	24	65
Independent-samples *t*-test																
Reported	-	-	-	-	-		1	100	2	33	1	33	3	38	7	37
Not reported	-	-	1	100	-		-		4	67	2	67	5	63	12	63
One-way between-groups ANOVA																
Reported	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Not reported	-	-	-	-	-	-	1	100	1	100	-	-	1	100	3	100
One-way repeated-measures ANOVA																
Reported	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Not reported	-	-	-	-	-	-	-	-	-	-	-	-	1	100	1	100
One-way ANCOVA																
Reported	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Not reported	-	-	-	-	-	-	-	-	1	100	-	-	-	-	1	100
Two-way ANCOVA																
Reported	-	-	-	-	-	-	-	-	-	-	-	-	1	100	1	100
Not reported	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
One-way mixed-design ANOVA																
Reported	-	-	-	-	-	-	-	-	-	-	-	-	2	100	2	100
Not reported	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
Two-way mixed-design ANOVA																
Reported	-	-	-	-	-	-	1	50	-	-	-	-	-	-	1	33
Not reported	-	-	-	-	-	-	1	50	1	50	-	-	-	-	2	67

Note. Numbers inside the parentheses are percentages.

3.13. Statistical Powers

Regrettably, of the 68 articles reviewed, none reported the statistical power of the parametric tests used. This indicates that statistical power is an overlooked statistical measure for the experimental STEM education articles published in national indexed journals.

4. DISCUSSION and IMPLICATIONS

In this study, we examined the methodological quality of experimental STEM education articles published in the refereed Turkish journals from 2014 to 2020. In this way, we attempted to reveal the degree to which current research reporting practices in these journals are in agreement with the standards of educational research described in the commonly used methodology textbooks. In what follows, we discuss the findings in light of the literature on methodological quality.

4.1. Contributions to the Literature

The present study found that 40% of the articles did not clarify how they contributed to the STEM education literature. This implies that national indexed journals are largely lacking a systematic effort to build a cumulative knowledge base in the area of STEM education. What is more, researchers who are publishing in these journals may not be aware that similar or related research might have been or is being conducted elsewhere by other colleagues and it is most likely that they will continue to conduct isolated studies in the future. The researchers' failure to relate their studies to past research may have serious consequences on the quality of STEM education research. As emphasized by Nelson and Shaver (1985), not clarifying contributions to the literature may lead to the "repetition of unproductive prior research and a disconnectedness of studies on similar topics" (p. 410). Thus, conducting isolated studies on STEM education may highly be counterproductive to knowledge building in this area.

We also found that none of the articles reported a direct or systematic replication of previous research. Although replication can be used as a strategy to compensate for weaknesses in generalizability (Horton et al., 1993), it was not accepted as a research reporting practice by the STEM education researchers who publish in the refereed Turkish journals. Several other researchers also found that replication is a neglected practice in educational research. For instance, Horton et al. (1993) examined the methodological quality of articles published in the Journal of Research in Science Teaching from 1985 to 1989 and found that only four of them (3%) replicated past research. Similarly, Shaver and Norton (1980) examined two social studies journals and found that only four (13%) and three (14%) of the articles in these journals replicated previous research.

4.2. Purpose Statements, Research Questions, and Hypotheses

We found that around 80% of the STEM education articles contained clearly formulated purpose statements. It is encouraging to find that a large proportion of the articles included purpose statements. These statements clarify the primary objective or focus of our research and thus are the most important ones in research studies. Moreover, they signal the procedures we should use during data collection and they point to the types of findings we expect to obtain in our research (Creswell, 2015).

It is also good news that roughly 75% of the articles formulated their research questions. Research questions hint at the methodology used in a research study and to the data analysis methods that are relevant to that study (Aktemur, 2015). For example, the research question "what is the effect of X on Y?" infers an experimental research methodology and subsequently the statistical tests used during the analysis of data such as *t*-tests, ANOVAs, and ANCOVAs. Therefore, it can be said that three-quarters of the STEM education articles enabled the audience

to easily determine whether the research methodologies and corresponding data analysis procedures used in them were correct or not.

On the other hand, the findings revealed that only three (4%) articles formulated their hypotheses. It seems that formulating a hypothesis is not within the STEM education authors' research reporting routines. Hypotheses refer to researchers' expectations about how specific phenomena work and affect, while experiments are procedures conducted to confirm, rebut, or ascertain the validity of these hypotheses (Horváth, 2016). Thus, hypotheses are an important key tenet of experimental designs. For this reason, well-formulated hypotheses are compulsory for carrying out more rigorous experimental studies.

4.3. Types of Experimental Research Designs

It is sobering to find that none of the STEM education articles used true experimental designs. One possible reason for the absence of true experimental designs in the STEM education articles might be that researchers find it difficult to obtain random samples of students for their studies. However, as argued by Campbell and Boruch (1975), school settings do provide natural laboratories in which random assignment could usually be used.

We also found that the most frequently used research design was a quasi-experimental design (56%). This finding is in line with the findings of previous research that examined the design quality of experimental studies (e.g., Sung et al., 2019) or research trends in education (e.g., Baydaş et al., 2015; Duman et al., 2015). For instance, Sung et al. (2019) investigated the quality of experimental designs in mobile learning research from 2006 to 2016 and revealed that 63% of the articles used a quasi-experimental design. Baydaş et al. (2015) examined educational technology research trends from 2002 to 2014 and similarly found that 48% of the experimental studies used a quasi-experimental design. Moreover, Duman et al. (2015) analyzed research trends in the mobile-assisted language learning articles published from 2000 to 2012 and found that 12 out of 26 experimental studies (46%) employed a quasi-experimental design.

Alarmingly, we found that almost half of the articles (41%) used weak experimental designs. However, an important drawback of weak experimental designs is that they are subject to numerous threats to validity. Mills and Gay (2019) cautioned that weak experimental designs should be avoided as much as possible. They further indicated that the findings obtained from studies using weak experimental designs are very questionable and thus "they are not useful for most purposes except, perhaps, to provide a preliminary investigation of a problem" (p. 310). Weak experimental designs provide little or no control of extraneous variables (Ary et al., 2014) and consequently, it is almost impossible to refute rival hypotheses or explanations. Thus, extreme caution must be exercised in interpreting and generalizing the findings obtained from the weak experimental STEM education articles that were analyzed in the current study.

4.4. Description of Sampling Strategies, Sample Characteristics, and Populations

The present study found that 42% of the articles used nonrandom sampling strategies. More specifically, nearly one-third of the articles (32%) used convenience sampling. These high percentages signal that many of the articles published in national indexed journals from 2014 to 2020 used biased samples because such samples almost always differ systematically from the population with respect to particular characteristics (Johnson & Christensen, 2020). Besides, it is a major error to use significance tests in studies where samples are obtained nonrandomly (Wallen & Fraenkel, 1988). Random sampling is the fundamental basis of inferential statistics and "one must raise serious questions about the use of inferential statistics when the lack of randomness makes probability statements indeterminate" (Shaver & Norton, 1980, p. 8). In the present study, only seven articles (10%) used random samples. Thus, it can be said that only these articles can appropriately use the significance tests.

We also found that almost 90% of the articles did not define their populations and only a very small portion of the articles provided rich descriptions for their samples (i.e., one article for K-12 students and three articles for pre-service teachers). Providing rich descriptions for the samples being studied is crucial because it offers researchers some basis for deciding whether their findings are generalizable to the relevant contexts (Shaver & Norton, 1980). More explicitly, describing the details of experimental studies or the contexts of nonexperimental studies as vividly as possible may help other researchers evaluate the applicability of particular findings to their own situations (Fraenkel, 1987). Thus, our findings demonstrate that relevant generalizations beyond the samples described may indeed not be possible for most of the STEM education articles published in the national indexed articles. These findings are not surprising given that such reporting practices were found to be extremely widespread in many of the previous studies conducted on methodological quality (e.g., Aktemur, 2015; Horton et al., 1993; Shaver & Norton, 1980; Wallen & Fraenkel, 1988).

4.5. Validity and Reliability of Data Collection Instruments

Sound measurement is the keystone of rigorous research and it is very significant for high-quality experimentation (D'agostino, 2005). In addition, sound measurements rely on the validity and reliability of instruments. Thus, improving the validity and reliability of instruments enhances the rigor of research and the quality of experiments in particular (D'agostino, 2005). In the current study, the selected STEM education articles mainly used Likert scales and achievement tests as data collection instruments. Despite the emphasis on using valid instruments, validity information was not documented for 37% and 38% of these Likert scales and achievement tests, respectively. Therefore, a considerable proportion of the Likert scales and achievement tests used in the STEM education articles might have jeopardized the validity of conclusions drawn from these articles. Frankly speaking, the STEM education articles might have used faulty Likert scales and achievement tests and these instruments might have yielded biased outcomes (D'agostino, 2005). Similar findings were obtained in previous studies about the documentation of instrument validity (e.g., Sung et al., 2019; Wallen & Fraenkel, 1988). For instance, Wallen and Fraenkel (1988) examined articles published in Theory and Research in Social Education over an eight-year period and found out that only 30% of the articles empirically checked the validity of instruments used. Similarly, Sung et al. (2019) revealed that only 24% of the mobile-learning articles provided information about instrument validity.

Encouragingly, we found that reliability information was provided for 94% of the Likert scales and 88% of the achievement tests used in the STEM education articles. More specifically, Cronbach's alpha reliabilities were reported for all types of instruments without any exceptions. However, test-retest reliability and inter-rater reliability were seldom reported and split-half reliability was never reported in these articles. These reliability types deal with different kinds of test consistencies. For instance, test-retest reliability measures the stability of scores over time and in high-quality journals, it is almost always reported in company with internal consistency reliability (Johnson & Christensen, 2020). Moreover, Mills and Gay (2016) stress that test-retest reliability is particularly crucial for instruments that are used for making predictions because predictions are based to a large extent on the assumption that the scores are stable over time.

4.6. Basic Assumptions of Parametric Tests

In this study, we used Sung et al.'s (2019) judgment tree to determine how well the selected STEM education articles fulfilled the basic assumptions of the parametric tests used. We found that 17 out of 37 articles (46%) met the basic assumptions of the paired-samples *t*-test and 7 out of 19 articles (37%) met the basic assumptions of the independent samples *t*-test. Moreover, for

one-way repeated-measures ANOVA, one-way ANCOVA, and one-way mixed-design ANOVA, the basic assumptions were not satisfied by any one of the articles. This finding shows that the selected STEM education articles largely overlooked the significance of basic assumptions when conducting certain parametric tests. However, violation of the basic assumptions gives rise to invalid probability inferences from these parametric tests (Aron et al., 2019). For instance, ANCOVA mandates that the relationship between the dependent variable and the covariate must be the same for each group (i.e., homogeneity of regression slopes). Unequal regression slopes demonstrate that there is an interaction between the treatment and the covariate and that the findings will be misleading in case ANCOVA is conducted (Tabachnick & Fidell 2019).

Similarly, one-way repeated-measures ANOVA necessitates that "variance of the population difference scores for any two conditions are the same as the variance of the population difference scores for any other two conditions" (i.e., sphericity; Pallant, 2016, p. 287). If this assumption is violated, one-way repeated-measures ANOVA will become too liberal and provoke inflation of Type I error rates (Tabachnick & Fidell, 2019). In such cases, significance tests such as Greenhouse-Geisser or Huynh-Feldt may be used as alternatives to avoid biased conclusions (Tabachnick & Fidell, 2019). Nevertheless, the articles that violated the basic assumptions of one-way repeated-measures ANOVA (i.e., A22), one-way mixed-design ANOVA (i.e., A15 and A24), and two-way mixed-design ANOVA (i.e., A43 and A56) did not report the use of such corrective measures.

Our findings show, by and large, that fulfilling the basic assumptions of the statistical tests is not a cut-and-dried practice for most of the STEM education articles that we examined. The findings also demonstrate that the STEM education articles failed to make appropriate adjustments when they did not meet the basic assumptions of the parametric tests and therefore, they seriously threatened the validity of findings on STEM education.

4.7. Sample Sizes, Effect Sizes, and Statistical Powers

Although many educational research textbooks (e.g., Ary et al., 2014; Fraenkel et al., 2012; Mills & Gay, 2016) recommend a minimum of 30 participants in each of the experimental and control groups, the current study revealed that the most proportion of the articles (38%) had 20–29 participants in their groups. This shows that using 20–29 participants for each cell might have been accepted as a sample size standard for the STEM education articles published in national indexed journals. What is worse, 22% of the articles had 10–19 participants in their groups. This is especially alarming because studies that recruit an inadequate number of participants for each cell or group are most likely to produce low statistical power (Cheung & Slavin, 2012). For instance, a medium effect size ($0.2 < d < 0.8$), as most commonly reported in educational research studies, with 25–30 participants produces a statistical power of 0.47 (Cohen, 1988). Since this value is below 50%, it appears that for 60% of the selected STEM education articles, the probability of correctly detecting a real treatment effect is worse than guessing. This alerts that the selected STEM education articles may highly be prone to Type II errors.

Moreover, due to the existence of *between-persons errors*, recruiting a fewer number of participants for between-subjects designs may have far more serious consequences on the accuracy of findings compared to within-subject designs and mixed-designs (Privitera, 2019). For between-subjects designs and the related parametric tests (e.g., independent-samples *t*-tests, between-groups ANOVAs, and ANCOVAs) at least 128 participants must be recruited to obtain a statistical power of 0.80 when there is a medium effect size (see Table 1). In the present study, we found that 19 articles (28%) used independent-samples t-tests, three articles used between-groups ANOVAs (4%), and 2 articles (3%) used ANCOVAs for their between-subjects designs. Since such parametric tests were often used in STEM education articles, future studies in this

area must recruit a greater number of participants to achieve the widely accepted standard of statistical power of 0.8 for these tests.

Our findings also indicated that more than 60% of the articles did not report effect sizes for each of the parametric tests used (see Table 15). This demonstrates that the authors of these articles relied solely on statistical significance tests when reporting their findings. However, statistical significance tests alone would be misleading and may lead to many different conclusions. Namely, studies with large sample sizes can readily reach statistical significance despite demonstrating very small practical significance. Conversely, studies with very few sample sizes may not achieve statistical significance despite having very large practical significance. Unlike statistical significance tests, effect sizes are not contingent upon sample sizes (Gravetter et al., 2021). Besides, since effect sizes add "a more exact numerical statement of facts" (Hanel & Mehler, 2019, p. 469), they are more informative than statistical significance (Cohen, 1994). Thus, for a more accurate interpretation of findings, STEM education researchers who publish in national indexed journals should pay more attention to supplementing statistical significance tests with effect sizes.

Ultimately, we found out that none of the STEM education articles reported statistical powers for the parametric tests they conducted. This suggests that the authors of these articles may not have sufficient knowledge and awareness of statistical power. There are several software packages such as G*Power, PASS, Power and Precision, and nQuery to calculate powers of statistical significance tests based on sample sizes, effect sizes, and the alpha levels. There are also different ways to assess statistical power. For instance, G*Power 3 allows for the calculation of the following five different types of power analysis: a priori power analyses, post hoc power analyses, compromise power analyses, sensitivity analyses, and criterion analyses (Faul et al., 2007). Currently, the use of statistical power analysis is absent in the national STEM education literature. However, given the accumulated body of information about statistical power and the diversity of computer programs available, there is not any reason to overlook statistical power when planning research studies and analyzing their findings.

4.8. Implications

Our findings demonstrate that the experimental STEM education articles published in the refereed Turkish journals from 2014 to 2020 suffer from serious methodological flaws. Thus, the methodological quality of these articles should remain a concern for the STEM education community including authors, journal editors, editorial board members, reviewers, practitioners, readers, and particularly for policymakers and curriculum developers who are responsible for developing and reforming national curricula in the Ministry of National Education.

To improve the methodological quality of STEM education articles, first, authors should develop some competence and awareness in experimental research designs. Universities or other institutions may design workshops and deliver some training to the authors to have them gain substantial expertise in experimental research methodology. In these workshops, several novel and fruitful approaches may be used to help the authors gain a more thorough understanding of experimental research designs. For instance, LaCosse et al.'s (2017) active-learning approach may be used. LaCosse et al. (2017) examined the impact of project-oriented active-learning techniques on psychology undergraduates' understanding of research methods and found that these techniques increased the participants' understanding significantly.

Scholarly journals are a primary means for disseminating research findings. Thus, to improve the rigor of STEM education research, editors, editorial board members, and reviewers may adopt clear and efficient quality guidelines or criteria such as the standards and procedures specified by the What Works Clearinghouse (2020a, 2020b), the Consolidated Standards of

Reporting Trials (Schulz et al., 2010), and the Study Design and Implementation Assessment Device (Valentine & Cooper, 2008). They may use these guidelines, standards, or criteria as requirements for the manuscripts submitted to the journals. Meanwhile, authors may use the coding framework developed in the current study and similar evaluation tools developed in previous studies such as the Checklist for the Rigor of Education-Experiment Designs (Sung et al., 2019) to check the experimental design quality of their manuscripts and remedy the deficiencies existing in their manuscripts before submission. Thus, the coding framework proposed in the present study may particularly serve experimental research authors as an effective self-checking and self-improvement tool.

Ultimately, using valid designs is significant for maintaining the sustainability and practicability of STEM education research. Using rigorous experimental designs will undoubtedly contribute to the development of theories and practices in the area of STEM education. On the other hand, the deficiencies in the experimental designs will be an obstacle to the sustainability of STEM education research. In addition, these deficiencies may mislead the STEM education community about the effectiveness of STEM education practices. In the current study, we found that the selected articles had considerable deficiencies in their research designs. Thus, policymakers and curriculum developers in the Ministry of National Education should be very cautious when using the findings of the STEM education articles published in the national journals and ruminate much on these findings before implementing STEM education curricula in K–12 schools. This is because initiating such reforms in educational environments requires too much time, energy, and resources and if STEM curricula do not lead to superior outcomes, contrary to the existent literature, then all the investment will be wasted.

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. Ethical approval for the current study was provided by Nevşehir Hacı Bektaş Veli University Ethics Committee (Date of Application: 31.03.2021, Number of Application: 2100014216).

Authorship Contribution Statement

The authors worked collaboratively in all phases of the manuscript preparation.

Orcid

Ramazan AVCU https://orcid.org/0000-0002-0149-5178
Seher AVCU https://orcid.org/0000-0003-4938-7325

REFERENCES

Ahern, K.J. (1999). Ten tips for reflexive bracketing. *Qualitative Health Research, 9*(3), 407–411. https://doi.org/10.1177/104973239900900309

Akgündüz, D., Aydeniz, M., Çakmakçı, G., Çavaş, B., Çorlu, M.S., Öner, T., & Özdemir, S. (2015). *STEM eğitimi Türkiye raporu: Günün modası mı yoksa gereksinim mi?* [A report on STEM education in Turkey: A provisional agenda or a necessity?]. Scala Press.

Aktemur, Ş. (2015). *Review of aviation research: A content analysis of articles published in the Collegiate Aviation Review, 2007–2012* [Unpublished master's thesis]. Florida Institute of Technology.

Aron, A., Coups, E.J., & Aron, E.N. (2019). *Statistics for the behavioral and social sciences: A brief course* (6th ed.). Pearson.

Ary, D., Jacobs, L.C., Sorensen, C.K., & Walker, D. (2014). *Introduction to research in education* (9th ed.). Wadsworth Cengage Learning.

Aydın Günbatar, S., & Tabar, V. (2019). Türkiye'de gerçekleştirilen STEM araştırmalarının içerik analizi [Content analysis of Science, Technology, Engineering and Mathematics (STEM) research conducted in Turkey]. *Yüzüncü Yıl University Journal of Education, 16*(1), 1054–1083. http://dx.doi.org/10.23891/efdyyu.2019.153

Baydaş, Ö., Küçük, S., Yılmaz, R. M., Aydemir, M., & Göktaş, Y. (2015). Educational technology research trends from 2002 to 2014. *Scientometrics, 105*, 709–725. https://doi.org/10.1007/s11192-015-1693-4

Borman, G.D., Slavin, R.E., Cheung, A., Chamberlain, A.M., Madden, N.A., & Chambers, B. (2005). Success for all: First-year results from the national randomized field trial. *Educational Evaluation and Policy Analysis, 27*(1), 1-22. https://doi.org/10.3102/01623737027001001

Brown, J. (2012). The current status of STEM education research. *Journal of STEM Education: Innovations and Research, 13*(5), 7–11.

Brown, J.R., & Dant, R.P. (2008). On what makes a significant contribution to the retailing literature. *Journal of Retailing, 84*(2), 131-135. https://doi.org/10.1016/j.jretai.2008.05.002

Campbell, D.T., & Boruch, R.F. (1975). Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In C. A. Bennett & A. A. Lumsdaine (Eds.), *Evaluation and experiment: Some critical issues in assessing social programs* (pp. 195–296). Academic Press.

Çavaş, P., Ayar, A., Bula Turuplu, S., & Gürcan, G. (2020). Türkiye'de STEM eğitimi üzerine yapılan araştırmaların durumu üzerine bir çalışma [A study on the status of STEM education research in Turkey]. *Yüzüncü Yıl University Journal of Education, 17*(1), 823–854. https://doi.org/10.33711/yyuefd.751853

Çevik, M. (2018). Impacts of the project based (PBL) science, technology, engineering and mathematics (STEM) education on academic achievement and career interests of vocational high school students. *Pegem Journal of Education and Instruction, 8*(2), 281–306. http://dx.doi.org/10.14527/pegegog.2018.012

Cheung, C.K., & Slavin, R.E. (2012). How features of educational technology applications affect student reading outcomes: A meta-analysis. *Educational Research Review, 7*(3), 198–215. https://doi.org/10.1016/j.edurev.2012.05.002

Coe, R.J. (2021). Effect size. In J. Arthur, M. Waring, R. Coe, & L. V. Hedges (Eds.), *Research methods and methodologies in education* (3rd ed., pp. 368–377). Sage.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education* (8th ed.). Routledge.

Creswell, J.W. (2015). *Educational research: Planning, conducting and evaluating quantitative and qualitative research* (5th ed.). Pearson Education.

Creswell, J.W., & Creswell, J.D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage.

D'Agostino, J. (2005). Measuring learning outcomes: Reliability and validity issues. In G. D. Phye, D. H. Robinson, & J. Levin (Eds.). *Empirical methods for evaluating educational interventions* (pp. 113–145). Elsevier Academic Press.

Daşdemir, İ., Cengiz, E., & Aksoy, G. (2018). Türkiye'de FeTeMM (STEM) eğitimi eğilim araştırması [An investigation of research trends in the field of STEM education in

Turkey]. *Yüzüncü Yıl University Journal of Education, 15*(1), 1161–1183. http://dx.doi.org/10.23891/efdyyu.2018.100

Dochy, F. (2006). A guide for writing scholarly articles or reviews for the Educational Research Review. Retrieved March 15, 2021, from https://www.elsevier.com/__data/promis_misc/edurevReviewPaperWriting.pdf

Duman, G., Orhon, G., & Gedik, N. (2015). Research trends in mobile assisted language learning from 2000 to 2012. *ReCALL, 27*(2), 197-216. https://doi.org/10.1017/S0958344014000287

Elmalı, Ş., & Balkan Kıyıcı, F. (2017). Türkiye'de yayınlanmış FeTeMM eğitimi ile ilgili çalışmaların incelenmesi [Review of STEM studies published in Turkey]. *Sakarya University Journal of Education, 7*(3), 684–696. https://doi.org/10.19126/suje.322791

Erdoğan, M., Marcinkowski, T., & Ok, A. (2009). Content analysis of selected features of K–8 environmental education research studies in Turkey, 1997–2007. *Environmental Education Research, 15*(5), 525–548. https://doi.org/10.1080/13504620903085776

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Finn, J.D., & Achilles, C.M. (1999). Tennessee's class size study: Findings, implications, misconceptions. *Educational Evaluation and Policy Analysis, 21*, 97–109. https://doi.org/10.3102/01623737021002097

Fraenkel, J.R. (1987). Toward improving research in social studies education. *Theory & Research in Social Education, 15*(3), 203-222. https://doi.org/10.1080/00933104.1987.10505546

Fraenkel, J.R., Wallen, N.E., & Hyun, H.H. (2012). *How to design and evaluate research in education* (8th ed.). McGraw-Hill.

Frey, B.B. (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation.* Sage.

Gall, J.P., Gall, M.D., & Borg, W.R. (2014). *Applying educational research: How to read, do, and use research to solve problems in practice* (6th ed.). Pearson Education.

Gall, M., Gall, J., & Borg, R. (2007). *Educational research: An introduction* (8th ed.). Pearson Education.

Gravetter, F., Wallnau, L., Forzano, L., & Witnauer, J. (2021). *Essentials of statistics for the behavioral sciences* (10 ed.). Cengage Learning.

Hair, J.F., Black, W.C., Babin, B. J., & Anderson, R.E. (2019). *Multivariate data analysis* (8th ed.). Cengage Learning.

Hanel, P.H., & Mehler, D.M. (2019). Beyond reporting statistical significance: Identifying informative effect sizes to improve scientific communication. *Public Understanding of Science, 28*(4), 468–485. https://doi.org/10.1177/0963662519834193

Hedges, L.V., & Schauer, J. (2018). Randomised trials in education in the USA. *Educational Research, 60*(3), 265–275. https://doi.org/10.1080/00131881.2018.1493350

Horton, P.B., McConney, A.A., Woods, A.L., Barry, K., Krout, H.L., & Doyle, B.K. (1993). A content analysis of research published in the Journal of Research in Science Teaching from 1985 through 1989. *Journal of Research in Science Teaching, 30*(8), 857–869. https://doi.org/10.1002/tea.3660300805

Horváth, I. (2016). Theory building in experimental design research. In P. Cash, T. Stankovic, & M. Storga (Eds.), *Experimental design research: Approaches, perspectives, applications* (pp. 209–231). Springer International Publishing.

Howell, D.C. (2017). *Fundamental statistics for the behavioral sciences* (9th ed.). Cengage Learning.

Johnson, R.B., & Christensen, L.B. (2020). *Educational research: Quantitative, qualitative, and mixed approaches* (7th ed.). Sage.

Kaya, A., & Ayar, M.C. (2020). Türkiye örnekleminde STEM eğitimi alanında yapılan çalışmaların içerik analizi [Content analysis of STEM education studies in Turkey]. *İstanbul Aydın University Journal of Education, 6*(2), 275–306.

Kirk, R.E. (2013). *Experimental design: Procedures for the behavioral sciences* (4th ed.). Sage Publications.

LaCosse, J., Ainsworth, S.E., Shepherd, M.A., Ent, M., Klein, K.M., Holland-Carter, L.A., Moss, J.H., Licht, M., & Licht, B. (2017). An active-learning approach to fostering understanding of research methods in large classes. *Teaching of Psychology, 44*(2), 117–123. https://doi.org/10.1177/0098628317692614

Li, Y., Wang, K., Xiao, Y., & Froyd, J. E. (2020). Research and trends in STEM education: A systematic review of journal publications. *International Journal of STEM Education, 7*, 1. https://doi.org/10.1186/s40594-020-00207-6

Lincoln, Y.S., & Guba, E. (1985). *Naturalistic inquiry.* Sage.

Miles, M.B., Huberman, M.A., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). Sage.

Mills, G.E., & Gay, L.R. (2016). *Educational research: Competencies for analysis and applications* (11th ed.). Pearson Education.

Mills, G.E., & Gay, L.R. (2019). *Educational research: Competencies for analysis and applications* (12th ed.). Pearson.

Ministry of National Education. (2016). *STEM eğitimi raporu* [STEM education report]. Innovation and Educational Technologies General Directorate.

Ministry of National Education. (2018). *Küresel bağlamda STEM yaklaşımları* [STEM approaches in a global context]. Innovation and Educational Technologies General Directorate.

Mizell, S., & Brown, S. (2016). The current status of STEM education research 2013-2015. *Journal of STEM Education*, 17(4), 52–56.

Nelson, J.L., & Shaver, J.P. (1985). On research in social education. In W. B. Stanley (Ed.), *Review of research in social studies education: 1976-1983* (pp. 401–433). National Council for the Social Studies.

Orne, M.T. (1981). The why and how of a contribution to the literature: A brief communication. *International Journal of Clinical and Experimental Hypnosis, 29*(1), 1–4. https://doi.org/10.1080/00207148108409137

Özcan, H., & Koca, E. (2019). The impact of teaching the subject "pressure" with STEM approach on the academic achievements of the secondary school 7th grade students and their attitudes towards STEM. *Education and Science*, *44*(198), 201–227. http://dx.doi.org/10.15390/EB.2019.7902

Pagano, R.R. (2013). *Understanding statistics in the behavioral sciences* (10th ed.). Wadsworth Cengage Learning.

Pallant, J. (2016). *SPSS survival manual: A step by step guide to data analysis using the SPSS program* (6th ed.). Open University Press.

Plano Clark, V.L., & Creswell, J.W. (2015). *Understanding research: A consumer's guide* (2nd ed.). Pearson Education.

Privitera, G.J. (2019). *Essential statistics for the behavioral sciences* (2nd ed.). Sage.

Randolph, J.J., Griffin, A.E., Zeiger, S.R., Falbe, K.N., Freeman, N.A., Taylor, B.E., Westbrook, A.F., Lico, C.C., Cristy, N. S., Sprull, N. M., Holt, C., Smith, K., & McAnespie, H. (2013). A methodological review of the articles published in Georgia Educational Researcher from 2003-2010. *Georgia Educational Researcher Online Edition, 10*(1), Article 1. https://doi.org/10.20429/ger.2013.100101

Schulz, K.F., Altman, D.G., & Moher, D. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *Trials, 11*(1), Article 32. https://doi.org/10.1136/bmj.c332

Schweinhart, L.J., Barnes, H.V., & Weikart, D.P. (1993). *Significant benefits: The High/Scope Perry Preschool Study through age 27*. High/Scope Press.

Shapiro, G.M. (2008). Sample size. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (Vol. 2, pp. 781–783). Sage.

Shaver, J.P., & Norton, R.S. (1980). Populations, samples, randomness, and replication in two social studies journals. *Theory & Research in Social Education, 8*(2), 1–10. https://doi.org/10.1080/00933104.1980.10506078

Shukla, A. (2017). Literature review: An oblivious yet grounding task of research. *Management Insight, 13*(1), 7–15. https://doi.org/10.21844/mijia.v13i01.8363

Slavin, R.E. (2008). Perspectives on evidence-based research in education—What works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37*(1), 5–14. https://doi.org/10.3102/0013189X08314117

Stebbins, L.F. (2006). *Student guide to research in the digital age: How to locate and evaluate information sources*. Libraries Unlimited.

Sung, Y.T., Lee, H.Y., Yang, J.M., & Chang, K.E. (2019). The quality of experimental designs in mobile learning research: A systemic review and self-improvement tool. *Educational Research Review, 28,* 100279. https://doi.org/10.1016/j.edurev.2019.05.001

Tabachnick, B.G., & Fidell, L.S. (2019). *Using multivariate statistics* (7th ed.). Pearson.

United States Department of Education. (2020). *ED delivers historic investment in STEM.* https://content.govdelivery.com/accounts/USED/bulletins/2ad85c3

Valentine, J.C., & Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The study design and implementation assessment device (Study DIAD). *Psychological Methods, 13*(2), 130–149. https://doi.org/10.1037/1082-989X.13.2.130

Wallen, N.E., & Fraenkel, J.R. (1988). An analysis of social studies research over an eight year period. *Theory & Research in Social Education, 16*(1), 1-22. https://doi.org/10.1080/00933104.1988.10505553

Warner, R.M. (2013). *Applied statistics: From bivariate through multivariate techniques* (2nd ed.). Sage.

What Works Clearinghouse. (2020a). *Standards handbook, version 4.1.* U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Standards-Handbook-v4-1-508.pdf

What Works Clearinghouse. (2020b). *Procedures handbook, version 4.1.* U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Procedures-Handbook-v4-1-508.pdf

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 2, 319–336
https://doi.org/10.21449/ijate.946978
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

An Analysis of Differential Bundle Functioning in Multidimensional Tests Using the SIBTEST Procedure

Didem Ozdogan[1,*], **Hulya Kelecioglu**[2]

[1]Istanbul Kultur University, Faculty of Education, Department of Educational Sciences, İstanbul
[2]Hacettepe University, Faculty of Education, Department of Measurement and Evaluation in Education, Ankara

ARTICLE HISTORY
Received: June 02, 2021
Revised: Jan. 06, 2022
Accepted: Jan. 24, 2022

Keywords:
Differential item functioning,
Differential bundle functioning,
Multidimensionality,
SIBTEST,
Type 1 error,
Power rate.

Abstract: This study aims to analyze the differential bundle functioning in multi-dimensional tests with a specific purpose to detect this effect through differentiating the location of the item with DIF in the test, the correlation between the dimensions, the sample size, and the ratio of reference to focal group size. The first 10 items of the test that is comprised of 30 items were acknowledged as the bundle. The data in line with the parameters were generated via SAS program as two categories (1-0) and multidimensional through an extended 2PL model. Differential bundle functioning was detected via the SIBTEST procedure. The results of the study were interpreted according to the criteria of the power rate and the type I error. When the results were reviewed, the analysis of the bundle revealed that the more the correlation between the two dimensions increased, relatively the less the power rates became. It was observed that the power rates, which were obtained according to two different sample sizes in the study, increased as the sample size increased. Another result as to the SIBTEST's power for detecting DIF was the highest when the ratio of reference to focal group size was equal. According to the results of the type I error rate, the error rate was observed to be relatively decreasing as the correlation between the dimensions increased and it was observed to be increasing as the sample size increased. Also, the highest error rate was obtained when the ratio of the samples was equal.

1. INTRODUCTION

In Classical Test Theory (CTT) and Item Response Theory (IRT), which are used to construct educational and psychological tests and to interpret scores, the assumption is that the individual has a characteristic or ability below the test performance. Uni-dimensional models included in IRT out of these theories comprise a single ability parameter that expresses the location of an individual on the tested characteristic. However, interactions between individuals and items are not generally easy enough to be expressed with these models. Answering a test item or solving a problem generally requires the use of multiple skills and abilities. For this reason, although one-dimensional IRT models are useful under certain conditions, IRT models which reflect the interaction between individuals and test items more accurately are needed. Such IRT models describe the interaction between individuals and characteristics of test items with multiple abil-

*CONTACT: Didem OZDOGAN ✉ reyhandidem@gmail.com Istanbul Kultur University, Faculty of Education, Department of Educational Sciences, İstanbul, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

ities. These models are defined as multidimensional IRT models for having more than one ability parameter for an individual (Reckase, 2009). The fact that only one ability is required for a test item to be answered is not well suited to actual test situations. For example, it is probable to encounter a two-dimensional structure in which a dimension in a mathematical problem reflects mathematical competency whereas the other reflects reading competency. In the related studies conducted it is stated that, besides the primary factors under the test performance of the individual, there is at least one secondary factor and that the tests generally show a multidimensional structure (Camilli, 1992).

Dimension can be defined as a characteristic that affects the probability of answering an item correctly. The main structure the test aims to measure is called as the primary dimension of the test. Since the cause of DIF is defined as the presence of multidimensionality in items showing DIF, such items measure at least one more dimension in addition to the primary dimension the items aim to measure (Cronbach, 1990; Wiley, 1990). In other words, when an item measures more than one dimension and the groups differentiate on the structure or structures that are not primarily measured by the item, the item shows DIF. If the groups do not differentiate on the dimension or dimensions that are not primarily measured, no DIF is detected even if the data are multidimensional (Ackerman, 1992a).

Other dimensions considered as the cause of DIF are called as secondary dimension. The secondary dimensions are the factors that may or may not be related to the dominant dimension. Each secondary dimension intended to be measured is called as the *auxiliary*; whereas each secondary dimension that is not intended to be measured is called as the *nuisance*. DIF caused by the auxiliary dimension is called as benign DIF (which refers to benign effect) because the test also aims to measure the auxiliary dimension. On the other hand, DIF caused by nuisance dimension is called as adverse DIF (which refers to bias) because the item is less valid for one group of individuals than the other for evaluating individual differences on the dimensions measured (Roussos & Stout, 1996).

Shealy and Stout (1993) suggested that DIF occurs due to the presence of two factors; namely,

(1) The item is sensitive not only to the structure θ, which the item aims to measure, but also to a secondary η structure;
(2) At a constant value of θ, there is a difference between the conditional distributions of groups of interest on the η structure.

In general, studies on the detection of DIF are carried out in two stages:
(1) The statistical determination of whether an item provides an advantage for a particular group;
(2) Evaluation of potentially biased items by substantive analysis methods to identify the source of DIF.

Since statistical methods provide limited information in detecting the sources of DIF, new methods have been developed. Roussos and Stout (1996) pointed to the failure of interpretation of substantive DIF analyses followed by statistical methods and Engelhard, Hansche and Rutledge (1990) pointed out in many cases the incompatibility of expert decisions in substantive analyses carried out after statistical methods. Roussos and Stout (1996) proposed a two-step approach by correlating the multidimensional model developed by Shelay and Stout (1993) for DIF to eliminate the inconsistency between statistical and substantive analyses. The first stage of this approach which is called as the multidimensionality-based DIF analysis paradigm is the substantive analysis in which DIF hypotheses are formed and the second stage is the statistical testing of DIF hypotheses. In other words, Shelay and Stout (1993) suggested reversing the process for DIF analyses; they thus stated that reliable and supportive explanations can be obtained about why DIF occurs. Some researchers argue that this problem between statistical and substantive analyses arises from the nature of analyzing individual items in DIF analyses and

that more information can be obtained by analyzing the items in groups rather than analyzing a single item at a time (Boughton et al., 2000; Douglas et al., 1996). Taking this into consideration, Douglas et al. (1996) introduced the DIF in the bundle concept and the applications of Differential Bundle Functioning (DBF) in order to identify the details of DIF.

DBF means that different groups of equal ability levels differ in their probability of correctly answering a bundle. A bundle is a dimensionally homogeneous item set that is not necessarily adjacent or related to a common text (Douglas et al., 1996). Many tests may appear to consist of independent items. However, when carefully analyzed, it might be observed that these items have common topics or similar content. Such items may be found scattered in the test, but they cause the DIF to increase at bundle or test level.

Gierl et al (2001) found that DIF-related properties may be more obvious in a multi-item pattern than a single item. Douglas et al. (1996) stated that the amount of DIF in small amounts at item level (at A-level) could not be statistically detected but that this amount can be detected by the DIF procedure in a bundle. Similarly, Nandakumar (1993) stated that the analysis of a bundle or a group of items would yield stronger results than a single item analysis at a time. Consequently, detecting the potential sources of DIF by identifying groups of biased items by Differential Bundle Functioning (DBF) procedure can provide significant contributions to assessment and test development processes in education (Ross, 2008).

A bundle is formed from items intended to measure a primary dimension (e.g. vocabulary) and a secondary dimension to be measured by the test (e.g. mathematical ability). A premise, which explains why one of the two groups of equal ability levels is more advantageous in the bundle, is developed. The hypothesis developed suggests that one of the groups in a bundle is more advantageous on the secondary dimension than the other group compared (Douglas et al., 1996). Since the bundles are formed based on a hypothesis, DBF analysis can be considered as confirmatory analysis. On the basis of DBF analysis, there is the assumption that a test that measures a particular trait, skill or ability consists of small bundles (Ross, 2008).

Multidimensionality is a concept related to the interaction between an item and its ability. When a test includes items consisting of more than one ability or combination of abilities, several problems might occur if it is not aimed to measure all abilities comprising the test. Ordering individuals accurately according to their abilities primarily requires a valid and reliable measuring. Considering that the items with DIF might weaken validity and that DIF in multidimensional tests is based on the ability differences of individuals on secondary dimension, DIF studies and addressing DIF and multidimensionality together become more important. No study has been found in our country on the concept of DBF and there are few studies outside of our country. The different condition analyzed in this study, different from the literature, is how the location of the items showing DIF in the test affects DBF. In this study, DBF and multidimensionality were analyzed together and tested under various conditions. It is believed that the results obtained would contribute to other studies aiming at detecting the source of DIF.

The purpose of this study is to analyze the DBF concept in multidimensional tests under various conditions. Variables included in the study were the correlations between the dimensions (0.10, 0.45, 0.80), the sample size (2000, 5000), and the ratio of reference to focal group size (1/3, 1/1, 3/1). All these variables were analyzed under the following conditions according to whether the items in the bundle and outside of the bundle were items with DIF.

1) When all items in the bundle show DIF,

SIBTEST power rates according to the conditions are as follows:

a) There are items showing DIF outside of the bundle and
b) There is no item showing DIF outside of the bundle.

2) When the items which show and do not show DIF are present together in the bundle,

SIBTEST power rates according to the conditions are as follows:

a) There are items showing DIF outside of the bundle and
b) There is no item showing DIF outside of the bundle.
3) When there are no items showing DIF in the bundle
a) For the condition in which there are items showing DIF outside of the bundle, SIBTEST Type I error is analyzed.
b) For the condition in which there is no item showing DIF outside of the bundle, SIBTEST Type I error is analyzed.

2. METHOD

2.1. Research Type

In this study, type I error and power rate changes in the bundle were analyzed by differentiating the variables; namely, the location of the item showing DIF in multidimensional tests, the correlations between the dimensions, the sample size, and the ratio of reference to focal group size. Therefore, this study was considered as a simulation study. In this study, the DIF was analyzed with a different approach which is believed to contribute to the theory. In this sense, it can be suggested that this study is a basic research (Karasar, 2020).

2.2. Data of the Study

Within this study, the DBF concept in multidimensional tests was analyzed by considering various conditions in which individuals who apply the tests might encounter in actual test situations. Since it was difficult to realize these conditions in an actual data set at the same time, simulation data were used.

A 30-item test was formed in the study and the first 10 items of this test were considered as the bundle. Type I error and power rates in the bundle were analyzed within the context of the location of the item showing DIF in multidimensional tests, the correlations between the two dimensions, the sample size, and the ratio of reference to focal group size. Item parameters used for generating the research data were generated in ITEMGEN (Ackerman, 1994) program. The item parameters which had equal scattering in this program were formed between the range of the angle values determined by the researcher. The angle values (α) are about the item discriminations. The value of α can vary from 0° to 90° based on the degree to which an item measures each trait. For the case of two dimensions, if an item measures only the first ability, θ_1, then the item direction (α) is 0°. If an item measures only the second ability, θ_2, then the item direction (α) is 90°. When α = 45°, an item measures two abilities (θ_1 and θ_2) equally. Therefore, calculating an item's angular direction (α) provides information about what items are really measuring (Ross, 2007).

When generating item parameters in this study, suitable angle values were determined for items that show and do not show DIF in accordance with the research conditions. It is generally considered in the conditions when the angle value for the items exceeds 20^0. If the secondary dimension is an unintended dimension, a threat to validity will arise (Ackerman, Gierl & Walker, 2003).

The dimension measuring the θ_1 ability in the study was determined as the dimension intended to be measured by the test, whereas the dimension measuring the θ_2 ability was determined as the dimension which is not intended to be measured by the test. The angle values for the items which primarily measure the θ_1 ability and do not show DIF were adjusted to alternate between the range of 5^0-20^0, whereas the angle values for items which primarily measure the θ_2 ability and show DIF were adjusted to alternate between the range of 70^0-85^0. The discrimination parameter of the model is a measure of the differential capability of an item. An item is considered valuable if it well discriminates subjects with ability levels in a range of interest for

the exam (UI Hassan & Miller, 2020). Multidimensional discrimination (MDISC) parameter was generated to alternate between the range of 0.8 and 1.8, whereas item difficulty parameter (d), which is a measure of the ease of the item, was generated to alternate between the range of -2 and 2. The same item parameters were used for reference and focus groups.

Item discrimination power is related to a certain angle value which the item has in latent ability space (Reckase & McKinley, 1991). Calculating this angle provides information about what the item measures. For multidimensional items, this angle can be calculated in terms of the latent axes. Accordingly, α angle alternates between 0^0-90^0. The angle between the discriminant vector and x axis (which refers to the primary dimension which the item aims to measure) can be calculated by Equation 1, which is a simplified version of the formula proposed by Reckase and McKinley (1991) (Ross, 2008).

$$\alpha = tan^{-1}\frac{a_2}{a_1} \quad (1)$$

a_1: The item discrimination power for the primary dimension;

a_2: The item discrimination power for the secondary dimension.

In this study, confirmatory factor analysis was performed in analyzing the accuracy of the structure formed. Also, α angles were checked by using the formula in Equation 1 and parameters were generated in accordance with the location of the item with DIF in the test.

2.3. Simulation Conditions

In accordance with the purpose of the study, simulation conditions determined in the DBF analysis in multidimensional tests are given below. In all conditions, test length, number of items in the bundle, and ability distributions of focus and reference groups were kept constant.

2.3.1. *Test length*

In this study, the length of the test was set at 30 items to represent a mid-length test.

2.3.2. *Number of items in the bundle*

There is one bundle in the test and the first 10 items of the test form the bundle.

2.3.3. *Ability distributions of the groups*

In the study, a test with the items with DIF was formed as θ_1 and θ_2 to measure two dimensions. The characteristic that was intended to be measured by the test was called θ_1, whereas the characteristic that causes the item to show DIF was called θ_2. Accordingly, the data were generated in a way that the items showing DIF would primarily measure the θ_2 dimension, whereas the items which do not show DIF would primarily measure the θ_1 dimension.

The ability distributions of the reference and focal groups for the first dimension were equal and the standard normal distribution was $\theta \sim (N_F(0,1))$ and $\theta \sim (N_R(0,1))$; for the second dimension, a difference of 0.50 was created between the ratio of reference to focal group size, considering the studies in the literature (Ross, 2008; Oshima & Miller, 1992; Russell, 2005, Walker & Şahin, 2016). In the second dimension, the distribution of focus and reference groups was determined as the non-standard normal distribution $\theta(\sim(N_F(-0.25,1))$ and $\theta \sim (N_R(0.25,1))$.

2.3.4. *Test location of item with DIF*

Six conditions were identified for this situation. In all conditions, the items which do not show DIF primarily measure the θ_1 dimension, whereas the items showing DIF primarily measure the θ_2 dimension.

1) All the items in the bundle are items with DIF and there are items showing DIF outside of the bundle: 10 items of the bundle have DIF and 5 test items outside of the bundle have DIF, whereas 15 items do not have DIF.
2) All items in the bundle are items with DIF and there are no items with DIF outside of the bundle: In this condition, 10 items forming the bundle show DIF, whereas 20 items outside of the bundle do not show DIF.
3) In the bundle, there are items with and without DIF, and outside of the bundle, there are items with DIF: 5 items in the bundle have DIF, while 5 items do not have DIF; 5 test items outside of the bundle have DIF, whereas 15 items do not.
4) In the bundle, there are items with and without DIF, and outside of the bundle, there are no items with DIF: 5 items in the bundle have DIF, while 5 items do not have DIF and 20 items outside of the bundle do not show DIF.
5) In the bundle, there are no items with DIF, and outside of the bundle there are items with DIF: 10 items of the bundle do not have DIF and 5 test items outside of the bundle have DIF, whereas 15 items do not.
6) There are no items with DIF in the bundle and outside of the bundle: 10 items forming the bundle do not have DIF; 20 items outside of the bundle do not have DIF.

2.3.5. *The correlation between the primary and secondary dimensions* ($r_{\theta 1 \theta 2}$)

One of the variables analyzed in this study is the effect of the correlations between the primary and secondary dimensions. For this purpose, the correlations between the dimensions were detected as 0.10, 0.45 and 0.80, to represent low, medium, and high correlation values, respectively.

2.3.6. *The sample size*

It is suggested in the literature to work with at least 1000-people samples in multi-dimensional structures (Bolt & Lall, 2003; Yao & Boughton, 2007). Ackerman (1994) stated that multidimensional calibrations require at least 2000 samples. When studies on the subject were reviewed, it was observed that the sample size generally varied between 500 and 5000. Two different sample sizes, 2000 and 5000, were determined for this study.

2.3.7. *The ratio of reference to focal group size (R/F)*

In DIF detection studies conducted, the ratio of reference to focal group size is generally preferred to be equal or close to each other. However, it can also be observed that the ratio of reference to focal group size differs from each other in actual test situations.
Shealy and Stout (1993) stated that at least 250 individuals in each group are required for the SIBTEST procedure. In this study, the reference and focus group sizes were analyzed by differentiating R/F rates, determined as 1/3 (500/1500; 1250/3750), 1/1 (1000/1000; 2500/2500), and 3/1 (1500/500; 3750/1250).

The variables and simulation conditions analyzed in the study are given in Table 1. As seen in Table 1, 108 conditions in total were analyzed: six for the location of the items with DIF, three for the correlation between the dimensions, two for the sample size, and three for the ratio of reference to focal group size (6x3x2x3). Each condition in the study was repeated 100 times. In the literature, it was stated that at least 25 replications are required for simulation studies (Harwell et al.,1996). In this study, a total of 10800 data sets were obtained with 100 replications done for each condition.

Table 1. *The variables and simulation conditions analyzed in the study.*

Variables	Simulation Conditions
Location of the DIF Item	1) All of the items in the bundle are items with DIF and there are items with DIF outside of the bundle 2) All of the items in the bundle are items with DIF and there are no items with DIF outside of the bundle 3) There are items with and without DIF in the bundle and there are items with DIF outside of the bundle. 4) There are items with and without DIF in the bundle and there are no items with DIF outside of the bundle. 5) There are no items with DIF in the bundle and there are items with DIF outside of the bundle 6) There are no items with DIF in the bundle and outside of the bundle
Correlation between dimensions ($r_{\theta 1 \theta 2}$)	1) 0.10 2) 0.45 3) 0.80
Sample size	1) 2000 2) 5000
The ratio of the samples (R/F)	1) 1/3 2) 1/1 3) 3/1

2.4. Analysis and Evaluation Criteria of the Data

In line with the aims of the study, the item parameters were obtained in the ITEMGEN program regarding the location of the items with DIF in the test. According to these parameters, the data were generated in SAS program in accordance with the extended two-parameter logistic model for multidimensionality and two categories (1-0).

Differential Bundle Functioning was identified using the SIBTEST procedure. SIBTEST was developed as an extension of the multidimensional DIF model developed by Shealy and Stout (1993) and is a non-parametric procedure that models the relationship between the latency and item performance measured by the test. After completing the DBF analyses for 108 conditions addressed in the study by using SIBTEST program, SAS program was used to calculate the Type I error and power rates. The effect of the conditions in the study for detecting the DBF was evaluated by the Type I error and power rate criterion. Power rate gives a measure of how accurate the DIF is detected for each item and the bundle using the SIBTEST procedure. It is generally expected that the power rates obtained are equal to and greater than 0.80. However, Type 1 error occurs when the DIF is detected in the item and bundle that do not contain DIF. Generally, in DIF studies, the criterion of type 1 error is 0.05 nominal alpha value (Ross, 2007; Atalay Kabasakal et.al.)

In this study, a variance analysis was also performed to detect how the type I error and power rates obtained regarding the differentiating location of the item with DIF in the test changed according to the conditions studied.

3. FINDINGS

The findings obtained from the analysis of the data generated according to the conditions specified in this section are presented in the context of the location of the item showing DIF in the bundle.

3.1. The Situation When All the Items in the Bundle Show DIF

The situation in which all the items in the bundle show DIF was analyzed in two conditions: 1) there are items with DIF outside of the bundle and 2) there are no items with DIF outside of the bundle. The power of the test was calculated regarding the DIF results according to the correlation between the dimensions, the sample size, and the ratio of reference to focal group size obtained in both conditions. The results are demonstrated in Table 2.

Table 2. *Power rates for the conditions in which all the items in the bundle show DIF.*

Condition			OUT DIF^{+}	OUT DIF^{-}
Correlation between dimensions ($r_{\theta 1 \theta 2}$)	Sample Size	Sample Ratio (R/F)	Power Rate	Power Rate
0.10	2000	1/3 (500/1500)	1	1
		1/1 (1000/1000)	1	1
		3/1 (1500/500)	1	1
	5000	1/3 (1250/3750)	1	1
		1/1 (2500/2500)	1	1
		3/1 (3750/1250)	1	1
0.45	2000	1/3 (500/1500)	1	1
		1/1 (1000/1000)	1	1
		3/1 (1500/500)	1	1
	5000	1/3 (1250/3750)	1	1
		1/1 (2500/2500)	1	1
		3/1 (3750/1250)	1	1
0.80	2000	1/3 (500/1500)	1	1
		1/1 (1000/1000)	1	1
		3/1 (1500/500)	1	1
	5000	1/3 (1250/3750)	1	1
		1/1 (2500/2500)	1	1
		3/1 (3750/1250)	1	1

Notes. OUT DIF^{+}: There are items with DIF outside of the bundle, OUT DIF^{-}: There is no DIF outside of the bundle.

In two different conditions (there are items and there are no items which contain DIF outside of the bundle) which were analyzed in the situation that all of the items in the bundle show DIF, SIBTEST detected the DIF in the bundle as 100% correct for all conditions. For this reason, the power rates of the bundle did not differ according to the variables analyzed in the study. All power rates obtained were above the acknowledged limit.

3.2. The Situation When the Items which Show DIF and do not Show DIF are Present Together in the Bundle

The situation when the items which show DIF and do not show DIF are present together in the bundle was analyzed in two conditions: 1) There are items showing DIF outside of the bundle and 2) There is no item showing DIF outside of the bundle. The power of the test was calculated regarding the DIF results according to the correlation between the dimensions, the sample size, and the ratio of reference to focal group size obtained in both conditions. The results are demonstrated in Table 3.

Table 3. *Power rates for the conditions in which the items that show DIF and do not show DIF are present together in the bundle.*

Conditions			OUT DIF^{+}	OUT DIF^{-}
Correlation between dimensions ($r_{\theta 1 \theta 2}$)	Sample Size	Sample Ratio (R/F)	Power Rate	Power Rate
0.10	2000	1/3 (500/1500)	0.33	1
		1/1 (1000/1000)	0.45	1
		3/1 (1500/500)	0.41	0.99
	5000	1/3 (1250/3750)	0.62	1
		1/1 (2500/2500)	0.84	1
		3/1 (3750/1250)	0.66	1
0.45	2000	1/3 (500/1500)	0.37	0.97
		1/1 (1000/1000)	0.30	1
		3/1 (1500/500)	0.35	0.97
	5000	1/3 (1250/3750)	0.58	1
		1/1 (2500/2500)	0.76	1
		3/1 (3750/1250)	0.69	1
0.80	2000	1/3 (500/1500)	0.26	0.97
		1/1 (1000/1000)	0.33	0.99
		3/1 (1500/500)	0.28	0.96
	5000	1/3 (1250/3750)	0.57	1
		1/1 (2500/2500)	0.74	1
		3/1 (3750/1250)	0.54	1

Notes. OUT DIF^{+}: There are items with DIF outside of the bundle, OUT DIF^{-}: There is no DIF outside of the bundle.

1) It was observed that the power rates obtained from SIBTEST varied between 0.26 and 0.84 when the items which show DIF and do not show DIF were present together in the bundle, and there are items with DIF outside of the bundle. The minimum power rate (0.26) was observed in the condition when the correlation between the two dimensions was 0.80, sample size was 2000, and sample ratio was 1/3. The largest power rate (0.84) was observed in the condition in which the correlation between two dimensions was 0.10, the sample size was 5000, and the ratio of reference to focal group size was 1/1.

In this condition analyzed regarding the differentiating location of the item with DIF, it was observed that the power rates obtained were detected below the acknowledged limit for other conditions except for one condition.

According to the variables analyzed in the study:

- The average power rates obtained from the bundle at different correlations between the dimensions were 0.55 when $r_{\theta 1 \theta 2}$ =0.10, 0.51 and when $r_{\theta 1 \theta 2}$ =0.45, and 0.45 when $r_{\theta 1 \theta 2}$ =0.80. It was observed that the DIF detection power of the SIBTEST in the bundle decreased as the correlation between the dimensions increased. In other words, the DIF in the bundle was more accurately detected using the SIBTEST method in the condition when the correlation between the dimensions was minimum.
- When the average power rates of the bundle at different sample sizes were examined, the results were obtained as 0.34 when the sample size was N=2000 and 0.67 and when the sample size was N=5000. DIF detection power of the SIBTEST in the bundle increased as the sample size increased. Using the SIBTEST procedure, the DIF in the bundle was more accurately detected in the large sample.

• When the average power rates of the bundle were analyzed in terms of the ratio of reference to focal group size, the results were obtained as 0.45 when R/F: 1/3; 0.57 and when R/F: 1/1 and 0.49 when R/F: 3/1. DIF detection power of the SIBTEST in the bundle was higher when the ratio of the reference and focal group size was equal. Using the SIBTEST procedure, the DIF was more accurately detected in the condition when the ratio of reference to focal group size was equal.

2) When the items which show DIF and do not show DIF were present together in the bundle and there was no item with DIF outside of the bundle, it was observed that power rates obtained from the SIBTEST were relatively lower compared to the condition when all of the items in the bundle were items with DIF. The power rates obtained varied between 0.96 and 1. The minimum power rate (0.96) was observed in the condition when the correlation between the two dimensions was 0.80, the sample size was 2000, and sample ratio was 3/1. The maximum power rate (1) was observed in all conditions when the sample size was 5000 and under some conditions when the sample size was 2000.

In this condition analyzed according to the differentiating location of the item with DIF in the test, the obtained power rates were above the acknowledged limit.

According to the variables analyzed in the study:

• The average power rates obtained from the bundle at different correlations between the dimensions were 1 when $r_{\theta 1 \theta 2}$ =0.10; 0.99 and when $r_{\theta 1 \theta 2}$ =0.45, and 0.99 when $r_{\theta 1 \theta 2}$ =0.80. It was observed that the highest DIF detection power of the SIBTEST in the bundle was obtained in the condition when the correlation between the dimensions was the lowest. In other words, the DIF in the bundle was most accurately detected using the SIBTEST procedure in the condition when the correlation between dimensions was the lowest.

• When the average power rates of the bundle at different sample sizes were analyzed, the results were obtained as 0.98 when the sample size was N=2000 and 1 when the sample size was N=5000. DIF detection power of the SIBTEST in the bundle increased as the sample size increased. Using the SIBTEST procedure, the DIF in the bundle was more accurately detected in the large sample.

• When the average power rates of the bundle were analyzed in terms of the ratio of reference to focal group size, the results were obtained as 0.99 when R/F: 1/3; 1 when R/F: 1/1 and 0.99 when R/F: 3/1. DIF detection power of the SIBTEST in the bundle was relatively higher when the ratio of the reference to focal group size was equal. Using the SIBTEST procedure, DIF was more accurately detected in the condition when the ratio of reference to focal group was equal.

3.3. The Situation When No Items Show DIF in the Bundle

The situation in which no items in the bundle show DIF was analyzed in two conditions: 1) there are items with DIF outside of the bundle and 2) there are no items with DIF outside of the bundle. The Type I Error of the SIBTEST was calculated according to the correlation between the dimensions, the sample size, and the ratio of reference to focal group size obtained in both conditions. The results are demonstrated in Table 4.

Table 4. *Error rates for the conditions in which there are no items showing DIF in the bundle.*

Conditions			OUT DIF^{+}	OUT DIF^{-}
Correlation between dimensions ($r_{\theta 1 \theta 2}$)	Sample Size	Sample Ratio (R/F)	Error Rate	Error Rate
0.10	2000	1/3 (500/1500)	0.33	0.28
		1/1 (1000/1000)	0.46	0.42
		3/1 (1500/500)	0.39	0.35
	5000	1/3 (1250/3750)	0.65	0.74
		1/1 (2500/2500)	0.79	0.80
		3/1 (3750/1250)	0.69	0.62
0.45	2000	1/3 (500/1500)	0.32	0.34
		1/1 (1000/1000)	0.37	0.38
		3/1 (1500/500)	0.36	0.36
	5000	1/3 (1250/3750)	0.64	0.67
		1/1 (2500/2500)	0.66	0.77
		3/1 (3750/1250)	0.53	0.63
0.80	2000	1/3 (500/1500)	0.35	0.37
		1/1 (1000/1000)	0.34	0.31
		3/1 (1500/500)	0.34	0.28
	5000	1/3 (1250/3750)	0.57	0.66
		1/1 (2500/2500)	0.67	0.69
		3/1 (3750/1250)	0.55	0.62

Notes. OUT DIF^{+}: There are items with DIF outside of the bundle, OUT DIF^{-}: There is no DIF outside of the bundle.

1) It was observed that Type I errors in the bundle varied between 0.32 and 0.79 when there are no items which show DIF in the bundle and when there are items which show DIF outside of the bundle. The lowest error rate (0.32) was observed in the condition when the correlation between the two dimensions was 0.45, sample size was 2000, and sample ratio was 1/3. The largest error rate (0.79) was observed in the condition when the correlation between the two dimensions was 0.10, sample size was 5000, and the ratio of reference to focal group size was 1/1.

It was observed that the error rates obtained in this condition were significantly higher than the nominal alpha level (0.05). It is thought that this situation is caused by the fact that the DIF levels increase when the items are analyzed in a bundle, whereas the items do not show DIF when each item is analyzed individually.

According to the variables analyzed in this study:

- The average error rates obtained from the bundle at different correlations between the dimensions were 0.55 when $r_{\theta 1 \theta 2}$ =0.10; 0.48 and when $r_{\theta 1 \theta 2}$ =0.45 and 0.47 when $r_{\theta 1 \theta 2}$ =0.80. It was observed that the average Type I error rates in the bundle relatively decreased as the correlation between the dimensions increased. In other words, Type I error rate decreased as the test approached unidimensionality.
- When the average Type I error rates of the bundle at different sample sizes were analyzed, the results were obtained as 0.36 when the sample size was N=2000 and 0.64 when the sample size was N=5000. It was observed that the error rates of the bundle increased as the sample size increased. A lower error rate was obtained in the small sample when detecting DIF in the bundle by using the SIBTEST procedure.
- When the average Type I error rates of the bundle were examined in terms of the ratio of reference to focal group size, the results were obtained as 0.48 when R/F: 1/3; 0.55 and when

R/F: 1/1 and 0.48 when R/F: 3/1. It was observed that the error rates obtained were higher when the ratio of reference to focal group was equal.

2) When there were no items which show DIF in the bundle and the test, it was observed that Type 1 errors of the bundle varied between 0.28 and 0.80. The lowest error rate (0.28) was observed in two different conditions when the sample size was 2000 and the correlation between the two dimensions was 0.10 and the sample ratio was 1/3, and the correlation between the two dimensions was 0.80 and the sample ratio was 3/1. The highest error rate (0.80) was observed in the condition when the correlation between the two dimensions was 0.10 and the sample size was 5000, and the ratio of reference to focal group size was 1/1.

In this condition analyzed regarding the differentiating location of the item with DIF in the test, it was observed that the error rates obtained were significantly higher than the nominal alpha level (0.05). It is thought as in the former sub-problem that this situation was caused by the fact that the DIF levels increase when the items are analyzed in a bundle, whereas the items do not show DIF when analyzed individually.

According to the variables analyzed in the study:

- The average error rates obtained from the bundle at different correlations between the dimensions were 0.53 when $r_{\theta 1 \theta 2}$=0.10; 0.52 and when $r_{\theta 1 \theta 2}$ =0.45 and 0.48 when $r_{\theta 1 \theta 2}$ =0.80. It was observed that the average Type I error rates in the bundle relatively decreased as the correlation between the dimensions increased. In other words, Type I error rate of the bundle decreased as the test approached unidimensionality.

- When the average Type I error rates of the bundle at different sample sizes were analyzed, the results were obtained as 0.34 when the sample size was N=2000 and 0.69 when the sample size was N=5000. It was observed that the average error rates of the bundle increased as the sample size increased. A lower error rate was obtained in the small sample when detecting the DIF in the bundle by using the SIBTEST procedure.

- When the average Type I error rates of the bundle were analyzed in terms of the ratio of reference to focal group size, the results were obtained as 0.51 when R/F: 1/3; 0.56 and when R/F: 1/1 and 0.48 when R/F: 3/1. It was observed that the error rates obtained were higher when the sample ratio of the reference and focal groups was equal.

A variance analysis was performed to determine how the power rates of the bundle with DIF and type 1 error rates of the bundle with DIF differ in terms of the variables analyzed in the study. Since the power rates of the bundles did not differ in the conditions when all the items in the bundle contained DIF, a variance analysis could not be performed on these bundles. ANOVA results for the power rates are demonstrated in Table 5 and ANOVA results for the error rates are demonstrated in Table 6.

When the ANOVA results for the power rates in Table 5 were analyzed, the difference of power rates was found significant for the variables: the correlation between the dimensions, the sample size, and the ratio of reference to focal group size in the condition when the items that show DIF and do not show DIF was present together in the bundle and also there were items with DIF outside of the bundle. In this condition, the largest effect size belongs to the sample size. The sample size had a medium effect size over the power rates analyzed in this condition.

The variables, the correlation between the dimensions and the ratio of reference to focal group size, had a small effect size on the power rates. In this condition, only the sample size and sample ratio interaction were significant among the interaction effects between the variables (F=.704, η^2=.005). According to Post-hoc test results, the power rates that belong to the condition in which the correlation between the dimensions was 0.10 were significantly higher than the condition in which correlation was 0.80. No significant difference was found between the correlation 0.45 and other correlations. The power rates which belong to N=5000 value were

significantly higher than the power rates of N=2000 value. The rate R/F: 1/1 was determined to be significantly higher than the ratio of the other two samples. No significant difference was observed between the ratios R/F: 1/3 and 3/1.

Tablo 5. *ANOVA results for the power rates of the bundles.*

		$Bundle_{2a}$		$Bundle_{2b}$	
Effects	Sd	F	η^2	F	η^2
CD*SS*SR	4	.681**	.001	.408	.001
SS*SR	2	4.704**	.005	2.654	.003
CD*SR	4	1.273	.002	.408	.001
CD*SS	2	.124	.0001	2.654	.003
SR	2	9.558**	.009	2.654	.003
SS	1	215.533**	.11	15.309**	.008
CD	2	6.631**	.006	2.654	.003

Notes. **: p<0.05; $Bundle_{2a}$: The condition in which the items that show DIF and do not show DIF are present together in the bundle and also there are items with DIF outside of the bundle; $Bundle_{2b}$: The condition in which the items that show DIF and do not show DIF are present together in the bundle and also there are no items with DIF outside of the bundle; CD= Correlation between Dimensions; SS= Sample Size; SR=Sample Ratio (R/F).

The difference of power rates was found to be significant only for the main effect of the sample size in the condition in which the items that show DIF and do not show DIF were present together in the bundle and there were no items which show DIF outside of the bundle. Sample size had a small effect size on the power rates analyzed in this condition. The power rates when N=5000 were significantly higher than the power rates when N=2000. The power rates of the bundle analyzed in this condition did not make a difference for the interactions between the other main effects and variables.

Tablo 6. *ANOVA results for the type 1 error rates of the bundles.*

		$Bundle_{3a}$		$Bundle_{3b}$	
Effects	Sd	F	η^2	F	η^2
CD*SS*SR	4	.508	.001	.934	.002
SS*SR	2	1.373	.001	1.493	.001
CD*SR	4	.750	.002	.881	.002
CD*SS	2	.792	.0008	.314	.0003
SR	2	4.476**	.005	5.022**	.005
SS	1	150.087**	.08	245.214**	.12
CD	2	5.188**	.005	1.653	.002

Notes. **: $p<0.05$; $Bundle_{3a}$: ANOVA results for the Type I error rate in the condition in which there are no items showing DIF in the bundle and there are items showing DIF outside of the bundle; $Bundle_{3b}$: ANOVA results for the Type I error rate in the condition in which there are no items showing DIF in the bundle and outside of the bundle; CD= Correlation between Dimensions; SS= Sample Size; SR=Sample Ratio (R/F).

When the ANOVA results for the type 1 error rates in Table 6 were analyzed, the error rates were found significant for the variables: the correlation between the dimensions, the sample size, and the sample ratio in the condition in which there were no items showing DIF in the bundle and there were items showing DIF outside of the bundle. In this condition, the largest effect size belongs to the sample size. The sample size had a medium effect size on type 1 error rate analyzed in this condition. The variables, the correlation between dimensions, and the ratio of reference to focal group size, had a small effect size on the error rates. In this condition, the interaction effects between the variables were not found to be significant. According to Post-hoc test results, the error rates that belong to the condition in which the correlation between

dimensions was 0.10 are significantly higher than the error rates with other correlation values. No significant difference was found between the correlations 0.45 and 0.80. The error rates were significantly higher when N=5000 than the error rates when N=2000. The rate R/F: 1/1 was determined to be significantly higher than the ratio of the other two samples. No significant difference was observed between the rates R/F: 1/3 and 3/1.

In the condition in which there were no items showing DIF in the bundle and outside of the bundle, type 1 error rates were found to be significant for the main effects of the sample size and the ratio of the samples. The largest effect size belongs to the sample size. The sample size had a medium effect size on the type 1 error analyzed in this condition. The ratio of reference to focal group size had a small effect size on the type 1 error rate. In this condition, the interaction effects between the variables were not found to be significant. According to Post-hoc test results, R/F: 1/1 value had significantly higher error rates than R/F: 3/1. No significant difference was found in the other paired comparisons of the ratio of reference to focal group size. The power rates were significantly higher when N=5000 than the power rates when N=2000.

4. DISCUSSION and CONCLUSION

In this study, the results of the power rate and type 1 error rate were analyzed regarding the differentiating location of the item with DIF in the test in terms of the correlation between the dimensions (0.10, 0.45 and 0.80), the sample size (2000 and 5000), and the ratio of reference to focal group size (1/3, 1/1 and 3/1). The results obtained are discussed by reviewing the studies conducted on the topic.

In all conditions analyzed regarding the differentiating location of the item with DIF in the test, the power of SIBTEST to detect items with DIF was obtained as the highest in the conditions when the correlation between the two dimensions was the lowest. In general, the power rates of the bundle relatively decreased as the correlation between the dimensions increased. In other words, the power of SIBTEST to detect items with DIF decreased as the test approached unidimensionality. This result is expected when it is considered that DIF is caused by multidimensionality. Ross (2008), in her study, determined the correlation between the dimensions as 0.316, 0.632 and 0.837 and found that the power of SIBTEST to detect DBF relatively decreased as the correlation between the dimensions increased. This finding is similar with the results obtained in this study. In some of the DIF studies conducted at item level, the variation of the power rates is not very explicit in terms of the correlation between the dimensions (Walker & Şahin, 2016; Lee, 2004). All in all, it is possible to state that as the correlation between dimensions decreases, that is, as the test approaches unidimensionality, the power of determining DIF in the item cluster of SIBTEST increases.

Among all conditions analyzed regarding the location of the item with DIF in the test, it was found that the power rates of the bundle increased as the sample size increased. In the studies conducted by Finch (2012), Ross (2008), and Russell (2005), which addressed DIF in various conditions, the researchers analyzed the power rates of the bundle in various sample sizes and found that the power rates of the bundle increased as the sample size increased. These findings are similar with the results obtained in this study. In addition, in some of the DIF studies conducted at item level, the power rates increased as the sample size increased (Awuor, 2008; Lee, 2004; Bolt, 2002; Narayanan & Swaminathan, 1994; & Ackerman, 1992b).

Another condition examined in this study, besides the sample size, is the ratio of reference to focal group size. Among all conditions analyzed regarding the location of the item with DIF in the test, the highest power rates of the bundle were observed in the conditions in which the ratio of reference to focal group size was equal (R/F: 1/1). No clear pattern was observed in the conditions in which the ratio of reference to focal group size was R/F: 1/3 and 3/1. In their studies Finch (2012) and Ross (2008) analyzed DIF in the bundle and found that the power

rates of the bundle were higher in the conditions when the ratio of reference to focal group size was equal. These findings are similar to the findings obtained in this study. In some of the DIF studies conducted at item level, the power rates were found to be higher in the conditions when the ratio of reference to focal group size was equal (Awuor, 2008; Narayanan & Swaminathan, 1994). Sample size and the ratio of sample sizes of focus and reference groups is an important factor in DIF determination studies. It is suggested that the sample size should be at least 1000 in DIF determination methods based on ITC (Shepard et al., 1981), and that there should be at least 250 individuals in each of the focus and reference groups for the SIBTEST method (Shealy & Stout, 1993). Since SIBTEST is a method based on IRT in determining DIF, since the sample size increased, the power to detect DIF in the item cluster of SIBTEST increased as its sample size increased, and the highest power ratios were obtained under the conditions where the focus and reference group ratios were equal.

The study also analyzed the conditions in which there were items with DIF and there were no items with DIF outside of the bundle when there were no items with DIF in the bundle. It was observed in these conditions that the type 1 errors were notably higher than the nominal alpha level. In a study conducted by Russell (2005), the DIF in the bundle was analyzed in terms of the sample size, test length, and DIF size, and it was observed that the type 1 error rates of the bundle were notably higher than the nominal alpha level. In all conditions, the error rates varied between 0 and 47 in 50 repetitions carried out. In a study conducted by Ross (2008), the DIF in the bundle was analyzed in terms of the correlation between the dimensions, the sample size, the sample ratio, item angle, and DIF size.

The error rates obtained in all conditions in the study do not significantly exceed the nominal alpha level (it varies between 0.04 and 0.07). In a study conducted by Finch (2012), the DIF in the bundle was analyzed in terms of the sample size, the sample ratio, test length, and the item rate in the bundle. As a result, the error rates which do not significantly exceed the nominal alpha level were obtained (it varies between 0.053 and 0.072).

In this study, the error rates of the bundle relatively decreased as the correlation between the dimensions increased. In the study conducted by Ross (2008), the error rates did not show a significant difference in terms of the correlation between the dimensions.

Moreover, in the study, it was observed that the error rates increased as the sample size increased when the error rates were analyzed in terms of the sample size. This finding is not similar with the study results found by Finch (2012), Ross (2008), and Russell (2005).

Another important finding concerns the error rates analyzed in terms of the ratio of reference to focal group size in the study as it was observed that the highest error rates were obtained when the sample ratio of the groups was equal. This finding is similar with the results obtained by Finch (2012). However, the error rates of the bundle did not differ according to the sample ratio of the groups in the study conducted by Ross (2008).

In the study, the DIF detection power of SIBTEST was the highest when all the items forming the bundle showed DIF. When all the items forming the bundle showed DIF, the condition in which there were items with DIF or no items with DIF outside of the bundle in the rest of the test had no effect on the power rates obtained at bundle level.

In this study, the presence of items which do not show DIF in the bundle caused the power rates obtained at bundle level to slightly decrease. However, when the items which show and do not show DIF in the bundle were present together and there were no items which show DIF outside of the bundle, the power rates obtained at bundle level did not fall below the acknowledged limit. On the other hand, the power rates obtained at bundle level fell below the acknowledged limit when the items which show and do not show DIF in the bundle were present together and there were items which show DIF outside of the bundle.

In another condition analyzed in this study, it was observed that the error rates of the bundle were notably higher than the nominal alpha level when there were no items with DIF in the bundle and there were items with and without DIF outside of the bundle. This may be caused by the fact that the items do not show DIF when analyzed individually but they can cause DIF in high levels when they form a bundle.

In DIF studies, DIF does not occur in the item if the item is responsive to the secondary dimension and the ability distributions of the groups do not differentiate on the secondary dimension. Again, DIF does not occur in the item if the ability distributions of the individuals differentiate on the secondary dimension and the item is not responsive to the secondary dimension (Shealy & Stout, 1993; Ackerman, 1992b). In this study, the situations when the bundle and the test did not contain DIF were set by creating conditions in which the individuals had different ability distributions; however, the item was not responsive to the secondary dimension. This condition was assured with the angle values of the item. It was preferred due to one of the conditions analyzed in the study, which is, there are items which contain DIF outside of the bundle, whereas there are no items which contain DIF in the bundle. In future research for the conditions without DIF, the items can be created multidimensional and the error rates can be analyzed in the conditions created without differentiating the abilities of the individuals in the secondary dimension.

In the study, the first 10 items of the test were acknowledged as the bundle. In future research, the effect of the items, which form the bundle and are scattered in the test, for detecting DBF can be studied. Another condition for the items showing DIF in the study, the difference in ability distribution between the reference and focal groups, was kept constant in all conditions. In future research, the effect of the different ability distributions between the reference and focal groups for detecting DBF can be studied. Another suggestion is about the procedure. In this study, the DIF at bundle level was detected using SIBTEST; in future research, the DIF at bundle level can be studied using such procedures as MIMIC and DFIT.

Acknowledgments

This paper was produced from the first author's doctoral dissertation 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.

Authorship Contribution Statement

Didem Ozdogan: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing -original draft. **Hulya Kelecioglu**: Methodology, Supervision, and Validation.

Orcid

Didem Özdoğan https://orcid.org/0000-0002-6631-3996
Hülya Kelecioğlu https://orcid.org/0000-0002-0741-9934

REFERENCES

Ackerman, T.A. (1992a). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*(1), 67-91. https://doi.org/10.1111/j.1745-3984.1992.tb00368.x

Ackerman, T.A. (1992b). *An investigation of the relationship between reliability, power, and the type I error rate of the Mantel-Haenszel and simultaneous item bias detection*

procedures. Paper presented at the National Council on Measurement in Education (April 21-23), San Fransisco, CA. https://eric.ed.gov/?id=ED344937

Ackerman, T.A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7*(4), 255-278. https://doi.org/10.1207/s15324818ame0704_1

Ackerman, T.A., Gierl, M.J., & Walker, C.M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*(3), 37-51. https://doi.org/10.1111/j.1745-3992.2003.tb00136.x

Atalay Kabasakal K., Arsan N., Gök, B., & Kelecioğlu H. (2014). Değişen madde fonksiyonunun belirlenmesinde mtk olabilirlik oranı sibtest ve mantel-haenszel yöntemlerinin performanslarının (i. tip hata ve güç) karşılaştırılması [Comparing Performances (Type I error and Power) of IRT Likelihood Ratio SIBTEST and Mantel-Haenszel Methods in the Determination of Differential Item Functioning]. *Kuram ve Uygulamada Eğitim Bilimleri. 6*(14), 2175-2194. https://doi.org/10.12738/estp.2014.6.2165

Awuor, R.A. (2008). *Effect of unequal sample sizes on the power of dif detection: an irt based monte carlo study with SIBTEST and mantel-haenszel procedures.* [Doctoral dissertation, Virginia Polytechnic Institute and State University]. https://vtechworks.lib.vt.edu/handle/10919/28321

Bolt, D.M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 15*(2), 113-141. https://doi.org/10.1207/S15324818AME1502_01

Bolt, D.M., & Lall, V.F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov Chain Monte Carlo. *Applied Psychological Measurement, 27*(6), 395-414. https://doi.org/10.1177%2F0146621603258350

Boughton, K.A., Gierl, M.J., & Khaliq, S.N. (2000). *Differential bundle functioning on mathematics and science achievement tests: A small step toward understanding differential performance*. Paper presented at the Canadian Society for Studies in Education (May 24-27), Edmonton, Alberta, Canada. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.385.5167&rep=rep1&type=pdf

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement, 16*(2), 129-147. https://doi.org/10.1177%2F014662169201600203

Cronbach, L.J. (1990). *Essentials of psychological testing* (5 ed.). Harper & Row.

Douglas, J.A., Roussos, L.A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33*(4), 465-484. https://doi.org/10.1111/j.1745-3984.1996.tb00502.x

Engelhard, G., Hansche, L., & Rutledge, K.E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Measurement in Education, 3*(4), 347-360. https://doi.org/10.1207/s15324818ame0304_4

Finch, W.H. (2012). The MIMIC model as a tool for differential bundle functioning detection. *Applied Psychological Measurement, 36*(1), 40-59. https://doi.org/10.1177%2F0146621611432863

Gierl, M.J., Bisanz, J., Bisanz, G.L., Boughton, K.A., & Khaliq, S.N. (2001). Illustrating the utility of differential bundle functioning analysis to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice, 20*(2), 26-36. https://doi.org/10.1111/j.1745-3992.2001.tb00060.x

Harwell, M., Stone, C.A., Hsu, T.C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101-125. https://doi.org/10.1177%2F014662169602000201

Mahmood UI.H., & Frank, M. (2020). Discrimination with unidimensional and multidimensional item response theory models for educational data. *Communications in Statistics Simulatıon and Computation.* 1-21. https://doi.org/10.1080/03610918.2019.1705344

Karasar, N. (2020). *Bilimsel araştırma yöntemi, Kavramlar İlkeler (35. Baskı) Teknikler [Scientific Research Method, Concepts Principles Techniques (35 ed.)]*. Nobel Yayıncılık.

Lee, Y. (2004). *The impact of a multidimensional item on differential item functioning (DIF).* [Doctoral dissertation, University of Washington]. https://www.proquest.com/openview/2e24c73698bf27f10d35bd8b63e2cc31/1?pq-origsite=gscholar&cbl=18750&diss=y

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement, 30*(4), 293-311. https://doi.org/10.1111/j.1745-3984.1993.tb00428.x

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*(4), 315-328. https://doi.org/10.1177%2F014662169401800403

Oshima, T.C., & Miller, M.D. (1992). Multidimensionality and item bias in item response theory. *Applied Psychological Measurement, 16*(3), 237-248. https://doi.org/10.1177%2F014662169201600304

Reckase, M.D., & McKinley, R.L. (1991). The discrimination power of items that measure more than one dimension. *Applied Psychological Measurement, 15*(4), 361-373. https://doi.org/10.1177%2F014662169101500407

Ross, T.R. (2008). *The impact of multidimensionality on the detection of differential bundle functioning using simultaneous item bias test* [Doctoral dissertation, Georgia State University]. https://scholarworks.gsu.edu/eps_diss/14/

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*(4), 355-371. https://doi.org/10.1177%2F014662169602000404

Russell, S.S. (2005). *Estimates of type I error and power for indices of differential bundle and test functioning* [Doctoral dissertation, Graduate College of Bowling Green State University]. https://www.proquest.com/openview/25873a6f54d69f576b5c2d3ac61f3aa3/1?pq-origsite=gscholar&cbl=18750&diss=y

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Shepard, L.A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6*, 317-375. https://doi.org/10.2307/1164616

Wiley, D.E. (1991). Test validity and invalidity reconsidered. In R. Snow & D.E. Wiley (Eds.), *Improving inquiry in social science: a volume in honor of Lee J. Cronbach.* Routledge.

Yao, L., & Boughton, K.A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31*(2), 83-105. https://doi.org/10.1177%2F0146621606291559

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 2, 337–356
https://doi.org/10.21449/ijate.982666
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Investigation of affective traits affecting mathematics achievement by SEM and MARS methods

Cagla Kuddar[1,*], **Sevda Cetin**[2]

[1]Hacettepe University, Faculty of Education, Department of Educational Sciences, Educational Measurement and Evaluation, Ankara, Türkiye
[2]Hacettepe University, Faculty of Education, Department of Educational Sciences, Educational Measurement and Evaluation, Ankara, Türkiye

ARTICLE HISTORY
Received: Aug. 13, 2021
Revised: Jan. 26, 2022
Accepted: Mar. 01, 2022

Keywords:
Multivariate Adaptive Regression Splines,
Structural Equation Model,
Data mining,
TIMMS.

Abstract: The purpose of the study is to analyze the affective traits that affect mathematics achievement through Structural Equation Modeling (SEM) as a traditional regression model and Multivariate Adaptive Regression Splines (MARS), as one of the data mining methods. Structural Equation Modeling, one of the regression-based methods, is quite popular for social sciences due to the various advantages it offers; however, it requires very intensive assumptions. MARS method, on the other hand, is a multivariate and adaptive nonparametric statistical regression method used for data classification and modeling. MARS does not need any assumptions such as normality, linearity, homogeneity. It allows variables that do not provide linearity to be included in the analysis. The present study examines whether it is possible to use the MARS method, which is a more flexible method compared to SEM, taking both methods into account. Regarding this goal, the SEM model was created with the program R using the affective data and the achievement variable picked from TIMMS 2019 data. Then, the MARS method was created using the SPM (Salford Predictive Modeler) program. The results of the study showed that at certain points the MARS model gave similar results to the SEM model and MARS model is more compatible with the literature.

1. INTRODUCTION

TIMSS (The Trends in International Mathematics and Science Study) is an international and large-scale examination. The 4th and 8th grade students are able to participate into the examination organized in a four-year period. TIMMS 2019 was the 7th administration of the exam for the candidates from 58 different countries at the 4th-grade and the ones from 39 countries at the 8th-grade (MEB, 2019). Since TIMMS is administered at the international level, it also offers researchers the opportunity to make some possible comparisons among the countries, as well as the opportunity to make the evaluation of their educational systems. The TIMMS exam includes the surveys for the students, teachers and school administrator as well as the achievement tests. In these surveys, the affective traits of the student such as their attitudes towards the schools and the classes, their role in the family, their experience of

*CONTACT: Çağla Kuddar caglakuddar@gmail.com Hacettepe University, Faculty of Education, Department of Educational Sciences, Educational Measurement and Evaluation, Ankara, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

bullying at school are also measured. The factors underlying students' achievement may be cognitive, affective. Many related studies show that achievement is associated with affective traits (Güngör et al., 2007). This reveals that the affective aspect of learning is of great importance (Lehman, 2006). The data obtained from such examinations provide an opportunity to reveal the reasons why students succeed or fail from different perspectives. In the various studies conducted on this subject, the impact of various affective traits on mathematics achievement has been examined (Demir & Kılıç, 2010; Wang, 2007; Zakaria & Nordin, 2008). The results are also important to show the extent to which affective traits affect achievement.

The analysis of TIMSS, which has a high number of variables and specific features, and similar large-scale exam data, can be complex. Data science provides convenience that can be advantageous in the analysis of such large-scale examinations. Accordingly, data means a piece of information and the smallest constituent that carries information (Oruç, 2019). Technology helps multidimensional and wide-ranging data develop; therefore, it has been inevitable that new means of analysis have emerged to add various meanings and dimensions to data, to extract new information that has never been extracted before from the data, and to consider data from different angles.

Data mining means using special algorithms in extracting significant models or relationships from large data stacks (Fayyad et al., 1996). Large data refers to a high amount of information with its density and volume. In other words, how big the data is, how much information it carries about the person or the item it informs, and the information it gives in a second refers to the size of the information. Data has transformed into a subject that concerns not only academics but also everyone in the 21st century. The data is gradually growing. With the use of information communication technologies in almost every corner of life, rapid technological developments trigger an increase in the size and types of data (Emre & Erol, 2017). Data mining includes automatic data extracting, processing, and modeling through a set of methods and techniques (Plotnikova et al., 2020). Methods such as ANN (Artificial Neural Networks), SVM (Support Vector Machines), CART (Classification &Regression Trees), CHAID (Chi-square Automatic Interaction Detector), MARS (Multivariate Adaptive Regression Splines) can be given as an example to the methods used in data mining. MARS data mining method, one of the data mining methods, was used in the study.

The MARS (Multivariate Adaptive Regression Splines) technique was first developed in 1991 by physicist Jerome Friedman at Stanford. The MARS model has many advantages for researchers. It is a nonparametric regression method that does not have any assumptions under the functional relationship between dependent and independent variables. MARS analyzes the effect of independent variables on the dependent variable, the interactions of independent variables with each other, and the effects of these interactions on the dependent variable together (Zhang & Goh, 2016). The interaction of independent variables with each other, which is seen as a problem of multicollinearity in regression analysis, is not considered a problem in the MARS method (Lee & Chen, 2005). The MARS model is a stepwise regression method (Özfalci, 2008). The stepwise regression method can be considered as an advanced method of forwarding selection (Anıl, 2010). According to this method, the variables that may have the highest contribution to the prediction model based on the correlation between the dependent variable and the independent variable are selected and the trivial ones are eliminated. Thus, the deviations in the model are reduced and a model with a higher prediction accuracy is obtained. Regarding the correlation coefficient between dependent and independent variables, the independent variable with the highest correlation coefficient is first included in the model. The stepwise regression model produces the least erroneous prediction model with the highest accuracy (Zateroğlu & Kandırmaz, 2018).

Structural Equation Modeling (SEM) is a model that needs a strong theoretical structure (Kline, 2015). SEM; it is a method that is successful in the testing of complex models and performs many analyses at once. Suggests new arrangements if any, for the network of relationships in the model under examination. It is also a method used in the testing of many theories and the development of new models, since it facilitates to look over the mediation and moderation impacts, and it considers the measurement errors (Dursun & Kocagöz, 2010). Thanks to the many advantages and conveniences it provides, SEM is a common method used in many areas such as marketing, education, psychology, and health. The main feature of SEM is that it decides on the models supported by experimental data; if the data model is not supported, the model is set up and tested again; meaning that a theoretical model is both setup and developed (Candemir, 2018). SEM has a strong theoretical background based on which the regression analysis of the observed variables and the factorial analysis of implicit variables lay (Kline, 2015). SEM is a statistical method that involves intensive assumptions. As with many methods of analysis, it is necessary to verify that various assumptions and requirements are met before analysis also in SEM.

Although there is little studies comparing the results of MARS and SEM, there are some studies comparing MARS with different statistical analysis methods. Deichnmann et al., (2002) made a comparision between a logistic regression and MARS in their studies. It can be concluded that MARS almost all cases produces better results than logistical regression though it is also stated that MARS gives better results when MARS and logistic regression are compared. Another study found out that prediction models created with MARS can be more reliable (Orhan et al. 2018). In the studies conducted so far, MARS has been seen as a strong regression model.

Nonlinear strong prediction models can be established and the relationships between variables can be analyzed and interpreted through MARS (Temel et al., 2010). In this study, the interactions of the affective variables were examined. In a study on the prediction of the MARS model (Zhang & Goh 2016), the advantages of MARS in the BNN method were found out and it is emphasized that the MARS prediction equation is advantageous. Furthermore, it is shown that MARS can replace many types of regression so it can completely ease the analysis and interpretation.

Bolder & Rubin (2007) noted that the MARS method yielded more successful results compared to ordinary least squares, non-parametric Kernel regression, and projection pursuit regression. AL-Qinani (2016) stated that the MARS model showed a noticeable improvement in the accuracy of prediction compared to the multiple linear regression (MLR) method, while Muzır (2011) reported that the MARS method revealed more successful results compared to the binary logistic regression and the theory of artificial neural networks. Moroever, in another study, the results of MARS and CART were compared and it was emphasized that the two types of analysis were more advantageous than other types of regression (Lee and et al., 2006). The result of another study shows that the MARS model makes a more accurate prediction at the point of the accuracy of prediction and regression than models such as artificial neural networks, regression models, regression tree models, and gives as reliable results as other models (Zhou & Leung, 2017). Furthermore, artificial neural networks and the MARS model were compared and the MARS model gave slightly better results considering the procedure than artificial neural networks, and as a result of this study, the MARS model was a strong predictor (Parsai et al., 2016). Abde-Aty and Haleem (2011) used MARS to predict traffic accidents in their study. It has been noted that the MARS model creates strong prediction equations and is an important predictor in predicting traffic accidents.

The data we obtain in international examinations or through data collection methods may tend not to provide the necessary assumptions. If these assumptions are not provided, various

statistical methods cannot be used. Appropriate estimation method should be chosen according to the structure and distribution of the data. The maximum likelihood method is an estimation method that can be used for data measured at least on an equal interval scale with regard to the normal distribution; however, when these assumptions are violated, analyzes can be carried out using the estimation methods that will be preferred for data that do not show categorical and/or normal distribution (Finney et. al, 2006). In this case, the researchers may experience limitations statistically. In nonparametric data, the situation is different. Therefore, in cases where these assumptions are not provided, it is considered important to be able to use nonparametric methods. Data mining methods can be applied in a group of data that do not provide the necessary assumptions in the field of education and social sciences. In this study, it is planned to perform the analyses via SEM and one of the alternative nonparametric methods, MARS, and to discuss these two methods in terms of their advantages and limitations in the practice.

1.1. Purpose of Study

The general purpose of the present study is to examine various affective factors affecting mathematics achievements in the TIMMS 2019 study and the possible relations of such factors with achievement through MARS and SEM analysis methods over the established model. One of the regression-based methods, Structural Equation Modeling (SEM), is quite well-known for social sciences due to its various benefits; however, it requires very intensive assumptions. MARS method, on the other hand, multivariate and adaptive nonparametric statistical regression method used for data classification and modeling. MARS does not need any assumptions such as normality, linearity, and homogeneity. It allows variables that do not provide linearity to be included in the analysis. Comparisons of different statistical methods with MARS are found in the literature (AL-Qinani, 2016; Bolder & Rubin, 2007; Deichnmann et al., 2002; Lee et al., 2006; Muzır, 2011; Zhang & Goh, 2016; Zhou & Leung, 2017) but no comparison of MARS and SEM methods in terms of their advantages and limitations in practice, has been found. The present study seeks an answer to the question if it is possible to use the MARS method, which is a more flexible compared to SEM, taking both methods into account.

MARS can make it possible to study both the data obtained from large-scale exams and the complex relationships in multi-pattern research (Şevgin, 2020). With this aspect, MARS is an efficient method of analysis not only for educational sciences but also for many disciplines. For this reason, in the present study, the results of MARS, which can be considered a relatively new method, were tried to be compared with those of SEM, a conventional method. This comparison may provide convenience to the researchers in social sciences from various aspects and add perspective to analyses. This aspect of the study is expected to contribute to the literature.

Considering the TIMMS 2019 report, it was seen that the measurement of cognitive traits was addressed in general. The subject distributions in mathematics and science were given and the performance of Turkey in such distributions was stated (MEB, 2019). It was noted that cognitive traits were considered in general in the TIMMS assessments; however, the affective traits were not included enough. The effects of the affective constructs on education and the interactions among these contructs are not adequately examined (Meteroğlu, 2015). It is highly believed that the current study can contribute to the affective assessments of TIMMS.

1.2. Research Problem

In this study, the impact of various affective traits on mathematics achievement was examined. These affective traits were "*interest in mathematics*", "*attitude towards school*", "*attitude towards teachers*" and "*bullying*". For this purpose, the following question was identified as the main research question:

Do the various affective factors affecting the mathematics achievement in the TIMMS 2019 study and their possible relations with achievement have predictive differences when analyzed by MARS and SEM analysis methods?

Sub-problems

1. How are the variable interactions when data is analyzed with SEM?
2. How are the variable interactions when data is analyzed with MARS?

To respond to sub-problems, the following hypotheses have been established in light of the literature on the affective data of TIMMS 2019 in the model established.

H1: Bullying significantly affects the math achievement of the students.

H2: The students' attitude towards the school positively affects the math achievement at a significant level.

H3: The students' attitude towards the teachers positively affects their math achievement at a significant level.

H4: The students' interest in mathematics positively affects their math achievement at a significant level.

H5: A statistically significant impact was considered on the math achievement in the mediator variable of the interest in mathematics between the attitude towards the school and the attitude towards the teacher.

H6: For the moderation, a statistically significant impact was considered on the math achievement in the moderator variable of bullying between the attitude towards the school and the attitude towards the teacher.

H7: A statistically significant impact was considered on math achievement in the mediator variable of interest in mathematics and the moderator bullying variable between the attitude towards the school and the attitude towards the teacher.

2. METHOD

2.1. Research Method

The present study bears the characteristics of basic research as it aims to conduct comparative data analysis using the TIMMS 2019 assessment and in doing so, it uses the ready-made package programs. Basic research refers to experimental or theoretical investigations that have no specific applications or purposes and are carried out to gather new information, primarily about the foundations of situations and observable incidents (Karasar, 2015). Basic research is experimental or theoretical research that helps discover information focusing on the process rather than the result, and that has the goal of discovery and allows us to better understand and make sense of the world. In addition, the research where the results of the MARS data mining model and SEM analysis is based on a relational scanning model.

2.2. Population and Sample of Research

Turkey participated in the TIMMS 2019 with 180 schools and 4,028 students at the 4th-grade level. At the 8th-grade level, the application was conducted with 4,077 students at 181 schools. The sample of the present study includes the items selected among the answers of these students given to the affective survey who participated into the mathematical assessment of TIMMS 2019 and the BSMMAT01-05 variables showing the level of achievement (plausible values). (MEB, 2019).

2.3. Data Collection Tools and Process

All data used for the study have been taken from the official site of the TIMMS exam (https://timss2019.org/international-database/). TIMMS 2019 was conducted with the fourth

and eighth grade students in mathematics and science. The exam included the achievement tests and the affective surveys. In the exam, the questions on mathematics and science were asked as a part of the achievement test. The affective surveys were five-point Likert-scale ones that measure the socio-economic level of the student, the attitude towards the teacher, the students' interest and motivation towards the course, and the level of bullying for the students suffered. These affective surveys were prepared not only for the students but also for the teachers and the school administrators. In this study, data consisted of the items selected from the affective surveys of the 8th-graders who participated in the TIMMS 2019 and their achievement scores. The items selected and the factors represented by these items are all given in Table 1, and in the other parts of this study, these items are given with their codes as below.

Table 1. *Selected items for the model.*

Factors	*Items*
Interest in Mathematics	BSBM16C Maths is boring.
	BSBM16E I love maths.
	BSBM16G I love maths problems.
	BSBM16B I wish I did not study maths
Attitude Towards Teacher	BSBM17A Teacher expects us to do.
	BSBM17B Teacher explains clearly.
	BSBM17C Teacher has clear answers.
	BSBM17G Teacher explains again.
	BSBM17F Teacher is associated with the course.
Attitude Towards School	BSBG13A I am present at school.
	BSBG13B I feel safe at school.
	BSBG13C I feel that I belong to the school.
	BSBG13E I am honored to attend this school.
Bullying	BSBG14B Lies about me have spread.
	BSBG14K I was threatened.
	BSBG14M I was excluded from society.
	BSBG14L I was hurt.

2.4. Analysis of Data

Analysis of the data took place in three steps. First, the suitability of the data for analysis was tested and the data was made suitable for analysis. Later, the established model was analyzed by SEM and MARS methods. Finally, the results of the two analyses were interpreted. For affective data, first of all, an explanatory factor analysis (EFA) was performed. A confirmatory factor analysis was then performed to confirm the model.

For the analysis, explanatory factor analysis was performed primarily for affective variables taken from the TIMMS 2019 exam. Later, the analysis was continued with confirmatory factor analysis. EFA and parallel analyzes were carried out with the Jamovi (The Jamovi Project, 2021). According to the KMO results (KMO=0.831) and Barlett Sphericity test results (p<0.05) for EFA analysis, it was decided that the data was suitable for factor analysis.

Given the parallel analysis results provided with the scree plot in Figure 1, it is seen that the model is 4-dimensional and after point 4, the Eigenvalues are similarly distributed. These values account for 54.57% of the variance.

Figure 1. *Scree plot.*

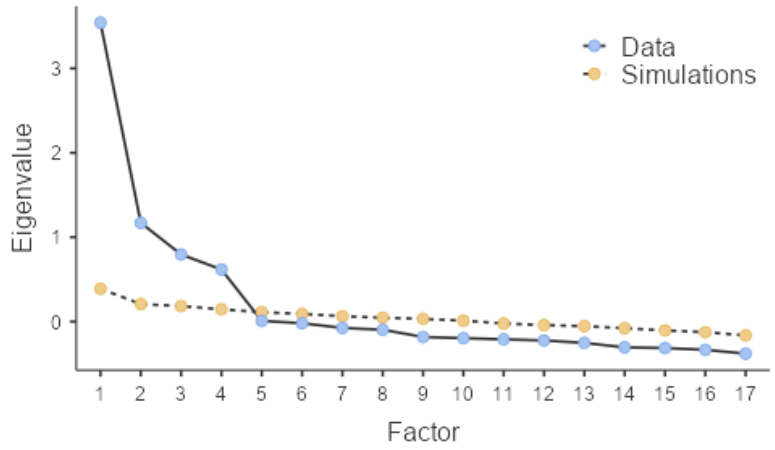


Table 2. *Measurement model dimension matrix.*

	1	2	3	4
BSBG13A			0.483	
BSBG13B			0.558	
BSBG13C			0.810	
BSBG13E			0.669	
BSBG14B				0.445
BSBG14K				0.563
BSBG14L				0.598
BSBG14M				0.435
BSBM16B	-0.679			
BSBM16C	-0.840			
BSBM16E	0.842			
BSBM16G	0.737			
BSBM17A		0.448		
BSBM17B		0.723		
BSBM17C		0.748		
BSBM17F		0.551		
BSBM17G		0.565		

When conducting EFA analysis principal axis factoring extraction method was used in combination with promax rotation. As a result of the EFA analysis, it was continued with 17 items. There were 4 items in the interest variable for mathematics; 5 items in the attitude variable for the teacher, 4 items in the interest variable for mathematics, and 4 items in the bullying variable. The achievement variable had 5 sub-dimensions. The dimensions were called "*Interest in Mathematics*", "*Attitude towards Teacher*", "*Attitude towards School*" and "*Bullying*". The analysis continued later on with 17 items and 5 dependent variables. As seen in Table 2, two items have negative factor loads. When calculating the total score, the item scores are added together to obtain the total score. These two items should not be included in the total score due to their negative factor loads. Since this study was not carried out on total scores, the negative factor loads of the items could not be taken into account. The analysis was proceeded with confirmatory factor analysis (CFA). The values of fit indices are given in Table 3 and CFA model estimations are given in Table 4.

Table 3. *Measurement model fit indices.*

Fit Index	*Calculated Value*
X^2	$p<0.05$
X^2/sd	*2.90*
RMSEA	0.041
SRMR	0.045
GFI	0.97
TLI	0.97
CFI	0.98

Table 4. *CFA model estimations.*

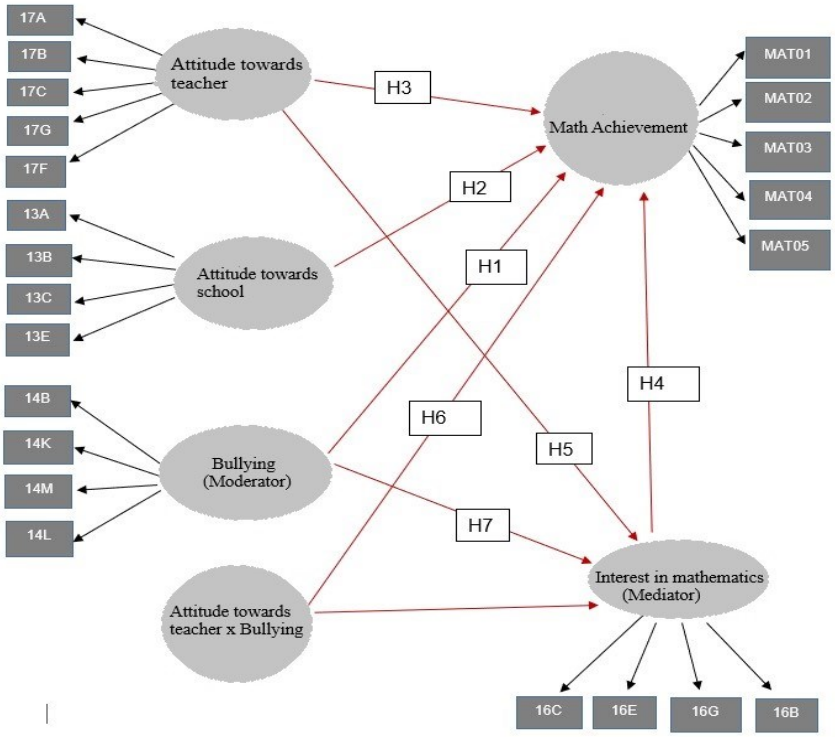
	Estimate	Std.Err	z-value	*p*
Interest in Mathematics =~				
BSBM16B	1.000			
BSBM16C	1.092	0.051	21.438	0.000
BSBM16E	-1.014	0.045	-22.673	0.000
BSBM16G	-0.969	0.045	-21.575	0.000
Attitude Towards Teacher =~				
BSBM17A	1.000			
BSBM17B	1.219	0.084	14.565	0.000
BSBM17C	1.126	0.078	14.448	0.000
BSBM17F	0.999	0.078	12.813	0.000
BSBM17G	0.615	0.05	12.377	0.000
Bullying =~				
BSBG14B	1.000			
BSBG14K	0.395	0.044	9.007	0.000
BSBG14M	0.258	0.036	7.212	0.000
BSBG14L	0.505	0.056	9.017	0.000
Attitude Towards School =~				
BSBG13A	1.000			
BSBG13B	0.958	0.066	14.586	0.000
BSBG13C	1.219	0.079	15.488	0.000
BSBG13E	1.157	0.076	15.277	0.000

As shown in the table, it seems that the degree of fit is excellent or acceptable. The model established in this part, where the four-factor structure is tested, is confirmed. The research model established after factor analysis is presented in Figure 2.

According to the research model, the following hypotheses have been established in light of the literature on the affective data of TIMMS 2019.

H1: Bullying significantly affects achievement.
H2: The attitude towards the school positively affects the achievement at a significant level.
H3: The attitude towards the teachers positively affects the achievement at a significant level.
H4: The interest in mathematics positively affects the achievement at a significant level.

Figure 2. *Research model.*



According to the model, the interest and attitude variables have a significant positive relationship with achievement, while bullying variables have a significant relationship. In other words, the students with a high level of interest towards the course, their attitudes towards the course, and the school also have a high level of achievement (*H2, H3*). The students who are subjected to bullying have significantly lower achievement levels (*H1)*.

It was thought that mediator and moderator effects should also be examined in the model. A mediator variable is a cause variable that has the potential to affect the result, while a moderator variable is a third variable that has the potential to affect the result. In this line, the interest in mathematics was considered as a mediator variable, while bullying was a moderator variable. The reason why these variables are selected is because the interest variable can be a cause variable that can affect the achievement, and the bullying variable can be an effect variable that can affect the achievement. The hypotheses are as follows:

H5: A statistically significant impact was considered on the math achievement in the mediator variable of interest in mathematics between the attitude towards the school and the attitude towards the teacher.

H6: A statistically significant impact was considered on math achievement in the mediator variable of interest in mathematics between the attitude towards the school and the attitude towards the teacher.

H7: A statistically significant impact was considered on the math achievement in the mediator variable of interest in mathematics and the moderator bullying variable between the attitude towards the school and the attitude towards the teacher.

3. RESULT

3.1. Results of the First Sub-Problem

In this section, the first research question *"How are variable interactions when data is analyzed with SEM?"* is dealt with.

Mediator variable analysis: Mediator analysis can be defined as the explanation of the relationships between variables that are related to another mediator variable. Mediator analysis can be explained as a process in which the variable X affects the variable Y and, accordingly, also affects the variable Z. In this case, the mediator variable Y is the cause variable. First, the predictor value between X and Z is checked, and then when the variable y intervenes, it is checked whether a certain part of this predictor value will be explained by this variable. The mediator variable is the one that affects the dependent variable (Şen, 2020). The "Interest in Mathematics" variable was determined in terms of comparison with the MARS model as a mediator variable. In the MARS model, this variable was the most interacting variable in the SPM (Salford Predictive Modeler) program. Since this variable is associated with other variables, it has been considered to what extent it predicts math achievement as a mediator.

Interest in mathematics variable has been taken as a mediator variable. The goodness of fit values of the mediator model are (χ^2 :0.00, *df*:0, RMSEA:0.000, CFI:1.00, TLI:1.00, SRMR: 0.000). Given the model data-fit indices, it can be seen that model data fit provides the necessary criteria. The regression equation is as follows:

Achievement = interest x -. 393 – attitude x 5.56 + school x 6.37 + bullying x 2.81

Moderated variable analysis: A moderator variable also acts as a dependent variable and the relationship between a dependent variable and an independent variable is affected by a third variable. This third variable is called the moderator variable. The effect that occurs in moderator variable analysis occurs only in the presence of this variable (Şen, 2020). The "bullying" variable has been set as the moderator variable.

The goodness of fit values of the moderator model are (χ^2 :0.000, *df*:0, RMSEA:0.000, CFI:1.00, TLI:1.00, SRMR:0.000). Given the goodness of fit values, it is possible to say that the model-data fitness has been ensured.

The regression equation is as follows:

Achievement = interest~attitude x 1.156 + interest~bullying x 1.467.

Mediated moderation analysis: In the mediator analysis, the impact of the attitude towards teacher and bullying interaction on achievement over the interest in mathematics.

As a result of the analysis, the impact of the attitude towards the teacher, interest in mathematics, and bullying variables on achievement over the attitude towards school variable was not found statistically significant (p>0.05). H7 was rejected. Since it consists of a combination of two analyses, it is not specified in the hypothesis table.

The overall results of the established model are given in Table 5. Among the hypothesis established, *H3, H4, H5, H7* resulted in rejection and all the other hypotheses were accepted. Since *H7* is a combination of both methods, it is not included in the table. In other words, the interest variable for mathematics is not a statistically significant variable that predicts achievement. The bullying variable is selected as the mediator variable. When the bullying variable is included in the analysis, it affects the significance of the analysis.

Table 5. *SEM results.*

	Variable	*p*	Hypothesis
Moderator Analysis	Interest~ attitude towards teacher	0.248	H6 Accepted
	Interest~bullying	0.142	H6 Accepted
	Interest~ attitude towards teacherxbullying	0.350	H4 Rejected
Mediator Analysis	Achievement~attitude towards teacher	0.908	H2 Accepted
	Achievement~attitude towards school	0.000	H2 Accepted
	Achievement~bullying	0.842	H1 Accepted
	Achievement~attitude towards teacherxbullying	0.381	H3 Rejected
	Achievement~interest in Mathematics	0.676	H3 Rejected

Note. $p<0.05$; Interest: interest in Mathematics

3.2. Results of the Second Sub-Problem

This part focuses on the results of the second research question "*How are variable interactions when data is analyzed with MARS?*".

At this stage, the SPM program was used to establish the MARS model. At the establishment stage of the model, variables were included in the model as 17 categorical (affective variables) and 5 continuous data (achievement variables). The same model established in SEM is also set here.

Table 6. *Mars model variable interactions.*

Variables	Basic Function Value	Coefficient
Attitude Towards teacher	1	49.97
Attitude Towards School	3	-36.23
Attitude Towards teacher	5	26.52
Bullying	7	-21.30
Interest in Mathematics	9	-49.27
Attitude Towards School	11	-19.00

The contribution of variables to the analysis was primarily studied. Table 6 shows the effects of variables on the achievement-dependent variable. The attitude variable exists with interaction values 49.97 and 26.52 in the basic equation values 1 and 5. The school variable exists with interaction values of -36.23 and-19.00 in the basic equation values of 3 and 11. The bullying variable exists with an interaction value of -21.30 in the basic equation value of 7. The interest variable exists with an interaction value of -49.27 in the basic equation value of 9.

Results of the MARS model are given in Table 7. The MARS model is a stepwise regression method and primarily analyzes all variables, and at the trimming stage, only variables that affect the dependent variable are included in the analysis. Thus, the variables that most affect the dependent variable remain in the analysis, and the others are eliminated. In other words, it does not include the variables that do not affect the dependent variable or variables that have little effect on the analysis. These variables are sorted out under the name of the importance table. It can be said that the variable taken into the final model has a statistically significant impact on the dependent variable. Here are the regression table and the relationship table formed in this way. In summary, the variable that the MARS model receives in the final model has a significant effect on the dependent variable and the level of relationship with the dependent variable is significant. Thus, "bullying significantly affects achievement." *H1* hypothesis is accepted.

Accordingly, the hypothesis that “attitude towards school significantly affects achievement in a positive way" can be explained as follows: the MARS model does not yield a positive or negative relationship. It only gives a statistically significant relationship. In this respect, the level of a relationship, such as positive or negative, can be determined in the analysis such as correlation or others. MARS gives zero to the variables which it does not take into the interaction model. It is therefore considered that the acceptance of this hypothesis is not correct. Although the variable yields a statistically significant effect, it cannot be commented on its direction, therefore the *H2* hypothesis is rejected. Likewise, the hypothesis *H3* "attitude towards teacher positively affects achievement at a significant level” is also rejected. The hypothesis that “interest in mathematics positively affects achievement at a significant level” is also evaluated in this context and the *H4* hypothesis is also rejected.

Table 7. *Results of the Mars model.*

Models	R^2	*GCVR-SQ*
MARS Model		
Mediator Analysis	0.08370	0.06397
Moderator Analysis	0.07488	0.06517
Mediated Moderation Analysis	0.07743	0.06012
Variable Importance Table	*Scores*	
Mediator Analysis		
Attitude Towards Teacher	100	
Attitude towards school	94.20	
Moderator Analysis		
Attitude Towards Teacher	100	
Attitude towards school	84.50	
Mediated Moderation Analysis		
Attitude Towards Teacher	100	
Attitude towards school	90.81	
Interest in Mathematics	33.49	

Mediation analysis in the MARS model: In the MARS model, the “interest in Mathematics” variable is regarded, which actively interacts in many analysis as a mediator variable and has a high coefficient.

Basic function equations for MARS;

Bf1 = Max (0, attit.to.teacher - 7); Bf4 = Max (0, 7 – attit.to.school); Bf6 = Max (0, 8 – attit.to.teacher) x Bf4; Bf8 = Max (0, 6 – attit.to.teacher) x Bf1

Y = 533.548 – 13.7375 x BF1 - 43.8386 x BF4 + 11.4032 x BF6 + 20.3738 x BF8

Model Mat. average = Bf1 Bf4 Bf6 Bf8

Basic function values are equations that aim to reveal the relationships among the variables. The variable constant is 533.54; the interaction of the first basic equation is -13.73; the interaction of the second basic equation is 43.83. Based on this, the relationship between the attitude towards the school and the attitude towards the teacher is 11.40. The attitude towards teacher and the attitude towards school relationship value is 20.37. The model average for the relationship value is 139.74. The relationship level here is contribution-based. A relationship level like correlation is not considered.

The amount of contribution of variables to the model can be seen in Table 6. According to the table, the contribution of the variable of attitude towards teacher to the dependent variable result is 100; while the contribution of the variable of attitude towards school is 94.20. After the variable of interest in mathematics was included in the analysis as a mediator variable, the significance of the variable and the changes in the value of the variable are indicated in Table 6. The coefficients changed after the mediator variable was included in the analysis.

Basic Function Equations for MARS Mediator Variable Analysis;

Bf1 = Max (0, attittoteacher - 7); Bf4 = Max (0,7 - attittoschool); Bf6 = Max (0,8 - attittoteacher) x Bf4; Bf8 = Max (0,6 - attittoschol) x Bf1

Y = 533.014 – 13.6083 x Bf1 – 44.1021 x Bf4 + 11.5904 x Bf6 + 20.6042 x Bf8

Model Mat. average = Bf1 Bf4

The attitude towards the teacher variable decreased from a coefficient of -13.73 to 13.60, and the attitude towards the school variable decreased from -43.83 to -44.10. The first interaction value increased from 11.40 to 11.59 and the second interaction value increased from 20.37 to 20.60. Here, it can be seen that the corresponding variable explains part of the relationship. On the other hand, the node values were 7 and 8 in the first case, while they were 6, 7, and 8 here.

After including the corresponding variable in the analysis, the basic function values also changed. The variable constant decreased from 533.548 to 533.014. The first variable value was 13.60; the second variable value was 44.10; the first interaction value was 11.59, and the second interaction value was 20.60. The average relationship value of the model was 238.75. After including the corresponding variable in the analysis, the significance table in which the contribution levels of the variables were determined also changed. The attitude towards the teacher increased from 100; the attitude towards the school increased from 94.20 to 95.82.

In conclusion, there was a 0.08344-degree interaction between the attitude towards the school and the attitude towards the teacher variables at first, while this interaction was 0.08370 when the variable of interest in mathematics was included in the analysis. The level of variable interaction was increased slightly. Therefore, it is possible to talk about mediation. Including the corresponding variable in the analysis reduced some values but increased some of them. The reason it increased slightly may be because of the low number of variables and the low variance described. As mentioned above, the amount of variance described in international exams is generally low. In this case, this can be shown as the cause of such a slight decrease. Consequently, hypothesis *H5* stating that "there is a statistically significant impact on achievement in the mediator variable of interest in mathematics between the attitude towards the school and the attitude towards the teacher" was accepted.

Moderator analysis in the MARS model: The “bullying” variable was analyzed as the moderator variable in the MARS model. It is seen in Table 6 that the model determinant value decreases when the bullying variable is analyzed (R^2=0.074). The estimated error value is also 0.06517. The lower this value, the lower the error amount. The bullying variable can be said to have a statistically significant impact on achievement. This effect is in the direction of reducing achievement.

Basic Function Equations for MARS Moderator Variable Analysis are as follows;

Bf2 = Max (0, 8- Attittoteacher); Bf4 = Max (0, 7- Attittoschool)

Y = 487.802 + 20.8284 x Bf2 - 17.5118 x Bf4

Model Mat. average = Bf1 Bf4

Looking at the basic function equations, it can be seen that the fixed term value is 487.802. The attitude value towards teacher is 20.82. The attitude value towards school is 17.51. These values

are relationship coefficients. After the bullying variable is included in the analysis, it can be seen that many values, including the constant variable, has changed. Given the interaction information in Table 5, the decrease in achievement was statistically confirmed when the bullying variable was included in the analysis. Bullying has also been a mediator variable for the MARS model. Consequently, hypothesis *H6* stating that "there is a statistically significant impact on achievement in the moderator variable of bullying between the attitude towards school and the attitude towards teacher" was accepted.

Mediated moderation analysis in MARS model: The effect of the variables of attitude towards school and attitude towards teacher in mediator variable of bullying and interest in mathematics was analyzed. Table 6 shows the results.

In moderator and mediator analysis, both mediating and moderating variables were analyzed. Here, the model determinant value was found to be 0.07743. It can be seen that the value R^2 decreased due to bullying; however, it did not lose much value in mediator interest in mathematics. This analysis also shows that it is the right decision for the bullying variable to participate in the analysis as the moderator variable and the variable of interest in mathematics as the mediator variable.

Basic function equations for MARS analysis are as follows;

Bf1 = Max (0, Attittoteacher- 5); Bf2 = Max (0, Attittoschool- 4); Bf4

= Max (0,7- Interestinmaths); Bf5 = Max (0, Interestinmaths- 10)

Y = 501.186- 12.278 x Bf1 + 10.9823 x Bf2- 30.4638 x Bf4- 16.1954 x Bf5

In the variable significance table in Table 6, the attitude variable towards teacher remained the same. The attitude variable towards school was 90.81. The interest variable for mathematics is 33.49. Regarding the coefficient ranking, it is observed that the highest coefficient belongs to the variable of attitude towards the teacher and the lowest coefficient belongs to the variable of interest in mathematics. Since bullying is a weighted variable, it is not specified in the table. Consequently, hypothesis *H7* stating that "there is a statistically significant impact on achievement in the mediator variable of interest in mathematics and moderator variable of bullying between the attitude towards school and the attitude towards teacher" is accepted.

The statistical analysis result comparison of SEM and MARS model is as follows;

Table 8. *Comparison of hypotheses.*

Hypotheses	SEM	MARS
H1	Accepted	Accepted
H2	Accepted	Rejected
H3	Rejected	Rejected
H4	Rejected	Rejected
H5	Rejected	Accepted
H6	Accepted	Accepted
H7	Rejected	Accepted

As shown in Table 8, it is clear that the results of the hypotheses except *H2, H5, H7* are the same. The difference here may be due to the fact that the MARS program does not provide direction information. As a result, although there are some differences between SEM and MARS, it seems they often give similar effects to the same hypotheses.

4. DISCUSSION and CONCLUSION

The purpose of the present study is to examine various affective factors affecting mathematical achievements in the TIMMS 2019 study and the possible relations of such factors with the achievement through MARS and SEM analysis methods over the established ones. For this purpose, the following results have been reached.

As a result of the study, a significant relationship between bullying and achievement was found according to SEM and MARS analyses, and the hypothesis *H1 "Bullying significantly affects math achievement"* is accepted. Pekel (2015) noted that the academic achievement of children who were bullied fell. Also, Kestel and Akbiyik (2016) expressed that bullying negatively affected the academic achievements of the students, as well as their emotional difficulties. Özdinçer and Savaşer (2009) stated that bullying in school was a variable that negatively affected the student's academic achievement. In a research thesis by Sarier (2020), it was stated that not only the academic achievements of the students bullied but also their social and psychological were negatively influenced. In addition, Karataş (2011) agreed the negative effects of bullying and added that this effect might continue for many years. The findings in the literature support the accuracy of both models. Both of the models have similar results to each other. This study concluded that achievement rates decreased significantly when the bullying variable was analyzed in the relationship between the attitude towards the school and the attitude towards the teacher in the frame of the results of the analysis conducted through both SEM and MARS methods. In both methods, the hypothesis *H6: For the moderator variable, the attitude towards the school and the attitude towards the teacher have a statistically significant impact on math achievement in the moderator variable of bullying* has been accepted. In the literature, there has been no study in which bullying is a moderator variable, interest in mathematics is a mediator variable, and they are present together (mediator-moderator).

According to the results of the SEM analysis in the study, it was concluded that the positive attitude of the student towards school positively affects achievement at a significant level and the hypothesis *H2" The attitude towards the school positively affects math achievement at a significant level"* is accepted. The result of the MARS analysis indicated that the student's attitude towards the school significantly affected their achievement. However, the hypothesis was rejected even if it gave statistically significant results in the established MARS model so no comment on this finding could be made. Adıgüzel and Karadaş (2013) stated that the attitude towards the school significantly predicted the achievement in their study, while Bahçetepe and Giorgetti (2015) stated that the school variable significantly predicted the achievement in their study. Atik (2016) stressed that attitude towards school significantly affected the course achievement. These findings in the literature support both models. Both models significantly explained the impact of the attitude towards school on the student achievement.

According to the results of the SEM study, a negative relationship between attitude towards teacher and achievement was found and the hypothesis *H3* "*Attitude towards teacher positively affects math achievement at a significant level"* is rejected. MARS analysis gave this hypothesis a significant relationship, but since no comment can be made on the direction of this relationship, the *H3* hypothesis was again rejected. His study (Cumhur, 2018) concluded that the attitude towards the teacher positively affected the achievement. Güneş et al., (2012) found in their study that the attitude towards the teacher significantly predicted the achievement. Eraslan (2009) emphasized that educating teachers was important in the achievement in his work on PISA. Huyut (2017) stated in his study that the teacher is an important factor in the student achievement. Regarding the findings of studies in the literature, it can be said that the MARS model gives more accurate results than SEM and gives a more consistent result with the literature. In addition, based on the significance table, it can be seen that the MARS model makes a higher contribution to this variable.

According to the results of the SEM analysis, it was concluded that interest in mathematics does not positively affected the achievement at a significant level. According to the results of the MARS analysis, it was concluded that interest in mathematics affected achievement, but it could not be commented on whether it was in a positive direction. According to both methods, the hypothesis *H4* "*Interest in mathematics positively affects math achievement at a significant level"* was rejected. In his study, Güzel (2014) found that interest in mathematics significantly predicted mathematical achievement. To conclude, it can be said that the results of SEM analysis are inconsistent with the literature, while the results of MARS analysis give more consistent results with the literature.

As a result of the SEM analysis, a statistically significant impact was not found on achievement in the mediator variable of interest in mathematics between the attitude towards the school and the attitude towards the teacher. In other words, it was concluded that the interest in the course does not predict achievement and the hypothesis *H5* "*A statistically significant impact was found on math achievement in the mediator variable of interest in mathematics between the attitude towards school and the attitude towards teacher"* was rejected. As a result of the MARS analysis, a statistically significant impact was found on the achievement in the mediator variable of interest in mathematics between the attitude towards the school and the attitude towards the teacher, and the hypothesis *H5* was accepted. In other words, it was concluded that interest in the course predicts achievement in the relationship between attitude towards school and attitude towards teacher.

Considering the advantages of the MARS method, the following can be said: The MARS model does not require an assumption in the cause-effect relationship and does not seek any mathematical relationship. On the contrary, MARS establishes these relations itself. There are no definite judgments about the variables in the MARS model. Variables can be categorical or continuous. In addition, although various assumptions such as normality, linearity and homogeneity are sought in other regression models, assumptions are not sought in the MARS model. MARS is less affected by the multicollinearity problem and enables the model to be established quickly (Le et al., 2009). The MARS algorithm constructs flexible models by using simpler linear regression and data-driven stepwise searching, adding, and pruning. Furthermore, the MARS models developed are easier to interpret (Zhang & Goh, 2016). In addition, it can be said that the use of a package program for MARS analysis is limited as a disadvantage.

The results of this study showed that at certain points the MARS model gave similar results to the SEM model. Considering the advantages of MARS mentioned above, this comparison may be useful for the social science researchers in a variety of ways, including adding perspective to the analyses. Nonetheless, as this study is limited to the analysis of the current data, it is not valid to make a comparison about the estimations, the coefficients or the power of the study. Many other studies show that the MARS model is a powerful predictor but simulation studies are required to make the certain comparisons between SEM and MARS methods.

Acknowledgments

This paper was produced from the part of the first author's master's dissertation 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: Hacettepe University Ethical Committee, No:35853172-300-E.00001313691, Date: 04.11.2020.

Authorship Contribution Statement

Cagla Kuddar: Investigation, Software, Methodology, Formal Analysis, Visualization, Resources, and Writing the original draft. **Sevda Cetin:** Software, Methodology, Supervision, and Validation.

Orcid

Cagla Kuddar https://orcid.org/0000-0001-8734-6722
Sevda Cetin https://orcid.org/0000-0001-5483-595X

REFERENCES

Abdel-Aty, M., & Haleem, K. (2011). Analyzing angle crashes at unsignalized intersections using machine learning techniques. *Accident Analysis & Prevention*, *43*(1), 461-470. https://doi.org/10.1016/j.aap.2010.10.002

Adıgüzel, A., & Karadaş, H. (2013). Ortaöğretim öğrencilerinin okula ilişkin tutumlarının devamsızlık ve okul başarıları arasındaki ilişki [The level of effect of high school students' attitudes towards school on their absenteeism and school success] *Yüzüncü Yıl University Journal of the Faculty of Education*, *10*(1), 49-67. https://dergipark.org.tr/en/pub/yyuefd/issue/13705/165929

Akbıyık, C., & Kestel, M. (2016). Siber zorbalığın öğrencilerin akademik, sosyal ve duygusal durumları üzerindeki etkisinin incelenmesi [An investigation of effects of cyber bullying on students' academic, social and emotional states] *Mersin University Journal of the Faculty of Education*, *12*(3), 844-859. https://doi.org/10.17860/mersinefd.282384

AL-Qinani, I.H. (2016). Multivariate adaptive regression splines (MARS) heuristic model: Application of heavy metal prediction. *International Journal of Modern Trends in Engineering and Research*, *3*(8), 223-229. https://doi.org/10.21884/IJMTER.2016.3027.7NUQV

Anıl, D. (2010). Uluslararası öğrenci başarılarını değerlendirme programı (PISA)'nda Türkiye'deki öğrencilerin fen bilimleri başarılarını etkileyen faktörler [Factors effecting science achievement of science students in programme for international students' achievement (PISA) in Turkey]. *Education and Science, 34*(152).

Arslan Ö.S., & Savaşer, S. (2009). Okulda zorbalık [School bullying] *Milli Eğitim*, *38*(184), 218-227. https://dergipark.org.tr/en/pub/milliegitim/issue/36201/407174

Atik, S. (2016). *Akademik başarının yordayıcıları olarak öğretmene güven, okula karşı tutum, okula yabancılaşma ve okul tükenmişliği [Trust in teacher, attitude towards school,alienation from school and school burnout aspredictors of academic achievement]* (Unpublished Doctoral dissertation) İnönü University

Bahçetepe, Ü., & Giorgetti, F.M. (2015). Akademik başarı ile okul iklimi arasındaki ilişki [The relation between the academic achievement and the school climate]. *İstanbul Eğitimde Yenilikçilik Dergisi*, *1*(3), 83-101. https://dergipark.org.tr/en/download/article-file/436183

Bolder, J., & Rubin, T. (2007). Optimization in a simulation setting: Use of function approximation in debt strategy analysis, *Bank of Canada Working Paper*, 1-92. http://dx.doi.org/10.2139/ssrn.1082840

Candemir, M. (2018). *Antecedents and consequences of Turkish Millennials' ELoyalty: A Structural Equation Modeling Application* [Unpublished master's thesis]. Marmara University

Clement, K.A., Victor, A.T., & Yao, Y.Z. (2020). Multivariate Adaptive Regression Splines (MARS) approach to blast-induced ground vibration prediction, *International Journal of Mining, Reclamation and Environment*, *34*(3), 198-222. https://doi.org/10.1080/17480930.2019.1577940

Cumhur, F. (2018). The investigation of the factors affecting the mathematical success of students in the context of teachers' opinions and suggestions. *Journal of Social and Humanities Sciences Research (JSHSR), 5*(26), 2679-2693. http://dx.doi.org/10.26450/jshsr.647

Deichmann, J., Eshghi, A., Haughton, D., Sayek, S., & Teebagy, N. (2002). Application of multiple adaptive regression splines (MARS) in direct response modeling. *Journal of Interactive Marketing*, *16*(4), 15-27. https://doi.org/10.1002/dir.10040

Demir, İ., & Kılıç, S. (2010). Öğrencilerin matematiğe karşı tutumlarının matematik başarısı üzerine etkisi. [Effects of students' self-related cognitions on matematics achievement] *İstanbul Aydın Üniversitesi Fen Bilimleri Dergisi*, *2*(4), 50-70. https://dergipark.org.tr/en/pub/iaud/issue/30050/32446

Dursun, Y., & Kocagöz, E. (2010). Yapisal eşitlik modellemesi ve regresyon: karşılaştırmalı bir analiz [Structural equation modeling and regression: a comparative analysis] *Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, *35*, 1-17. https://dergipark.org.tr/en/pub/erciyesiibd/issue/5892/77926

Emre, İ.E., & Erol, Ç.S. (2017). Veri analizinde istatistik mi veri madenciliği mi? [Statistics or data mining for data analysis] *Bilişim Teknolojileri Dergisi, 10*(2), 161-167. https://doi.org/10.17671/gazibtd.309297

Eraslan, A. (2009). Finlandiya'nın PISA'daki başarısının nedenleri: Türkiye için alınacak dersler [Reasons behind the Success of Finland in PISA: Lessons for Turkey] *Necatibey Faculty of Education Electronic Journal of Science and Mathematics Education, 3*(2), 238-248. https://dergipark.org.tr/en/pub/balikesirnef/issue/3369/46514

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining.* AAAI/MIT Press.

Finney, S. J., DiStefano, C., Hancock, G. R., & Mueller, R. O. (2006). *Structural equation modeling: A second course.* LAP-Information Age Publishing Inc.

Güneş, S., Görmüş, Ş., Yeşilyurt, F. & Tuzcu, G. (2012). ÖSYS başarısını etkileyen faktörlerin analizi [The determinants of OSYS success] *Pamukkale Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *11*, 71-81.

Güngör, A.A., Eryılmaz, A., & Fakıoğlu, T. (2007). The relationship of freshmen's physics achievement and their related affective characteristics. *Journal of Research in Science Teaching, 44* (8), 1036-1056. https://doi.org/10.1002/tea.20200

Güzel, H. (2004). Genel fizik ve matematik derslerindeki başarı ile matematiğe karşı olan tutum arasındaki ilişki [The relationship between success in general physics and mathematics courses and attitude towards mathematics]. *Journal of Turkish Science Education*, *1*(1), 49- 58. https://www.tused.org/index.php/tused/article/view/41/16

Huyut, M.T., & Keskin, S. (2017). Matematik başarısına etki eden faktörlerin: çevresel faktörlerin çoklu uyum analizi ile belirlenmesi [Determination of factors affecting of mathematics success: environmental factors with multiple correspondence analysis] *Türkiye Teknoloji ve Uygulamalı Bilimler Dergisi*, *1*(2), 48-59. https://dergipark.org.tr/en/pub/tubid/issue/32796/303761

Karasar, N. (2015). *Bilimsel araştırma yöntemleri* [*Science research method*]. Nobel Akademik Yayıncılık.

Karataş, H. (2011). *İlköğretim okullarında zorbalığa yönelik geliştirilen programın etkisinin incelenmesi [Examining the effect of the program developed to address bullying in primary schools].* [Unpublished Doctoral dissertation] Dokuz Eylül University

Kline, R.B. (2015). *Principles and practice of structural equation modeling*. Guilford Publications.

Lee, T.S., Chiu, C.C., Chou, Y.C., & Lu, C.J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines.

Computational Statistics & Data Analysis, *50*(4), 1113-1130. https://doi.org/10.1016/j.csda.2004.11.006

Lee, T.S., & Chen, I.F. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications*, 28, 743-752. https://doi.org/10.1016/j.eswa.2004.12.031

Lehman, R. (2006). The role of emotion in creating instructor and learner presence in the distance education experience. *Journal of Cognitive Affective Learning*, *2* (2), 12-26.

Li, B., Bakshi, B.R., & Goel, P.K. (2009). *Other Methods in Nonlinear Regression*. In Comprehensive Chemometrics, edited by S. D. Brown, R. Tauler, and B. Walczak, 463-476. Elsevier.

Lu, C.J., Lee, T.S. & Lian, C.M. (2012). Sales forecasting for computer wholesalers: A comparison of multivariate adaptive regression splines and artificial neural networks. *Decision Support Systems*, *54*(1), 584-596. https://doi.org/10.1016/j.dss.2012.08.006

MEB. (2019). *TIMSS 2019 Ulusal Matematik ve Fen Ön Raporu: 4. ve 8. Sınıflar* [TIMSS 2019 National mathematics and sciences preliminary report 4th and 8 th grades]. http://odsgm.meb.gov.tr/meb_iys_dosyalar/2020_12/10175514_TIMSS_2019_Turkiye_On_Raporu_.pdf

Muzır, E. (2011). *Basel II düzenlemeleri doğrultusunda kredi riski analizi ve ölçümü: geleneksel ekonometrik modellerin yapay sinir ağları ve MARS modelleriyle karşılaştırılmasına yönelik ampirik bir çalışma* [*Credit risk analysis and measurement in accordance with Basel II regulations: An empirical study to compare traditional econometric models to Artificial Neural Networks and MARS models*] [Unpublished Doctoral dissertation]. İstanbul University

Orhan, H., Teke, E.Ç., & Karcı, Z. (2018). Laktasyon eğrileri modellemesinde çok değişkenli uyarlanabilir regresyon eğrileri (MARS) yönteminin uygulanması [Application of multivariate adaptive regression splines (MARS) for modeling the lactation curves] *Kahramanmaraş Sütçü İmam Üniversitesi Tarım ve Doğa Dergisi, 21*(3), 363-373. https://doi.org/10.18016/ ksudobil.334237

Oruç, M.A. (2019). *İstihbaratın geleceği: Siber uzayda istihbarat ve karşı istihbarat faaliyetlerinde yapay zekâ ve veri bilimi kullanımı* [*Future of intelligence: The using artificial intelligence and data science in intelligence and counter intelligence activities in cyber space*] [Unpublished master's thesis]. İstanbul Aydın University

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]. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *7*(1), 202-220. https://doi.org/10.21031/epod.34424

Özfalcı, Y (2008). *Çok değişkenli uygulanabilir regresyon kesitleri: MARS [Multivariate adaptive regression splines: MARS]* [Unpublished master's thesis]. Gazi University.

Parsaie, A., Haghiabi, A.H., Saneie, M., & Torabi, H. (2016). Prediction of energy dissipation on the stepped spillway using the multivariate adaptive regression splines. ISH Journal of Hydraulic Engineering, *22*(3), 281 292. https://doi.org/10.1080/09715010. 2016.1201782

Pekel-Uludagli, N., & Uçanok, Z. (2005). Akran zorbaligi gruplarinda yalnizlik ve akademik basari ile sosyometrik statüye göre zorba/kurban davranis türleri [Loneliness, academic achievement and types of bullying behavior according to sociometric status in bully/victim groups]. *Türk Psikoloji Dergisi*, *20*(56), 77. https://psycnet.apa.org/record/2006-01737-005

Plotnikova, V., Dumas, M., & Milani, F. (2020). Adaptations of data mining methodologies: a systematic literature review. *PeerJ. Computer science*, 6, 267. https://doi.org/10.7717/peerjcs.267

Rodríguez, C.M., & Wilson, D.T. (2002). Relationship bonding and trust as a foundation for commitment in US–Mexican strategic alliances: A structural equation modeling approach. *Journal of International Marketing*, *10*(4), 53-76. https://doi.org/10.1509/jimk.10.4.53.19553

Sarıer, Y. (2020). TIMSS uygulamalarında Türkiye'nin performansı ve akademik başarıyı yordayan değişkenler [Turkey's performance in TIMSS applications and variables predicting academic achievement]. *Temel Eğitim, 2*(2), 6-27. https://dergipark.org.tr/en/pub/temelegitim/issue/57288/745624

Şen, S. (2020). *Mplus ile yapısal eşitlik modellemesi uygulamaları.* Nobel Yayınları.

Şevgin, H. (2020). *ABİDE 2016 fen başarısının yordanmasında MARS ve brt veri madenciliği yöntemlerinin karşılaştırılması [Predicting the ABIDE 2016 science achievement: The comparison of MARS and BRT data mining methods]* [Unpublished master's thesis]. Gazi University.

Temel, G.O., Ankarali, H., & Yazici, A.C. (2010). Regresyon modellerine alternatif bir yaklasim: MARS [An alternative approach to regression models: MARS.] *Türkiye Klinikleri Biyoistatistik, 2*(2), 58.

The jamovi project (2021). *Jamovi (Version 1.6) [Computer Software].* https://www.jamovi.org/

Wang, J. (2007). A trend study of self-concept and mathematics achievement in a cross cultural context. *Mathematics Education Research*, *19*(3), 33-47. https://doi.org/10.1007/BF03217461

Yoon, S., Co, M.C., Jr, Suero-Tejeda, N., & Bakken, S. (2016). A Data mining approach for exploring correlates of self-reported comparative physical activity levels of urban latinos. *Studies in Health Technology and Informatics*, *225*, 553–557. https:// doi.org/10.3233/978-1-61499-658-3-553

Zakaria, E., & Nordin, N.M. (2008). The effects of mathematics anxiety on matriculation students as related to motivation and achievement. *Eurasia Journal of Mathematics, Science & Technology Education*, *4*(1), 27-30. https://doi.org/10.12973/ejmste/75303

Zateroğlu, M.T., & Kandırmaz (2018). Türkiye için güneşlenme süresi değişiminin izlenmesi, değerlendirilmesi ve bazı meteorolojik verilerle ilişkisinin belirlenmesi [Observation and evaluation of sunshine duration changes in Turkey and determination of relations with some meteorological parameters]. *Ç.Ü Fen ve Mühendislik Bilimleri Dergisi*, *35*(3), 105-114.

Zhang, W., & Goh, A.T. (2016). Multivariate adaptive regression splines and neural network models for prediction of pile drivability. *Geoscience Frontiers, 7*(1), 45-52. https://doi.org/10.1016/j.gsf.2014.10.003

Zhang, W., Wu, C., Li, Y., Wang, L., & Samui, P. (2021). Assessment of pile drivability using random forest regression and multivariate adaptive regression splines. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, 15*(1), 27-40. https://doi.org/10.1080/17499518.2019.1674340

Zhou, Y., & Leung, H. (2007). Predicting object-oriented software maintainability using multivariate adaptive regression splines. *Journal of Systems and Software*, *80*(8), 1349-1361. https://doi.org/10.1016/j.jss.2006.10.049

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 2, 357–375
https://doi.org/10.21449/ijate.958145
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Turkish preschool children's representations of friendship: Story completion method adaptation study

Imray Nur[1,*], **Yasare Aktas Arnas**[2]

[1]Osmaniye Korkut Ata University, Health Services Vocational School of Higher Education, Child Development Program, Türkiye
[2]Hasan Kalyoncu University, Faculty of Education, Department of Pre-School Teaching, Gaziantep, Türkiye

ARTICLE HISTORY
Received: June 27, 2021
Revised: Feb. 14, 2022
Accepted: Mar. 24, 2022

Keywords:
Preschool,
Friendship,
Mental representations,
Story completion.

Abstract: In the preschool period, children's friendships considered a crucial developmental task. Hence, it is critical to evaluate children's mental representations of friendships during this period. This study aims to evaluate the validity and reliability of the story completion protocol designed to evaluate preschool children's mental representations of their friendships in school settings for Turkish children. The Preschool Friendship Story Completion Task consists of six stories, one of which is warming. The stories were translated into Turkish by researchers and Turkish-English language experts, and expert opinion was taken for Turkish-English compatibility. Two pilot studies were conducted to evaluate children's participation in the task and following the instructions. Seventy children attending pre-school education institutions participated in the study. Children are asked to complete the unfinished stories presented through a scene and figures. The coders evaluated the video-recorded children's narratives using a detailed rating system. In addition, teachers evaluated the social skills and problem behaviors of children. The results affirmed that the friendship dimensions were conceptually related to each other and the inter-coder reliability was high. Furthermore, our findings indicate that the story completion method on the basis of the narratives of children is a comprehensive and developmentally appropriate way to assess children's friendships.

1. INTRODUCTION

In preschool education institutions, children take part in the peer group with which they interact daily for a long time. The interactions that occur in this group create opportunities for children to form close friendships over time. Friendship acts as an emotional and cognitive resource for children to learn about themselves and others, and the mutual regulation and intimacy required by close relationships are models for future relationships (Dunn, 2004; Hartup, 1992; Newcomb & Bagwell, 1996). Nonetheless, studies on preschool friendships have generally focused on the presence or absence of friendship (Kingery & Erdley, 2007; Ladd & Troop-Gordon, 2003) and its quality (Park & Waters, 1989; Sebanc, 2003; Youngblade & Belsky, 1992). Few studies have considered children's perceptions of their friendships (San Juan, 2006; Vu, 2015). Due to

*CONTACT: Imray Nur imraynur@hotmail.com Osmaniye Korkut Ata University, Health Services Vocational School of Higher Education, Child Development Program, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

the lack of this research, much remains to be understood about how young children make sense of friendships and how this affects their development.

Furman (1996) asserted that it is necessary to investigate how children perceive these relationships in friendship relationships. Because perceptions indicate how children interpret and affect their behaviors of friends and how this affects the functioning and quality of the relationship. This study, it was examined whether an assessment tool on the basis of the story completion method is reproducible for Turkish children to evaluate the quality of friendship in the preschool period. Verifying that the story completion test is valid and reliable in Turkish culture will meet the need for a developmentally appropriate and valid tool that can be used by researchers who aim to evaluate the friendship perceptions of children.

1.1. Friendship Quality

Friendship refers to the mutual relationship that individuals establish between themselves, both sides are voluntary and happy to spend time with each other (Bukowski et al., 1996). While parents, siblings, teachers, and peers are normal members of the child's social circle, volunteerism is essential in friendship (Laursen, 1996). Bukowski and Hoza (1989) proposed a model in which they define three different aspects or determinants of friendship: Existence of friendship (whether they participate in a mutual friendship), the number of friendships (the prevalence of the friendship network), and the nature of the friendship (the support that friendship provides for the child, the characteristics of friendship or conflict). The existence and stability of friends in early childhood have been frequently addressed by researchers. Studies have verified that friendless children are more anxious, less accepted by their peers (Parker & Seal, 1996), experience more victimization, and exhibit less prosocial behavior than children with friends (Wojslawowicz et al., 2006). Ladd (1990) validated that kindergarten children who have many friends have better academic performance. Proulx and Poulin (2013) pointed out that children who can maintain a long-term relationship with the same friend at school exhibit more social behavior, are more accepted by their peers, and are less shy than those who change friends during the school year. The results of previous studies are valuable in showing that the presence and stability of friends are a crucial for the development of children, but the quality of this friendship is as essential as the children's making friends (Hartup, 1996). Because friendship is also a potential source of negative impact on children's development.

Berndt (1996) elucidated that the quality of friendship is defined by positive characteristics such as companionship and sincerity and negative characteristics such as conflict and dominance. It can be said that friendship is qualified when positive characteristics are more dominant than negative characteristics. Negative experiences such as conflict and interactions that hurt emotions or can be embarrassing also play a significant role in these relationships. Negative friendships can teach harmful skills and lead to unnecessary or potentially damaging information and provide false models for subsequent relationships (Engle et al., 2011; Sebanc, 2003). For instance, studies with older children and adolescents highlighted that friendships are a risk factor for antisocial behavior and substance use, sometimes depending on the characteristics and interactions of friends (Dishion et al., 1995). In studies conducted with preschool children, it was determined that negative friendship quality was associated with externalizing behavioral problems and overt aggression (Engle et. al., 2011; Sebanc, 2003). Besides, high-quality friendships are associated with positive social behavior. For this reason, it is important to consider the qualities that define the friendship of children in terms of preventing many developmental problems.

Considering the limited peer interaction or cognitive and language development in the early stages of development, it is not easy to define friendship for this period (Bukowski et al., 1996; Howes, 1996). Due to the illiteracy of preschool children and some developmental limitations, researchers used more observational methods to determine the quality of friendship in this pe-

riod. One of them is Dyadic Relationships Q-set developed by Park and Waters (1989). Researchers observed pairs of friends in a laboratory setting and encoded children's games with the Q-set coding procedure. Similarly, Youngblade and Belsky (1992) developed the Dyadic Coding System to evaluate the friendship quality of preschool children. These two observational methods provide essential information about the characteristics of preschool friendships, but there are some limitations in their use. First of all, observing children's play is mostly artificial and requires intensive labor. Interactions between friends videotaped in a lab may not be compared to interactions between friends in preschool classrooms. Second, the positive and negative traits obtained from observations of children cannot easily be compared with data derived from the perceptions of older children (Sebanc, 2003). Furthermore, these studies provide an incomplete picture of the friendships of preschool children, as they focus only on observable features of friendship and have difficulty in capturing psychological aspects.

1.2. Using Storytelling to Understand Relationships of Children

Internal working models in attachment theory are defined as mental representations developed by individuals about the world, including the self, and people who are important to them (Bowlby, 1969/2012; Delius et al., 2008; Shaver et al., 1996). Researchers use the narrative method as a way to understand the internal working models or mental representations that children develop in their relationships with their parents. Storytelling and asking the child to complete the story are familiar to children, attracting much more attention than traditional methods such as interviews and questionnaires. Researchers using this method state that, during story completion, children are under the influence of mental representations developed by their family relationships and experiences, and thus story completion gives us information about children's perceptions (Bretherton et al., 1990; Oppenheim et al., 1997; Rydell et al., 2005). Studies highlighted that the story completion task is an appropriate way to assess children's mental representation in their relationships (Bretherton et. al., 1990; Muller et al., 2014; Oppenheim et. al., 1997; Page & Bretherton, 2001; Warren et al., 1996). Nonetheless, studies aimed at determining the mental representations of children about relationships have generally been limited to the parent-child relationship.

The importance and consistent nature of friendship, albeit differently from relationships with parents, shape children's mental representations (Howes, 1996). On the basis of this assumption, the Preschool Friendship Story Completion Task (PFSCT), developed by San Juan (2006), is a narrative procedure designed to evaluate the mental representations of preschool children about their close friendships. This test is an innovative approach on the basis of the past friendship research that examines children's relationship processes to evaluate preschool friendship quality (San Juan, 2006). Sebanc (2003) asserted that evaluations based on only observational methods in the preschool period focus on behaviors such as cooperation, play style, conflict, and conflict resolution. These behaviors are crucial for the functioning of friendships, but they fall short in evaluating the processes such as sincerity and love that distinguish friendship from a simple game partnership (Howes, 1996). PFSCT is on the basis of a model that includes both processes and provides children with a concrete context to express their thoughts and feelings about friendship (San Juan, 2006). This provides the mental representation of young children to be evaluated more effectively than surveys or interviews (Emde et al., 2003).

In Turkey, there are many studies on the perception of primary-school-age children and adolescents friendship (Akın et al., 2014; Demir, 2006; Ercan, 2015; Öztürk, 2009; Öztürk & Kutlu, 2017). In the preschool period, especially peer relations were discussed by researchers (Gülay, 2008; Gülay & Erten, 2011; Özmen, 2013; Ulutaş, 2016; Yoleri, 2015), but no research on children's friendships was available during this period. It is thought that one of the reasons for the lack of studies on friendship and children's perceptions in the early period is the lack of an appropriate assessment tool. This study mainly aimed to examine the validity and reliability of

PFSCT, which is a developmentally appropriate measurement tool for young children and based on their perception, for Turkish children. A methodology on the basis of the storytelling of PFSCT provides an opportunity to evaluate both the play dimension of the friendship relations and the emotional qualities of the relationship in the mental models of preschool children. In this respect, PFSCT offers researchers a friendship quality model similar to those used to examine the friendship relationships of school-age and adolescents (San Juan, 2006). Thus, it is thought that it will provide an opportunity for researchers to better understand the developmental effects of friendship experiences of young children.

2. METHOD

2.1. Participants

The participants of this study consisted of children in the classes of four teachers who volunteered to participate in the study from two kindergartens in Osmaniye city center ($N = 70$; 31 girls, 39 boys; $M_{months} = 67.45$, $SD = 2.80$). In the process of determining the children, a consent letter was first sent to the parents explaining the purpose of the research. 66% of the parents gave consent for their children to participate in the study. Parents also completed the form containing information about their child and themselves. 60% of mothers and 48.6% of fathers have high school or less, 14.3% of mothers and 5.7% of fathers have associate degree, 25.7% of mothers and 45.7% of fathers have undergraduate or graduate degrees. Monthly incomes of families range from 1400–10,000 TL ($M = 4081.45$, $SD = 2051.06$).

All teachers participating in the study are women and have undergraduate degrees (Mage= 31.75, $SD = 4.03$). The number of children in class ranges from 21 to 32 ($M = 26.25$, $SD = 4.50$).

2.2. Instruments

2.2.1. *Friendships of preschool children*

Preschool Friendship Story Completion Task (PFSCT) consists of stories that aim to describe the interactions between two close friends in the preschool period (San Juan, 2006). Developed on the basis of the story completion method created to examine mental representations in attachment relationships, PFSCT determines children's mental representations and perceptions in friendship. During the application of the test, two friends and a peer figure, a wagon, two bicycles, small blocks, toy animals, and different colors of fabrics or felt are used to represent different environments.

In the explanations about the stories and the following parts of the study, “child” represents the participant child, “FC” represents the figure representing the child participating, “BF” represents the figure representing the best friend and “P” represents the peer figure in the stories. PFSCT consists of a warming-up story and 5 stories, each presenting a different type of situation to the children. In the warm-up story (birthday party), FC and BF will go to celebrate P's birthday. The warming history is used to help children get used to the procedures and is not included in the assessment. The three story stems include conflict situations between close friends. The first story stem (Who Gets to Ride?), begins with the peer coming and wanting to get on the wagon while one friend pulling the other by wagon in the playground. The friend got on the carriage gets angry and states that only he/she was going to be pulled. The second story stem (Zoo Animals or Blocks) is based on a common conflict in classrooms. In this story, friends have to decide what to play. However, while FC wants to play with blocks, BF wants to play with toy animals. In the third story stem (Sandbox Betrayal), friends are playing in the sandbox. While they play in the sandbox, the peer comes and offers to ride a bike to one of the friends. The friend goes on a bike ride with the peer, and the other is left alone in the sandbox. The fourth story stem (Can’t Build the Tower) was created specifically to evaluate the helping

behavior of children. FC tries to build a big tower with blocks in the classroom, but building this tower is very difficult. The fifth story stem (Friend Moving Away) uses the loss of a friend so that children can express their feelings of intimacy and affection. FC and BF play games at home. BF says they will move to a very distant place (see Interviewer Protocol and "Who Gets to Ride?" story in Appendix).

PFSCT includes two different rating systems, friendship, and peer relations, to evaluate the processes that occur in children's responses to story stems. Friendship processes include Companionship, Exclusivity, Conflict, Conflict Resolution, Relationship Asymmetry, Helping Behavior, and Intimacy/Affection. The second rating, which focuses on peer interactions in children's narratives, includes Positive Peer Interaction, and Negative Peer Interaction. Furthermore, the narrative coherence of the children's responses to the story stems was rated. Narrative coherence refers to the degree to which the child is presenting a smooth storyline and whether or not she/he is addressing the dilemma in the story.

2.2.1.1. Adaptation Process. PFSCT was first translated into Turkish by researchers and two English experts. The appropriateness of the translations was compared and the translated scale was sent to three academicians (Ph.D. degree) who are experts in preschool education and have knowledge of English for evaluation of how well the translated stories correspond to the original content. The revised translation of the scale in line with the recommendations of the experts was sent to different three academics (Ph.D. degree) who are also experts in preschool education for evaluation in terms of understandability and suitability for purpose. The stories, which were rearranged in line with the recommendations, became ready for the validity and reliability study of the Turkish form of the test.

Two pilot applications have been carried out to evaluate whether the stories in PFSCT are understood by children and whether they can participate in the game and follow the instructions. 12 children participated in the first pilot. In this practice, it was observed that the children participated in the stories, able to continue the stories following the general theme and follow the instructions. However, the fact that the warming story took place in P's house and the "Friend Moving Away" story in FC's house confused the children. Participant children are asked to identify their close friends, while the closest friend in the class is asked. Some children do not have the opportunity to meet their closest classmates at home. For example, participant children used expressions such as "We cannot go to her/his house. My mother will not allow" or "but how will FC and BF go to his house?" in the warming story. Similarly, FC and BF play a game in FC's house in the "Friend Moving Away" story. Participating children used expressions such as "BF doesn't come to us anyway. I'll play with another friend, too" or "his/her home is far away. "he/she cannot come to us". There upon, the researchers rearranged the stories to be in the classroom. The warming story has been changed to celebrate P's birthday in the classroom. In the story "Friend Moving Away", FC and BF play games in the classroom. Meanwhile, FC told BF, "You know, my mom told me we were going to move. I will not come to this school anymore, I will go to another school." After the stories were organized, a second pilot study was conducted with seven children. It was determined that the children were able to maintain all the stories following the given theme.

2.2.1.2. Rating System for Scoring PFSCT Narrative of Children. The children's responses to the stories were recorded with a camera and evaluated for the themes of friendship, peer relations, and narrative coherence (Companionship, Exclusivity, Conflict, Conflict Resolution, Relationship Asymmetry, Helping Behavior, Intimacy/Affection, Positive Peer Interaction, Negative Peer Interaction and Narrative Coherence). Some stories to PFSCT were created to evaluate specific friendship processes, but all stories were evaluated for all friendship processes. For instance, although Intimacy/Affection are scored for all stories, this friendship process is quite evident in the "Friend Moving Away" story. Sub-dimensions for each story stem

were scored separately with a system including 4 ratings. For instance, the helping behavior dimension is encoded as 0 = no evidence of helping behavior, 1 = low helping behavior, 2 = moderate helping behavior, and 3 = high helping behavior. There is a detailed procedure for rating each sub-dimension.

The conflict and conflict resolution sub-dimension is scored in two different conflict and conflict resolution situations. First, the continuation and resolution of the conflict given in the story stem are scored. Secondly, unlike the conflict given in the story stem, when there is a new conflict situation between children, this conflict is scored separately as the created conflict and its resolution. A score of "0" (not appropriate at all, no conflict created) was given if there was no conflict between the children other than the conflict specified in the story stem. In the absence of the created conflict, the solution of the created conflict also received "0" points. However, this score was rewritten as "4" during the analysis, as this situation would be interpreted as the conflict that was created was not resolved.

2.2.2. *Social skills and problem behaviors*

Preschool and Kindergarten Behavior Scales (PKBS-2) was developed to evaluate the social skills and problem behaviors of children aged 3-6 (Merrell, 2003; adapted in Turkish by Özbey, 2009). The scale based on teacher perceptions consists of two independent scales: Social Skills and Problem Behavior. The Social Skills Scale consists of social cooperation (11 items), social independence and social acceptance (8 items), and social interaction (4 items). The Cronbach Alpha values of the Social Skills Scale sub-dimensions and total score were determined as .92, .88, .88, and .94, respectively (Özbey, 2009). The Problem Behavior scale consists of four sub-dimensions: externalizing problems (16 items), internalizing problems (5 items), antisocial (3 items), and egocentric (3 items). The Cronbach Alpha values of the sub-dimensions and total score of the scale are .95, .87, .81, .72 and .96, respectively (Özbey, 2009). The reliability of PKBS-2 sub-dimensions used in the current study was found to be .89 for social cooperation, .83 for social independence and social acceptance, .95 for social interaction, .96 for externalizing problems, and .81 for internalizing problems.

2.2.3. *Language skills*

Test of Early Language Development-Third Edition (TELD-3) was used to evaluate children's language skills (Hresko et al., 1999; adapted in Turkish by Güven & Topbaş, 2014). TEDİL consists of parallel forms, A and B. Each form includes two subtests, receptive and expressive. There are a total of 76 items in each form. This test requires skills such as showing the spoken word in picture booklets, understanding spoken instructions and answering questions verbally. Reliability measures for receptive and expressive language subtests revealed that test-retest reliability was .96−.93; inter-rater reliability was .99−.99; and internal consistency coefficient was .94−.92, respectively. The verbal language scores of children were used for analysis.

2.3. Working with Children

Individual interviews were conducted with the children participating in the study by the first researcher at the children's school. The researcher spent at least three hours in the children's classrooms to develop relationships with the children before starting the interviews. During this time, the researcher introduced herself, participated in the activities of the children, and played games with them. Later, children whose parents gave their consent were invited to the interview one by one. The children were informed about the research and their consent was obtained to participate. Besides, they were informed that their answers would be confidential and that they could stop answering whenever they wanted. All children agreed to participate. Explained to the children that the camera will record stories and the reasons for this. None of the children objected to the camera, and the vast majority ignored it.

Participant children's close friends in class were determined using a two-stage sociometric nomination procedure. Initially, children were shown photos of their classmates and asked to choose their three friends with whom they played the most. Children are encouraged to then reconsider these three choices to choose their best friend.

During the application of PFSCT, three child figures were presented to the children first and they were asked to decide which one would be in the stories. The child then chose a figure for BF (the figures are presented according to the gender of the child and the gender of the child chosen as a close friend). The remaining child figure was said to be one of the children in the class and his name was Mert (if the child is a boy) or Ayşe (if the child is a girl). Before starting the individual interviews, it was made sure that there was no child with these names in the classroom so that the child did not have direct contact with any peers in the classroom. The reason for this is to prevent situations that may affect children's stories and add another relationship to the stories (names are changed if there are children with the same name in the classroom). After the participant child learned who each figure belongs to and the stage materials, firstly, the story stem of "Birthday Party", which is a warming story, is presented. Then the children completed the story stems of "Who Gets to Ride?", "Zoo Animals or Blocks", "Sandbox Betrayal", "Can't Build the Tower" and "Friend Moving Away" respectively. After each story stem is presented to the child, it is said "show me and tell me what happens next". To evaluate the children's language skills, the test was administered 2-3 days after the stories were completed.

2.4. Analytic Strategy

All of the PFSCT data recorded by video were coded by the first researcher. To evaluate the reliability, two independent coder encoded 30 videos (42% of the sample) randomly selected. Before the reliability study, two coders discussed the scoring system on the stories of 5 children who were not included in the analysis to evaluate the PFSCT correctly. The inter-coder agreement was determined by the intra-class correlations (ICC) coefficient. Before starting the analysis, a series of preliminary analyzes were made. In these analyzes, means, standard deviations, and correlations were calculated for PFSCT sub-dimensions and TELD-3. Correlations between social skills, problem behaviors, and PFSCT were evaluated to examine the convergent and discriminating validity of PFSCT.

3. RESULT

3.1. Descriptive Analyses

As stated before, each story in PFSCT was scored for different sub-dimensions. Then, the scores obtained from the stories for each sub-dimension were collected and the total score for that sub-dimension was obtained. As seen in Table 1, average scores are relatively low. Besides, it is observed that the highest scores obtained in most of the sub-dimensions are also low. The reason for this is that each story in PFSCT is planned to reveal different friendship processes.

The exclusivity sub-dimension refers to the interactions in children's narratives where friends prefer to play together and P is somewhat excluded. The reason this dimension has a low average is due to the fact that although most of the children prefer their close friend, they do not exclude the peer aggressively. "Who Gets to Ride?" and the "Sandbox Betrayal" story stems specifically offer conditions that encourage peer exclusion. Nonetheless, 7.1% of the children in the story of "Who Gets to Ride?" and 14.2% of the children in the "Sandbox Betrayal" story exclude peers aggressively, respectively, 58% and 50% of the children included their peers in their games.

Table 1. *Descriptive information for PFSCT and TELD-3.*

Variable	Min-Max *(possible max)*	Mean (*SD*)	Median
Friendship Feature			
Companionship	2.00-15.00 (15.00)	6.77 (3.14)	6.0
Exclusivity	.00-11.00 (12.00)	3.50 (2.47)	3.50
Sustained Conflict	.00-6.00 (9.00)	1.63 (1.56)	1.00
Resolution of Stem Conflict	.00-9.00 (9.00)	5.20 (1.99)	6.00
Created Conflict	.00-6.00 (15.00)	0.37 (0.95)	0.00
Resolution of Created Conflict	.00-3.00 (15.00)	0.35 (0.74)	0.00
Relationship Asymmetry	.00-3.00 (15.00)	0.32 (0.73)	0.00
Helping Behavior	.00-9.00 (15.00)	1.64 (1.62)	2.00
Intimacy & Affection	.00-8.00 (15.00)	2.50 (1.91)	2.00
Peer Interaction			
Positive Peer Interaction	.00-12.00 (12.00)	4.01 (2.70)	4.00
Negative Peer Interaction	.00-10.00	1.20 (1.77)	1.00
Narrative Coherence	6.00-15.00 (15.00)	12.60 (2.73)	13.00
Language skills	80.0-132.0	102.3 (12.59)	103.50

Sustained conflict, the first of the conflict situations, is a continuation of three story stems (Who Gets to Ride?, Zoo Animals or Blocks, Sandbox Betrayal). When this sub-dimension was examined, it was determined that 72.9% of the children did not continue the conflict or decided what to do at the end even if they told a low-level conflict. The second conflict situation depicts a new conflict between friends that has nothing to do with story stems. Only 20% of the children told about the new conflicts that occurred between friends.

When the resolution of the root conflict, which is the first of the conflict resolution situations, was examined, 85% of the children responded to the conflict in the story of "Who Gets to Ride?" with more complex and appropriate solution strategies (rated 2 or 3). Similarly, in the "Zoo Animals or Blocks" story (65%) and the “Sandbox Betrayal” (58%), more than half of the children were able to resolve the conflict at the root of the story. The average of the solution of the created conflict has a relatively low score. This is because only 20% of the children create a new conflict. Furthermore, the conflict resolution score created for children who do not create new conflict is "0". This situation creates a problem for the analysis. For this reason, "4" was taken instead of "0" in further analysis. 64% of the children who created a new conflict were able to solve the problem.

The relatively low average in the relationship asymmetry sub-dimension reflects the equality-based interactions of friends throughout the stories. The coders gave a score of "3" only once, and above the score of "1", only 7 times were rated. Although the helping behavior sub-dimension was scored for all stories, it appeared mostly in the "Can’t Build the Tower" story, as expected. 54% of the children told a story where one of the friends helped or offered help. Similarly, the Intimacy/Affection sub-dimension was scored in all five stories, but, as expected, “Friend Moving Away” appeared the most. 10% of the children did not express any Intimacy/Affection among friends.

When the averages of the PFSCT peer interaction processes were examined, 21% of the children got low scores from positive peer interaction (0 or 1 points). Besides, 70% of the children got low scores from negative peer interactions. This situation affirmed that children do not exclude their peers and conflict between peers is low as mentioned before. Finally, when the

narrative consistency, which aims to evaluate whether the child's narratives are in a proper storyline, is examined, it is determined that children generally told consistent stories.

3.2. Correlations Analyses

Table 2 shows the correlations between PTT sub-dimensions and language skills of children. Significant relationships were determined between friendship, which is an important dimension of Companionship, and Exclusivity, Helping Behavior, Intimacy/Affection. This situation shows that children represent more private, close, and intimate relationships with their close friends by telling more interactive and more fun games with their close friends. At the same time, the Companionship was found to be correlated to the Resolution of Stem Conflict. In three conflict stories, children who used more complex strategies to resolve conflicts were told more interactive games with their friends.

Table 2. *Correlations analyses for PFSCT and TELD-3.*

Variable	2	3	4	5	6	7	8	9	10	11	12	LS
1.Companionship	**.35****	.03	**.72****	.18	-.16	.16	**.49****	**.74****	**.58****	.14	**.63****	.22
2.Exclusivity	-	.20	**.35****	.06	-.19	-.12	.13	**.32****	-.21	**.61****	.28*	.01
3.Sustained Conflict		-	-.14	**.42****	**-.24***	.09	.02	.23	-.06	**.43****	.09	**-.24***
4.Res. of Stem Conf.			-	-.03	.04	.10	.23	**.49****	**.34****	.12	**.58****	**.23***
5.Created Conflict				-	-.69	**.40****	.05	.17	.11	**.29***	.11	-.10
6.Res. of Created Conf.					-	**-.26***	-.08	-.17	-.09	**-.41****	-.04	-.02
7.Rel. Asymmetry						-	-.11	.02	.08	.01	.19	.22
8.Helping Behavior							-	**.42****	**.41****	.19	**.36****	.05
9.Intimacy & Affection								-	**.36****	.22	**.54****	.13
10.Peer Positive									-	**-.26***	**.39****	.22
11.Peer Negative										-	.08	-.10
12.Narrative Coherence											-	.19

Note: *$p < .05$; **$p < .01$; LS = Language skills

Significant correlations have been found between Exclusivity and Resolution of Stem Conflict, Intimacy/Affection. Since Exclusivity in friendship is related to excluding peers, this indicates that children stimulate finding solutions to the conflict at the root of the story by excluding the peer in prosocial or aggressive ways. As the peer exclusion scores of the children increase, the scores of Intimacy/Affection towards friends also increase. Similarly, the relationships between Helping Behavior and Intimacy/Affection were found to be significant. This situation reveals that generally friends who have strong emotional bonds also support each other by helping each other. As stated earlier, children generally did not represent highly asymmetry in their relationships with close friends. Similarly, a few children's stories have portrayed a new conflict creation situation. Nonetheless, the correlation between Relationship Asymmetry and Created Conflict is significant.

Significant positive correlations were determined between the positive peer interaction sub-dimension and Companionship, Resolution of Stem Conflict, Helping Behavior, Intimacy/Affection sub-dimensions. On the other hand, positive correlations were found between negative peer interaction and Exclusivity, Sustained Conflict and Created Conflict, and negative relationships were found with the Resolution of Created Conflict. This affirmed that children who foster close, sincere and supportive relationships with their close friends exhibit more positive

peer interactions. The correlation between Exclusivity and Negative Peer Interaction is as expected, as the sub-dimension of being special in friendship is scored for excluding peer.

In addition to these findings, story consistency also confirmed positive correlations with some sub-dimensions of friendship and positive peer interaction. While negative correlations were determined between children's language skills and Sustained Conflict, the correlations between the Resolution of Stem Conflict were positively significant.

3.3. Coding Reliability

ICC was calculated for each sub-dimension of PFSCT in order to evaluate inter-coder reliability. The ICC is .83 for Companionship, .81 for Exclusivity, .89 for Sustained Conflict, .94 for Resolution of Stem Conflict, .81 for Created Conflict, .77 for Resolution of Created Conflict, .79 for Relationship Asymmetry, .79 for Helping Behavior, .79 for Intimacy/Affection, .89 for Positive Peer Interaction, .91 for Negative Peer Interaction, and .90 for Narrative Coherence.

3.4. Concurrent Validity

To address the concurrent validity of PFSCT, correlations between sub-dimensions of PFSCT and sub-dimensions of PKBS-2 social skills and problem behavior scales were examined (Table 3). Positive correlations were found between Companionship, Resolution of Stem Conflict, Intimacy/Affection, and Positive Peer Interaction in narratives of children, and Social Interaction based on perceptions of teachers. Furthermore, teachers reported fewer Internalizing Problems for children who represented more Positive Peer Interaction in their stories. Similarly, more Negative Peer Interaction representations of children are associated with less Social Cooperation perceived by teachers.

Table 3. *Correlations analyses for PFSCT and PKBS-2.*

	Social Skills			Problem Behavior	
Variable	SC	SISA	SI	EP	IP
Companionship	-.04	.06	**.39****	-.02	-.11
Exclusivity	-.18	.00	.10	.03	-.10
Sustained Conflict	-.11	.19	.15	.03	-.16
Resolution of Stem Conflict	-.07	-.02	**.32****	-.02	-.15
Created Conflict	.01	.15	.11	-.08	-.11
Resolution of Created Conflict	-.06	-.08	-.07	.13	.09
Relationship Asymmetry	.07	.01	.10	-.02	.00
Helping Behavior	-.09	-.01	.16	.03	-.10
Intimacy & Affection	-.08	.21	**.34****	.00	-.08
Positive Peer Interaction	.19	.10	**.35****	-.16	**-.28***
Negative Peer Interaction	**-.30****	-.06	-.13	.18	.09

Note: *$p < .05$; **$p < .01$; SC: Social Cooperation; SISA: Social Independence and Social Acceptance; SI: Social Interaction; EP: Externalizing Problems; IP: Internalizing Problems

4. DISCUSSION and CONCLUSION

Building and maintaining friendships is a crucial developmental task for preschoolers. However, researchers emphasized that the quality of friendship is as important as the presence of a friend (Hartup, 1996). This study was designed to evaluate the validity and reliability of the story completion protocol, which was designed to evaluate preschool children's perceptions of their friendships in school environments for Turkish children. The PFSCT is on the basis of the story completion studies designed to assess children's mental representations in previous attachment studies. It is defined as a developmentally appropriate tool to assess the specific and close friendships of children.

A detailed coding guide and rating scale are available to assess whether children complete the videotaped stories reliably. The codes created to assess children's representations in friendships are on the basis of the past friendship research (e.g. Berndt, 2004; Ladd et al., 1996; Sullivan, 1953). Independent coders use this coding guide to rate friendship traits. In the current study, it was determined that the coding guideline was suitable for reliably assessing friendship characteristics, and the inter-coder reliability was found to be high.

The findings of the study confirmed that children can reflect the positive (Companionship, Resolution of Conflict, Helping Behavior, Intimacy/Affection) and negative (Exclusivity, Sustained Conflict, Relationship Asymmetry) characteristics in their friendships to their stories. In addition, there were relatively few negative relational representations in the children's narratives. There could be several reasons for this condition. First, each of the PFSCT stories is designed for specific friendship processes. Three story stems contain conflict situations, but there is no story stem prepared specifically for Exclusivity and Relationship Asymmetry. In addition, considering that the positive friendship narratives in the stories highlighted sufficient variability and formation, children may need some additional encouragement in the negative friendship themes that appear in the stories. For example, children can be given a chance to comment with questions such as “why did you feel like this?” or “what makes your friend feel this way?” Second, a meta-analysis by Newcomb and Bagwell (1995) highlights that children's interactions with friends are different from interactions with peers who are not friends. Children have a more positive relationship with their friends than with those who are not. That is, the intensity and frequency of talking, smiling, laughing, sharing, cooperating, and helping with friends are higher. Furthermore, while conflict situations between friends and non-friends are no different, disagreements between friends are more likely to be resolved through negotiation. As a result, conflict resolution strategies used by friends are more likely to lead to fair outcomes that help preserve relationships. Equality in friendships may be more pronounced than in peer relationships (Newcomb & Bagwell, 1995). Moreover, relationships with friends involve less intense competition and domination than relationships with non-friends. Finally, friends tend to be similar in demographic and behavioral characteristics, and children express more mutual affection, intimacy, and loyalty in their friendships (Erdley et al., 2001).

The friendship sub-dimensions scores obtained from the encodings evaluating the logic, details, and relevance of the children's narratives were correlated with each other to determine consistency. As expected, the findings confirmed that positive and negative friendship dimensions are related to each other. In addition, the relations between Exclusivity and Sustained Conflict, Intimacy/Affection are positively significant. Exclusivity emphasized that two friends prefer each other and exclude others. When it comes to friendships, Exclusivity seems to be a positive feature, but in recent years there have been opinions that it is a negative feature. Friends who prefer one another over others are quite likely to experience more conflict, given that they spend more time together. Previous research has validated that when pairs of friends play together, they develop higher levels of play and experience more conflict (Fonzi et al., 1997; Hartup et al., 1988; Simpkins & Parke, 2002). Similarly, it is not surprising that children who portrayed more Exclusivity showed more Intimacy/Affection. These associations in the current study indicate that the Exclusivity feature in friendship needs further investigation.

By coding the answers of the children to the PFSCT stories, two different structures were obtained, namely friendship characteristics and peer relations. The positive and negative peer relationships in the representations of children further provided evidence for consistency. Findings verify that children reflect the quality of their relationship with their friends to others who are not friends. This indicates that high-quality friendships are associated with higher social skills, and positive interactions with friends provide essential opportunities to learn and practice social skills (Engle et. al., 2011; Howes et al., 1988; Rose & Asher, 2000). Conversely, the

negative friendship quality is associated with higher levels of problem behavior because they pave the way for learning in aggressive and destructive behaviors among friends (Bagwell & Coie, 2004; Berndt, 2004)

Other studies in the parent-child literature indicate that children's language skills do not play any role in the story completion task because using material allows for describing without speaking (Miljkovitch et al., 2007). Nonetheless, some studies highlighted that children's verbal skills are related to secure attachment (Stievenart, et al., 2011), and their capacity to verbalize themselves is effective in dealing with conflicts in positive ways (Bretherton & Oppenheim, 2003; Von Klitzing et al., 2007). The findings of the present study support previous studies. Children's lower language skills were associated with more conflict representations, and higher language skills were associated with more conflict resolution representations. This finding indicates the importance of supporting children's language skills in preschool friendships. There also arises the necessity of adapting completely non-verbal procedures to assess children's friendships.

Narrative coherence in children's stories is associated with positive friendship qualities and positive peer relationships. Narrative coherence refers to whether children address the dilemma at the root of the story and whether their narrative is in a straight line. Narrative coherence was found to be a predictor of children's social competence in previous studies using the story completion procedure (Von Klitzing et. al., 2007). Aggressive content themes in children's story completions are typically not an appropriate response to the story stems and therefore reduce narrative coherence (Oppenheim, 2006). As a result, our findings corroborate that children who can cope with the conflicts in the stories and find a consistent way in their narratives are more comfortable in their interactions with their friends and peers, behave close and sincerely, cooperate, and deal with problems in a constructive way.

Since friendship and social skills are associated in previous studies, we expected a correlation between children's representations and teacher-reported problem behavior and social skills. Teachers' perceptions of children's relationships with peers and perceptions of children's friendships overlap to some extent. This provides moderate support for concurrent validity. At this point, it should be taken into account that teachers generally report peer relations. Friendships of children may differ from their relationships with others. Positive qualities in friendships may not represent positive features of a child's relationship with their peer group (Sebanc, 2003). Given that early childhood is when friendships are often first formed (Howes, 1983), children may be more likely to approve of a friend on the basis of the activities enjoyed by both individuals. At the same time, children may be more tolerant towards their close friends when they have conflicts with their peers.

Although there are relatively few significant correlations between PFSCT and teacher-reported social skills, the associations found are conceptually significant. Researchers have suggested that positive interactions are evident in early friendships (Park & Waters 1989; Youngblade, Park, & Belsky 1993). Considering that friendship is generally defined as a mutual preference for interaction (Howes 2009), it seems normal to have more harmony between friends, fun games, and more constructive solutions to conflicts. Additionally, positive social interactions can be an indicator of motivation to take more responsibility to maintain the friendship. No significant associations were found between the negative representations of children and teachers' perceptions of problem behavior. Children may represent in greater detail the Exclusivity, maintenance of conflict, or dominance in their friendships. Therefore, it may differ from teacher perceptions. In this regard, teachers and children can provide very valuable perspectives that differ from each other. These results clearly confirm that more research is needed on children's representations of friendship, and specifically on Exclusivity and Asymmetry.

In conclusion, our findings indicate that the story completion method is a valid and reliable way to access preschool children's representations of friendships from the attachment framework. The narratives used to evaluate mental representations of children show the importance of giving children the opportunity to express themselves through different means. In addition, PFSCT offers the opportunity to evaluate representations of children in a developmentally appropriate and comprehensive way, using both play and language.

The present study has some limitations. First, our study was limited to a small sample. Our findings and conclusions may not be valid in other circumstances for classrooms, teachers, and children with different characteristics. Examining the friendship representations of children at risk, especially those who exhibit high levels of problem behavior and peer conflict, may provide an opportunity to open a window to the perspectives of children. Thus, it can help teachers and parents be more sensitive to children's experiences.

Negative friendship themes were relatively underrepresented in narratives of children. Making adjustments to reveal these themes more consistently is an important direction for future research. In their narratives, many children who had conflicts with their friends did not openly express their feelings. To elicit emotions in the stories, the interviewer can ask additional questions about how they and their friends are feeling. The story stems that present different states for Conflict, Exclusivity, and Asymmetry can be added, and the rating system can be revised. Additionally, using methods such as children's drawings and interviews with narratives can help capture details from children's perspectives (Wolcott et al., 2019). The characteristics of the other friend not evaluated in the current study are also a crucial factor determining the quality of friendship. For instance, a child with behavioral problems may have different behavioral consequences if they have a friend with high social skills. For this reason, it is noteworthy to consider the characteristics of friends together with the quality of friendship. In addition, in the present study, it was not taken into account whether the children mutually preferred each other. Future research may reveal more about the nature of mutual friendship.

Although friendships are crucial context for children's development, children's social-emotional skills and other social relationships are likely to affect their friendships. Moderate harmony between social skills on teacher perceptions and children's friendship narratives points to the importance of children's representations in this regard. Representations can allow teachers to capture emotional details that they missed during observations. With this new perspective, teachers can identify more functional ways to provide the social and emotional support children need.

Acknowledgments

This research was conducted within the scope of the project titled “The Association of Preschool Children’s Mental Representation in the Relationship With Their Parents, Teachers and Friends With School Adjustment.” This project is supported by Scientific Research Projects Coordination Unit of Cukurova University with the project number of SDK-2017-9660.

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: Çukurova University, 95704281-604.02.02.

Authorship Contribution Statement

Imray Nur: Introduction, Review of Literature, Methodology (Data Collection and Analyses), Discussion and Conclusion. **Yasare Aktas Arnas:** Investigation, Resources, Methodology, Supervision

Orcid
Imray Nur https://orcid.org/0000-0002-1905-1655
Yasare Aktas Arnas https://orcid.org/0000-0002-0738-9325

REFERENCES

Akın, A., Adam-Karduz, F.F., & Akın, Ü. (2014). The validity and reliability of Turkish version of the friendship quality scale. *Journal of Research in Education and Teaching, 3*(4), 378–383.

Bagwell, C.L., & Coie, J.D. (2004). The best friendships of aggressive boys: Relationship quality, conflict management, and rule-breaking behavior. *Journal of Experimental Child Psychology, 88*(1), 5–24. https://doi.org/10.1016/j.jecp.2003.11.004

Bowlby, J. (2012). *Bağlanma* [Attachment]. (T.V. Soylu, Trans.). Pinhan Yayıncılık. (Original publication 1969).

Berndt, T.J. (1996). Exploring the effects of friendship quality on social development. In W.M. Bukowski, A.F. Newcomb, & W.W. Hartup (Eds.), *The company they keep: Friendship in childhood and adolescence* (pp. 346–365). Cambridge University Press.

Berndt, T.J., (2004). Children's friendships: Shifts over a half-century in perspectives on their development and their effects. *Merrill-Paimer Quarterly, 50*(3), 206–223. https://www.jstor.org/stable/23096162

Bretherton, I., & Oppenheim, D. (2003). The MacArthur Story Stem Battery: Development, directions for administration, reliability, validity and reflections about meaning. In R.N. Emde, D.P. Wolf, & D. Oppenheim (Eds.), *Revealing the inner worlds of young children: The MacArthur Story Stem Battery and parent-child narratives* (pp. 55–80). Oxford University Press.

Bretherton, I., Ridgeway, D., & Cassidy, J. (1990). Assessing internal working models of the attachment relationship. In M.T. Greenberg, D. Cicchetti, & E.M. Cummings (Eds.), *Attachment in the preschool years: Theory, research, and intervention* (pp. 273–310). University of Chicago Press.

Bukowski, W.M., & Hoza, B. (1989). Popularity and friendship: Issues in theory, measurement, and outcome. In T.J. Berndt & G.W. Ladd (Eds.), *Wiley series on personality processes. Peer relationships in child development* (pp. 15–45). John Wiley & Sons.

Bukowski, W.M., Newcomb, A.F., & Hartup, W.W. (1996). Friendship and its significance in childhood and adolescence: Introduction and comment. In W.M. Bukowski, A.F. Newcomb & W.W. Hartup (Eds.), *The company they keep: Friendships in childhood and adolescence* (pp. 1–15). Cambridge University Press.

Delius, A., Bovenschen, I., & Spangler, G. (2008). The inner working model as a 'theory of attachment': Development during the preschool years. *Attachment & Human Development, 10*(4), 395–414. https://doi.org/10.1080/14616730802461425

Demir, S. (2006). *The effect to the sociometric status of second degree primary school students of the group guidance program in order to develop friendship abilities* [Unpublished master's thesis]. İnönü University.

Dishion, T.J., Andrews, D.W., & Crosby, L. (1995). Antisocial boys and their friends in early adolescence: Relationship characteristics, quality, and interactional process. *Child Development, 66*(1), 139–151. https://doi.org/10.1111/j.1467-8624.1995.tb00861.x

Dishion, T.J., Capaldi, D., Spracklen, K.M., & Li, F. (1995). Peer ecology of male adolescent drug use. *Development and Psychopathology, 7*(4), 803-824. https://doi.org/10.1017/S0954579400006854

Dunn, J. (2004). Children's friendships: The beginnings of intimacy. Blackwell.

Erdley, C.A., Nangle, D.W., Newman, J.E., & Carpenter, E.M. (2001). Children's friendship experiences and psychological adjustment: Theory and research. *New Directions for Child and Adolescent Development, 91*, 5–24. https://doi.org/10.1002/cd.3

Emde, R.N., Wolf, D.P., & Oppenheim, D. (2003). *Revealing the inner worlds of young children: The MacArthur Story Stem Battery and parent-child narrativ*. Oxford University Press.

Engle, J.M., Mcelwain, N.L., & Lasky, N. (2011). Presence and quality of kindergarten children's friendships: Concurrent and longitudinal associations with child adjustment in the early school years. *Infant and Child Development, 20*(4), 365–386. https://doi.org/10.1002/icd.706

Ercan, H. (2015). Psychometric properties and the adaptation study of the adolescent friendship scale. *The Journal of Academic Social Science Studies, 38*, 227–240. https://doi.org/10.9761/JASSS3054

Fonzi, A., Schneider, B.H., Tani, F., & Tomada, G. (1997). Predicting children's friendship status from their dyadic interaction in structured situations of potential conflict. *Child Development, 68,* 496–506. https://doi.org/10.2307/1131674

Furman, W. (1996). The measurement of friendship perceptions: conceptual and methodological issues. In W.M. Bukowski, A.F. Newcomb, & W.H. Hartup (Eds.), *The company they keep friendship in childhood and adolescence* (pp. 41–65). Cambridge University Press.

Gülay, H. (2008). *Standartization of a scale for measuring peer relations among 5-6 years old children and studying the relations between some familial variables and peer relations of children at this age* [Unpublished doctoral dissertation]. Marmara University.

Gülay, H., & Erten, H. (2011). The predictive effects peer acceptance has on school adjustment variables in preschool children. *E-International Journal of Educational Research, 1*(2), 81–92. http://www.e-ijer.com/tr/download/article-file/89726

Güven, T., & Topbaş, S. (2014). Adaptation of the Test of Early Language Development - Third Edition (TELD-3) into Turkish: Reliability and validity study. *International Journal of Early Childhood Special Education, 6*(2), 151-176. https://doi.org/10.20489/intjecse.62795

Hartup, W.W. (1992). *Having friends, making friends, and keeping friends: Relationships as educational contexts*. ERIC Clearinghouse on Elementary and Early Childhood Education.

Hartup, W.W. (1996). The company they keep: Friendships and their developmental significance. *Child Development, 67,* 1–13. https://doi.org/10.2307/1131681

Hartup, W.W., Laursen, B., Stewart, M.I., & Eastenson, A. (1988). Conflict and the friendship relations of young children. *Child Development, 59,* 1590-1600. https://doi.org/10.2307/1130673

Hresko W.P., Reid D.K., & Hammill D.D. (1999). *Test of Early Language Development (TELD)* (3rd ed.). PRO-ED.

Howes, C. (1983). Patterns of friendship. *Child Development, 54*(4), 1041-1053. https://doi.org/10.2307/1129908

Howes, C. (2009). Friendship in early childhood. In K.H. Rubin, W.M. Bukowski & B. Laursen (Eds.), *Handbook of peer interactions, relations, and groups* (pp. 180–194). Guilford Press.

Howes, C. (1996). The earliest friendships. In W.M. Bukowski, A.F. Newcomb & W.W. Hartup (Eds.), *The Company they keep: Friendships in childhood and adolescence* (pp. 66–86). Cambridge University Press.

Howes, C., Hamilton, C.E., & Philipsen, L.C. (1998). Stability and continuity of child-caregiver and child-peer relationships. *Child Development, 69*(2), 418-426. https://doi.org/10.1111/j.1467-8624.1998.tb06199x

Kingery, J.N., & Erdley, C.A. (2007). Peer experiences as predictors of adjustment across the middle school transition. *Education and Treatment of Children, 30,* 73-88. https://doi.org/10.1353/etc.2007.0007

Ladd, G.W. (1990). Having friends, keeping friends, making friends, and being liked by peers in the classroom: Predictors of children's early school adjustment. *Child Development, 61,* 1081–1100. https://doi.org/10.2307/1130877

Ladd, G.W., Kochenderfer, B.J., & Coleman, C.C. (1996). Friendship quality as a predictor of young children's early school adjustment. *Child Development, 67,* 1103–1118. https://doi.org/10.2307/1131882

Ladd, G.W., & Troop Gordon, W. (2003). The role of chronic peer difficulties in the developmentof children's psychological adjustment problems. *Child Development, 74,* 1344–1367. https://doi.org/10.1111/1467-8624.00611

Laursen, B. (1996). Closeness and conflict in adolescent peer relationships: Interdependence with friends and romantic partners. In W.M. Bukowski, A.F. Newcomb & W.W. Hartup (Eds.), *The company they keep: Friendship in childhood and adolescence* (pp. 186–210). Cambridge University Press.

Merrell, K.W. (2003). *Preschool and kindergarten behavior scales. Examiner's manual* (2nd ed.). Pro-ed.and International Publesher.

Miljkovitch, R., Pierrehumbert, B., & Halfon, O. (2007). Three-year-olds' attachment play narratives and their associations with internalizing problems. *Clinical Psychology & Psychotherapy, 14*(4), 249–257. https://doi.org/10.1002/cpp.535

Muller, E., Perren, S., & Wustmann–Seiler, C. (2014). Coherence and content of conflict-based narratives: Associations to family risk and maladjustment. *Journal of Family Psychology, 28,* 707–717. https://doi.org/10.1037/a0037845

Newcomb, A.F., & Bagwell, C.L. (1995). Children's friendship relations: A meta-analytic review. *Psychological Bulletin, 117,* 306-347.

Newcomb, A.F., & Bagwell, C.L. (1996). The developmental significance of children's friendship relations. In W.M. Bukowski, A.F. Newcomb & W.W. Hartup (Eds.), *The company they keep: Friendships in childhood and adolescence* (pp. 289–321). Cambridge University Press.

Oppenheim, D. (2006). Child, parent, and parent-child emotion narratives: Implications for developmental psychopathology. *Development and Psychopathology, 18*(3), 771-790. https://doi.org/10.1017/S095457940606038X

Oppenheim, D., Nir, A., Warren, S., & Emde, R.N. (1997). Emotion regulation in mother–child narrative co-construction: Associations with children's narratives and adaptation. *Developmental Psychology, 33,* 284–294. https://doi.org/10.1037/0012-1649.33.2.284

Özbey, (2009). *Study of the validity and reliabilty of the Preschool and Kindergarten Behaviour Scales and to examine affect of the promoter education program* [Unpublished doctoral dissertation] Gazi University.

Özmen, D. (2013). *The analisis of peer relations of 5-6 years old children in terms of social problem solving abilities* [Unpublished master's thesis]. Selcuk University.

Öztürk, D. (2009). *The effects of friendship making skills training with board game on friendship making skills of fourth grade elementary school students* [Unpublished master's thesis]. Middle East Technical University.

Öztürk, N., & Kutlu, M. (2017). The impact of friendship skills psycho-education on the friendship quality of 9-12 year-old students. *Education and Science 42*(191), 397–413. https://doi.org/10.15390/EB.2017.7030

Page, T., & Bretherton, I. (2001). Mother- and father-child attachment themes in the story completions of pre-schoolers from post-divorce families: Do they predict relationships with peers and teachers? *Attachment and Human Development, 3*(1), 1–29. https://doi.org/10.1080/713761897

Park, K.A., & Waters, E. (1989). Security of attachment and preschool friendships. *Child Development, 60*(5), 1067–1081. https://doi.org/10.2307/1130781

Parker, J.G., & Seal, J. (1996). Forming, losing, renewing, and replacing friendships: Applying temporal parameters to the assessment of children's friendship experiences. *Child Development, 67,* 2248–2268. https://doi.org/10.2307/1131621

Rose, A.J., & Asher, S.R. (2000). Children's friendships. In C. Hendrick & S.S. Hendrick (Eds.), *Close relationships: A sourcebook* (pp. 47–57). Sage.

Proulx, M., & Poulin, F. (2013). Stability and change in kindergartners' friendships: Examination of links with social functioning. *Social Development, 22*(1), 111–125. https://doi.org/10.1111/sode.12001

Rydell, A., Bohlin, G., & Thorell, L.B. (2005). Representations of attachment to parents and shyness as predictors of children's relationships with teachers and peer competence in preschool. *Attachment and Human Development, 7,* 187-204. https://doi.org/10.1080/14616730500134282

San Juan, R.R. (2006). *Studying preschool friendship quality: A story completion task to examine young children's mental models of a specific best friendship* [Unpublished doctoral dissertation]. University of Wisconsin-Madison.

Sebanc, A.M. (2003). The friendship features of preschool children: Links with prosocial behavior and aggression. *Social Development, 12,* 249-268. https://doi.org/10.1111/1467-9507.00232

Simpkins, S.D., & Parke, R.D. (2002). Do friends and nonfriends behave differently? A social relations analysis of children's behavior. *Merrill-Palmer Quarterly, 48*, 263–283. https://www.jstor.org/stable/23093770

Shaver, P.R., Collins, N., & Clark, C.R. (1996). Attachment styles and internal working models of self and relationship partners. In G.J. Fletcher & J. Fitness (Eds.), *Knowledge structures in close relationships: A social psychological approach* (pp. 25-62). Lawrence Erlbaum Associates.

Stievenart, M., Roskam, I., Meunier, J.C., & van de Moortele, G. (2011). The reciprocal relation between children's attachment representations and their cognitive ability. International *Journal of Behavioral Development, 35*(1), 58-66. https://doi.org/10.1177/0165025410370790

Sullivan, H.S. (1953). *The interpersonal theory of psychiatry.* W. W. Norton & Company.

Ulutaş, A. (2016). The effect of the game-based training program on peer relations among five-year-old children receiving preschool education. *International Journal of Social Science, 42,* 61–74. https://doi.org/10.9761/JASSS3169

Von Klitzing, K., Stadelmann S., & Perren, S. (2007) Story stem narratives of clinical and normal kindergarten children: Are content and performance associated with children's social competence?. *Attachment & Human Development, 9*(3), 271-286, https://doi.org/10.1080/14616730701455445

Vu, J.A. (2015). Children's representations of relationships with mothers, teachers, and friends, and associations with social competence. *Early Child Development and Care, 185*(10), 1695–1713. https://doi.org/10.1080/03004430.2015.1022538

Warren, S.L., Oppenheim, D., & Emde, R.N. (1996). Can emotions and themes in children's play predict behavior problems? *Child and Adolescent Psychiatry, 35*, 1331–1337. https://doi.org/10.1097/00004583-199610000-00020

Wojslawowicz, J.C., Rubin, K.H., Burgess, K.B., Booth-LaForce, C., & Rose-Krasnor, L. (2006). Behavioral characteristics associated with stable and fluid best friendship patterns in middle childhood. *Merrill-Palmer Quarterly, 52*, 671-693. https://doi.org/10.1353/mpq.2006.0000

Wolcott, C.S., Williford, A.P., & Hartz Mandell, K. (2019): The validity of narratives for understanding children's perceptions of the teacher–child relationship for preschoolers who display elevated disruptive behaviors. *Early Education and Development, 30*(7), 887–912. https://doi.org/10.1080/10409289.2018.1556547

Yoleri, S. (2015). Preschool children's school adjustment: indicators of behaviour problems, gender, and peer victimisation. *Education 3-13, 43*(6), 630-640. https://doi.org/10.1080/03004279.2013.848915

Youngblade, L.M., & Belsky, J. (1992). Parent–child antecedents of 5-year-olds' close friendships: A longitudinal analysis. *Developmental Psychology, 28*, 700–713. https://doi.org/10.1037/0012-1649.28.4.700

Youngblade, L.M., Park, K.A., & Belsky, J. (1993). Measurement of young children's close friendship: A comparison of two independent assessment systems and their associations with attachment security. *International Journal of Behavioral Development, 16*, 563–587. https://doi.org/10.1177/016502549301600403

APPENDIX

Interviewer Protocol - Friendship Story Task

Materials used during

Small figures to represent target child, best friend, peer

Piece of felt to represent different settings for each story stems

Green = Playground

Tan = Sandbox

Red = Classroom

Props

Birthday Cake, Zoo Animals, Blocks, Wagon, Bikes

Obtaining the child's assent/introducing figures to child

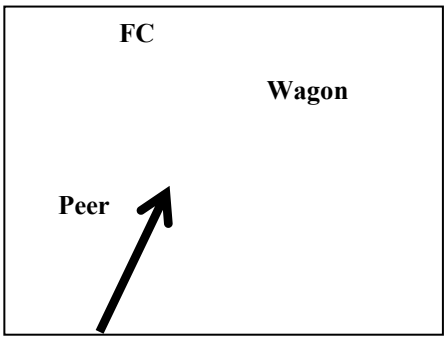
Say to child:

1. *"Today, I'd like you to help me make up some stories about your friend, (BF), and you. Would you like to help me make up some stories? (If child assents). Great, so let me show you some things that we're going to use to help us make up our stories. We have some people here who are going to be in our stories. (Stand three child figures). I want you to tell me which one is always going to be you in our stories (allow the child to choose). Great, so in our stories, this one will always be you. Now, which one is going to be BF (allow the child to choose). Great, so in our stories, this one will always be BF). We're also going to have one more kid in our stories. This child will always be Mert/Ayşe in our stories.*
2. *Now, let's see if you remember who each person is? (Point to each figure and ask) Who is this? (repeat each figure's name as child identifies). So we have, (name each figure)*

Great! Now, let's make up our first story. For each story, I 'm going to start the story, and then I want you to finish the story for me.

Who Gets to Ride? - Props used: FC, BF, peer figures, small red wagon, green felt to represent playground

Figure and prop configuration for Who Gets to Ride?



You are pulling BF in the wagon. Mert/Ayşe walks up to you and asks, "CHILD, can I have a turn being pulled in the wagon by you?" BF gets mad and says to you, "No! You said you were only going to pull me in the wagon!" Show me and tell me what happens next.

Prompt only once if child does not address how FC handles who gets to be pulled in the wagon: "What do they do about who gets to be pulled in the wagon?"

After child completes story:

Are you ready to go to the next story? For the next story, you and BF are playing here (piece of felt) at school. We'll leave Mert/Ayşe over here (place Mert/Ayşe figure off to side, visible to child, but not part of main action in story).

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 2, 376–396
https://doi.org/10.21449/ijate.949160
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Towards an Online Self-Assessment for Informed Study Decisions–A Mixed-Methods Validation Study

Laurie E. C. Delnoij[1,2*], **José P. W. Janssen**[1], **Kim J. H. Dirkx**[1,3], **Rob L. Martens**[1]

[1]Open Universiteit, Faculty of Educational Sciences, Valkenburgerweg, 177, 6419 AT, Heerlen, the Netherlands
[2]Maastricht University, School of Business and Economics, Department of Educational Research and Development, Tongersestraat 53, 6211 LM, Maastricht, the Netherlands
[3]Zuyd Hogeschool, Nieuw Eyckholt 300, 6419 DJ Heerlen, the Netherlands

ARTICLE HISTORY
Received: Sep. 09, 2021
Revised: Mar. 28, 2022
Accepted: Apr. 08, 2022

Keywords:
Study decisions,
Self-assessment,
Validity,
Mixed-methods,
Higher education.

Abstract: Informed study decisions are pivotal for student retention in higher online education. A self-assessment prior to enrolment has been proposed as a promising approach to enable informed decision-making and to build resources for retention. To determine whether such a self-assessment affects the decision-making process as intended, thorough and careful validation is a necessity. This study reports on two validity aspects that are less commonly addressed in that respect, but essential for evaluating effectiveness: response processes and consequences of (self-) testing. To map the response processes and consequences of the current self-assessment, a mixed-methods approach was used in which eight prospective students took a self-assessment in an observed think-aloud mode and were interviewed before and after that. Results show different response processes depending on the type of subtest that is taken. The results also indicate that consequential aspect of validity must be considered in the context of decision-making phases. The demonstrated evidence and possible threats to validity are discussed in light of refining the self-assessment and embedding it in counselling practice.

1. INTRODUCTION

Adequate, personalized information is pivotal for prospective students to make a well-informed study decision, to stay motivated, and to successfully complete their studies (Nicol, 2009; O'Regan et al., 2016; Tinto, 2005; Vossensteyn et al., 2015). Self-assessments prior to student enrolment can provide such information and are increasingly deployed for informed decision-making (Kubinger, et al., 2012; Nolden & Wosnitza, 2016; Nolden et al., 2019). To determine whether such assessment instruments fulfil their purpose, empirical evaluation is necessary, especially since the use of these instruments can have important consequences for individual decision making and student enrolment. However, empirical evidence is often implicit or completely lacking (Niessen & Meijer, 2017). We argue that such self-assessments should be validated explicitly and as fully as (standardized) summative assessments as well that such

*CONTACT: Laurie E. C. Delnoij ✉ l.delnoij@maastrichtuniversity.nl Maastricht University, School of Business and Economics, Department of Educational Research and Development, Tongersestraat 53, 6211 LM, Maastricht, the Netherlands.

e-ISSN: 2148-7456 /© **IJATE** 2022

validations yield important scientific information that can bring the field a step further. For that purpose, with this study, we show one step in the validation process of such a self-assessment in the context of open online higher education.

1.1. Self-assessments for Informed Study Decisions

Self-assessments for informed study decisions are advisory and informative instruments conducive to self-examination (Hornke et al., 2013). In general, these instruments aim to elicit reflection on study preparedness by informing prospective students about where they stand in regard to the demands of studying in higher education. One example is a self-reflection tool developed by Nolden et al., (2019). In this instrument, prospective students complete tests and receive feedback on, for instance, self-discipline, learning strategies, and emotional stability. In the feedback, respondents get information about how they scored in comparison to other students. In case the results indicate issues (e.g. lack of self-discipline), access to remediation is offered by topic-specific recommendations and information about university's support services. In another example, prospective students complete similar tests and receive program-specific feedback focused on their chances of success after enrolment (Broos et al., 2018; 2019; Fonteyne & Duyck, 2015). As self-assessments seems beneficial for retention, we also developed such a self-assessment (Delnoij et al., 2020a; Delnoij et al., 2020b; Delnoij et al., 2021). This self-assessment entails three categories of subtests (i.e. knowledge/skills, attitude, and social situation), which have shown to be predictive of obtaining study credits in the context of higher online education (Delnoij et al., 2021). Feedback is provided after each subtest and includes concrete tips and opportunities for remediation, to address possible risks for non-completion early (Delnoij et al., 2020b). Note that our self-assessment is generic; it does not differentiate between or provide an advice for specific study directions. Comparable to the examples given above, the self-assessment not committal and not aimed at selecting students. Rather, the aim is to enable informed decision-making (food for thought), and to encourage prospective students to start well-prepared (feedback for action).

1.2. (The quest for) Validity

These aims pose high demands on assessment validity, i.e. do the test scores, the feedback provided in relation to them, and prospective students' interpretations thereof and following actions all match the proposed use of the assessment?

Hence, to develop an effective self-assessment and feedback (hereafter called 'SA'), it is important to collect and evaluate sources of validity evidence. In the literature, five sources of validity evidence can be distinguished (American Educational Research Association [AERA] et al., 1999): content, predictive power, internal structure, response process, and consequences (effects). Investigating these five sources of validity evidence is not a 'once and for all' activity, but one that requires regular attention, as student populations and/or educational practice may evolve over time (Messick, 1988; Royal, 2017). However, a chronological order appears to exist when it comes to collecting evidences from these sources: investigating response processes and consequences makes sense only after the content, internal structure, and predictive power have been more or less secured.

So far, applied validation studies tend to mainly focus on the first three (Cook et al., 2014), also in the specific context of study decision making instruments. More specific, for self-assessments prior to student enrolment, the determination of which tests to include (content aspect of validity), their internal structure, and predictive value (e.g. for retention after enrolment) are often theory- and data-driven (e.g. see Nolden et al., 2019). However, scientific attention is lacking for how prospective students actually proceed through such instruments (response processes) and how these instruments affect their study decision (consequences). To

create a complete picture of the self-assessments' effectiveness, these validity aspects cannot be ignored (AERA et al., 2014; Cook et al., 2014).

Having established satisfactory results regarding content, internal structure and predictive aspects of validity in previous studies (Delnoij et al., 2020a; Delnoij et al, 2020b; Delnoij et al, 2021), the present study aims to investigate *response processes* and *consequences* of a self-assessment for informed study decisions.

1.3. Process and Consequential Aspects of Validity

The process aspect of validity comprises theoretical and empirical analyses evaluating how well test takers' actions (responses) align with the intended construct (Cook et al., 2014). The focus is on users' response processes, including the *actions, thought processes*, and *strategies* of individual respondents while taking the assessment (Beckman, et al., 2005). *Actions* provide insight into whether prospective students use the SA as intended. In the present study, we focus on the selection and order of subtests taken and the extent to which feedback information is consulted. Respondents' actions are often studied through observation (Cook et al., 2014; Goodwin & Leech, 2003). Additionally, by asking respondents to think-aloud, their *thought processes* (i.e. considerations for providing certain answers) and *reactions* (on a specific test or its items) can be investigated by interviews or asking respondents to think-aloud while they are taking the self-assessment (Cohen, 2006; Cook et al., 2014; Goodwin & Leech, 2003; Kutlu & Yavuz, 2019). In (concurrent) thinking aloud, participants verbalize their thoughts as they complete a task (Van den Haak et al., 2003). This research method has proved a valid source of data about participants' thinking (Charters, 2003). For securing trustworthiness, follow-up interview questions are proposed, to capture as many of respondents' experiences as possible and to validate researchers' interpretations of participants think-aloud verbalizations (Charters, 2003; Padilla & Benítez, 2014).

Using these methods, valid *strategies* to complete subtests can be estimated (Cohen, 2006; Kutlu & Yavuz, 2019; Padilla & Benítez, 2014). This is important as the validity of strategies depends the content and format of a test (Cohen, 2006). For cognitive tests (i.e., testing knowledge or skills, answers are right or wrong), for example, strategies such as cheating and guessing are clearly flawed (Cook et al., 2014). On the other hand, a common valid test taking strategy is to go back to a specific question or item for clarification (rereading or paraphrasing) (Cohen, 2006). Test-taking strategies may also be flawed by specific measurement techniques. Non-cognitive tests (i.e. measuring attitude or affect) involve test-takers to classify themselves in which self-knowledge and experience is called upon. Such self-report measures, in general, are more prone to socially desirable answers, especially in high-stakes contexts (Cook et al., 2014; Niessen et al., 2017). The relative 'low-risk', non-committal nature of the SA can be expected to reduce socially desirable answers. Nevertheless, investigating variations in response processes may reveal relevant evidence for the process aspect of validity and threats in the sense of variance that is irrelevant to the constructs being measured or the purpose of the SA (Downing & Haladyna, 2004). Thus, results gained from studying prospective students' response processes may reveal relevant implications for development and improvement of the SA.

A second focus of this study is the *consequential aspect* of the SA's validity. Though added later as a distinct source of validity evidence, the literature shows that the consequential aspect of validity is solidly embodied in the current Standards (AERA et al., 1999; Downing, 2003). The consequential aspect of validity pertains to anticipated and unanticipated consequences – both positive and negative – of measurement (Cook et al., 2014; Downing, 2003; Goodwin & Leech, 2003), which can support or challenge the validity of score interpretations and actions based upon them (Beckman et al., 2005). Consequence evidence can be evaluated both from an individual and societal perspective (St-Onge et al., 2016). In the context of the current SA,

anticipated individual consequences range from interpretations of the scores and feedback to the decision on whether or not to enroll. The extent to which consequences are valid requires interpretation of the context in which the consequences occur. Increased levels of study choice certainty, for example, are a valid consequence if one scores well on the SA. In this particular context, feeling affirmed in an already certain choice can also be considered valid. A valid consequence to a poor score would be (the intention) to take remedial measures as a follow up on the feedback or even to postpone or reconsider the study decision. Though of course, in the context of open education, we want to be particularly careful not to unnecessarily discourage prospective students. At a societal level, the anticipated consequence is a positive impact of the SA on completion rates. The latter, impact on completion rates, requires 'mainstream' deployment of the SA. Prior to the decision for a 'full release' of the SA, (i.e. making it available and evaluate it on a large scale), investigating individual consequences will help to shed light on the question whether the anticipated effects of the SA such as taking remedial measures, postponing and/or reconsidering enrolment, and study choice certainty are evoked as intended.

In the present study, the focus is on the consequences of the SA on the individual level. This means we investigate how prospective students respond on obtained scores and feedback, the extent to which they intend to follow up on the feedback they receive, as well as possible impact on their study choice and certainty thereof.

1.4. Research Questions

The transition and access to higher (online) education requires the best possible support for students in making a study decision. Therefore, self-assessments deployed for that purpose should be thoroughly validated. With this study, we aim to contribute to a standard for such validation processes by zooming in on two aspects of validity that have not received much attention in validation studies so far, but are important in determining the effectivity of such self-assessments (Cook et al., 2014; AERA et al., 2014): response processes and consequences of testing. The resulting evidence and threats to validity provide insight for the (re)design of a self-assessment for informed study decisions. In other words, we aim to answer the following central research question:

What evidence and threats to process and consequential aspects of validity do we find for the self-assessment and what implications does this have for its design?

To answer the central research question, several sub questions are formulated. Questions establishing a baseline/context:

- RQ1. *What are prospective students' expectations regarding the impact of the SA?*
- RQ2. *What are prospective students' obtained scores on the subtests of the SA?*

Questions regarding the response process, i.e. how prospective students proceed through the SA:

- RQ3. *Which tests are selected, in what order and which feedback is consulted while taking the SA and why?*
- RQ4. *What reactions are elicited while taking the SA?*

Questions regarding consequences: interpretations, intentions, decisions:

- RQ5. *How do prospective students respond to obtained scores and the feedback they receive?*
- RQ6. *To what extent do prospective students plan to follow up feedback provided, and what reasons do they have for this?*
- RQ7. *How does the SA affect prospective students' study choice and certainty thereof?*

2. METHOD

2.1. Context

The SA is designed and developed for prospective students of the Open University of the Netherlands (OUNL), which provides academic courses as well as full bachelor and master programs, mainly online, occasionally combined with face-to-face meetings. The open access policy of OUNL means that the only entry requirement is a minimum age of 18 years (though naturally, additional entry requirements may be formulated for more advanced courses).

2.2. Research Design

The present study represents a particular step in the design-based research approach, typically comprising iterative stages of analyses, design, development, and evaluation (Van den Akker et al., 2013). More particularly, this study evaluates evidence for response process and consequences through a convergent mixed-methods design (Creswell, 2014) involving observation, think-aloud and semi-structured interviews.

Quantitative data were collected through the subtests, observation and the semi-structured interviews. These data include the obtained subtest scores (RQ2), the number and order in which subtests were taken, consultation of feedback (RQ3), and study choice certainty expressed on a scale of 0 (certain not to enroll) to 10 (certain about enrolling)(RQ7).

Qualitative data were collected through think-aloud as well as semi-structured interviews. These data involve prospective students' expectations of SA's impact (RQ1), their reactions on the subtests (RQ4), their response to obtained scores and feedback (RQ5), and their reflections regarding consequences of the SA (RQ6 and 7).

2.3. Materials

In this section, we describe the SA (prototype), observation and think-aloud protocol as well as the semi-structured interview protocol.

2.3.1. *Self-assessment prototype*

The SA prototype, illustrated in Figure 1, consists of four constituent tests, completion of which results in a score and related feedback per subtest. The subtests measure numerical skills, discipline, social support, and hours planned to study (Delnoij et al., 2021). The numerical skills subtest involves nine items in either multiple choice or open-ended formats. One example item is '*Which of the following options is less than 1?*' with five answer options in which respondents have to add two fractions. The discipline subtest consists of three items on a 7-point scale ranging from totally disagree to totally agree. For instance, '*I find it hard to stick to a study schedule'*. Social support entails one item asking prospective students to indicate for three sources of social support (financial, emotional, practical) whether they receive this from their environment (i.e. partner, family, friends, co-workers, and/or employer). Examples for the three support sources are given and respondents can select multiple answers or a 'none of the above'-option. Hours planned to study is measured by a multiple-choice question with categorical answer options such as 0-5 or 6-10 hours per week.

The feedback design is based on related work in other contexts (Broos et al., 2018; 2019; Fonteyne & Duyck, 2015; Jivet et al., 2020; Nolden et al., 2019) and further informed by the results of an initial user study (Delnoij et al., 2020b). The feedback consists of three components: information on the obtained score, information on the test (what was measured and why), and an advice for further preparation (e.g., general tips, services and contact information of study advisors and opportunities for remediating tutorials at the OUNL.

Information on the obtained score is communicated by means of a visualization in which the obtained score, indicated by an arrow, is projected on a bar representing the possible range of

scores (scale of 0 – 100%). The color in the bar fades from white ('high risk' area) via light green ('medium risk' area) to dark green ('low risk' area) indicating increased odds of obtaining study credits. After completing a test, the arrow in the bar is presented on the overall self-assessment dashboard, additional feedback information can be consulted by clicking the result button that appears alongside (see Figure 1, C-E).

2.3.2. *Observation & think-aloud protocol*

To observe participants while taking the SA they were asked to share their screen, so that the following actions, related to the process aspect of validity, could be captured: number and order of subtests taken, feedback consultation (i.e., do prospective students consult the feedback or not and, if so, how quickly do they seem to go through it?). A think-aloud protocol was carried out to capture participants' test-taking strategies and reactions while taking the subtests (process aspect of validity) and gain insight into how they respond to their obtained scores and feedback (consequential aspect of validity). We based our think-aloud protocol on previous (related) work (e.g. Charters, 2003; Padilla & Benítez, 2014). In the present study, participants were instructed to express aloud anything coming to mind while taking the SA (e.g., considerations regarding the order in which they filled out the tests, spontaneous feelings and reactions evoked by the test items) and while consulting the obtained score and the feedback provided alongside. Furthermore, it was stressed to participants that it was the SA that was being tested in the present study, not them. Before the actual think-aloud procedure was carried out, it was briefly exercised to allow participants to become familiar with it. The protocol further contained the instruction that in case participants remained quiet for 5 seconds or longer, the researcher should kindly remind them to think-aloud, by asking *'What are you thinking right now?'*. The think-aloud procedure stopped when participants indicated that they had finished taking the subtests of their choice. Subsequently, questions were asked to validate the researcher's interpretation of the think-aloud utterances as a source of triangulation (Charters, 2003). After that, the researcher moved on to the interview questions on participants' experiences with the SA as described in the next section.

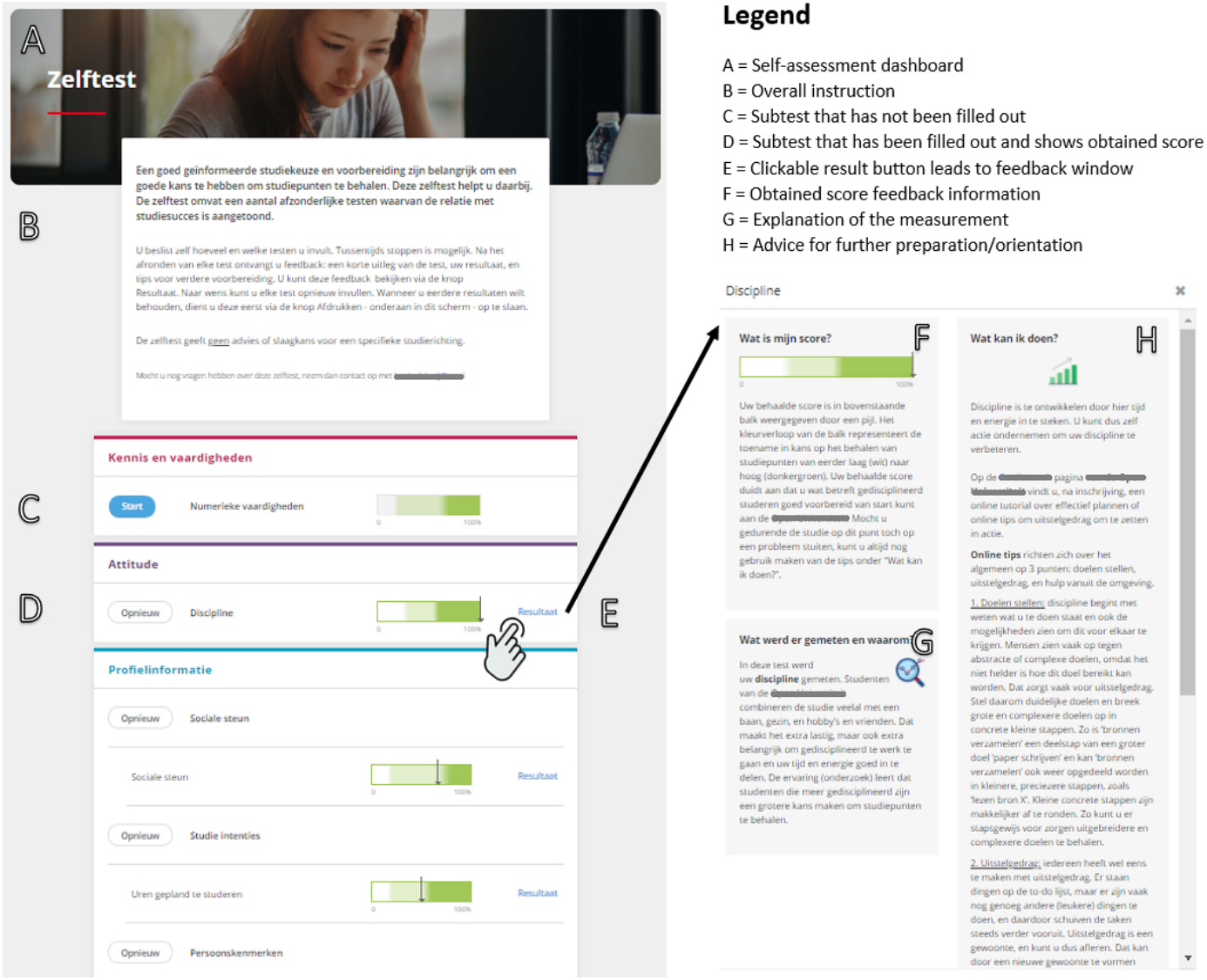
2.3.3. *Semi-structured interview protocol*

The interview protocol consisted of instructions for the interviewer (i.e., steps to take prior to the interview), instructions for the participant (e.g., there are no right or wrong answers, try to be as complete and honest as possible in answering the questions), and a list of pre-defined questions on which follow-up questions were asked if necessary. Pre-defined questions were formulated with a focus on both participants' expectancies prior to taking the SA, (e.g., *If so, to what extent do you expect an impact of the SA on your study choice?*) and their thoughts and reflections after taking the SA (e.g., *If any, which follow up actions will you be taking, based on the SA?*). Prospective students' certainty of their study decision was measured on a scale of 0 (certain not to enroll) to 10 (certain to enroll) both prior to and after taking the SA.

2.4. Participants

Eight prospective students participated in this study (6 Female, M_{age} = 36.25). One participant was interested in following a course, the other seven in following a full study program. Five participants were interested in the domain of law, two in management sciences and one in psychology. All, but one participants already possessed a degree in higher education (university of applied sciences).

Figure 1. *Prototypical Self-assessment.*



2.5. Procedure

2.5.1. *Sampling procedure*

Sampling took place in June and July of 2020. Prospective students who indicated their interest for a course or study program at the OUNL (e.g., by calling the service and information department for information on a certain course) were informed about the study and invited to leave their e-mail address if interested in participating. They received the information letter and link to the online consent form via e-mail. After signing the consent form, an appointment was made.

2.5.2. *Research procedure*

The sampling procedure was carried out after obtaining ethical approval of the study. In the meantime, a pilot session was conducted to test the research procedure and the latest prototype of the SA. When it comes down to trustworthiness of qualitative research, pilot tests contribute to enhancing credibility and confirmability (Guba, 1981; Krefting, 1990; Shenton, 2004). Based on this pilot session, no adjustments were made for the research protocol. The textual feedback provided with some of the subtests was adapted in order to make it more concise, without loss of content.

The research took place in BlackBoard Collaborate©, an online virtual conferencing tool providing functionalities for video calling (i.e., sharing camera and microphone) and virtual lectures (i.e., screen sharing, sharing content). In this session, participants first received explanations on the content and duration of the session. Any additional questions were answered after which the researcher inquired participants' expectations of the self-assessment. Next, the think-aloud procedure was practiced in a mock test very similar to those in the actual SA. Subsequently, participants were instructed login into the online SA environment, upon which the actual think-aloud procedure began. Participants were instructed to notify the researcher once they had taken the tests they wanted to take and read all the information they wanted to read. Afterwards, the follow-up interview took place. Finally, the researcher answered remaining questions and thanked participants for taking part in the study. Participants received a portable document format (PDF) of their obtained SA scores and feedback. All sessions (including the pilot) were recorded (of which participants were informed in the information letter and again during the session).

2.6. Analysis

The mixed-methods design of this study involved collection of various data, both quantitative and qualitative. The expected impact of the SA (RQ1), obtained subtest scores (RQ2), total number of subtests taken and feedback consulted (RQ3), intended follow-up actions (RQ6), and study choice certainty (RQ7) are summarized in descriptives. Participants' reactions while taking the SA (RQ4), responses to obtained scores and feedback (RQ5), and further reflections (RQ6 and 7) are analyzed using qualitative content analysis.

2.6.1. *Qualitative content analysis*

As a starting point of the qualitative data analysis, audio recordings were transcribed verbatim. All transcripts were first read in depth to allow familiarization with the data. Next an iterative coding process took place. Two researchers coded one part of the data separately first. For securing credibility and confirmability (Guba, 1981; Shenton, 2004), they discussed their coding results together and with a third researcher. Initial categories of codes and themes of categories emerged from this discussion. Based on that, the principal investigator coded the rest of the data. Ambiguities were solved in consultation with the other two researchers. The coding process was carried out in accordance to the steps of qualitative content analysis as described by Erlingsson and Brysiewicz (2017). The first step in that process was to split up the data in

(condensed) meaning units: a short text fragment, in which the core meaning is retained. These condensed meaning units were coded. A code is a label that most accurately describes what a condensed meaning unit is about, usually in 1 or 2 words. For example, "It has been a long time since I have had to keep track of such a schedule, so I don't know" was coded as "Lack of recent experience" and "I don't fully trust my own answers" was coded as "(Possibly) flawed answering". After that, codes were grouped into categories, e.g. a group of codes that are related to each other through content or context and is usually factual and short. For instance, the codes "Lack of recent experience" and "(Possibly) flawed answering" were grouped together as "Process threat". Subsequently, we inspected categories to elicit the main themes. These themes express an underlying meaning of 2 or more categories, and are descriptive in name. As an example, "Process threat" and "Process evidence" were grouped together as "Process aspect of validity".

3. RESULT

3.1. Expectations of SA Impact (RQ1)

Table 1 provides a summary of whether or not an impact of the SA on study choice was expected. Four participants did not expect the SA to have much impact on their study decision e.g., because they already had gone through an extensive orientation process, expressed as "I would say the assessment will not have much influence on my decision, as I already did a lot of research" (participant P†). Nevertheless, it can help improve their understanding of what studying in the specific educational context will entail. Participant L mentioned this as following: "I will definitely continue the study decision I already made, but then at least I will have a better picture of the time and effort it would cost me."

Four participants expected the SA to have an impact on their study decision in the sense that they are seeking affirmation on whether or not they are making the 'right' decision. Participant J said, "That I get a kind of confirmation whether or not my decision is a good idea" and participant E stated "Either a confirmation of what you already have in mind or of your insecurities and, therefore, a confirmation to look further and choose something else".

Table 1. *Overview of expected impact, test taking behaviour, obtained scores, feedback consultation and study choice certainty.*

	Participant							
	J	P	L	Y	I	K	E	Z
Impact on study choice expected	yes	no	no	no	no	yes	yes	yes
Test taking order[1] obtained score[2] and feedback consultation[3] per subtest								
Numerical skills	1 ✓	4 ✓	1 ✓	1 ✓	1	1	1	1 ✓
Discipline	2 ✓	1 ✓	2 ✓	2 ✓	2	2	2	2 ✓
Social support	3 ✓	2 ✓	3	3	3	3	3	3
Hours planned to study	4 ✓	3 ✓	4	4 ✓	4	4	4	4
Study choice certainty								
Prior to SA	5.0	8.0	7.0	10.0	8.0	7.0	10.0	7.0
After SA	7.0	8.0	7.0	10.0	8.0	7.0	10.0	7.0

[1] 1…4 Order of test taking from 1 (first test taken) to 4 (last test taken)

[2] 'high risk' score 'medium risk' score 'low risk' score

[3] ✓ Feedback consulted

† All participants were given an anonymous identifier, obtained via Randomwordgenerator©.

3.2. Obtained Scores (RQ2)

A summary of the obtained subtests scores is provided in Table 1. Overall, participants' scores were in the (relatively) safe areas on most subtests. One participant obtained a score in the 'high risk area' on the numerical skills test.

3.3. Test Taking Behavior and Feedback Consultation (RQ3)

A summary of the number and order of subtests taken and feedback consultation is provided in Table 1.

3.3.1. *Number of subtests taken*

Even though participants were instructed to be in charge of which subtests they would take and in which order, all participants completed all subtests. This is remarkable, as some participants commented that in particular the numerical skills did not seem relevant to them. Reasons for still taking this test were the few subtests in the SA:

> Normally I would have skipped the numerical skills test, as I do not think it is relevant for my study decision (…). Now I filled it out, because there were not that many tests and the other tests did not consist of many questions, so I decided to see what insights the numerical skills test might provide me. (Participant P)

And the lack of clarity (despite instruction) that it was possible to skip subtests: "I thought I had to fulfil it, or I would not be able to continue with other tests" (Participant L).

3.3.2. *Order of taking subtests*

In general, participants took the tests in the order in which they were presented from top to bottom. The (incidental) reason to diverge from this order was the drive to first take the test they felt most insecure about: "Study intentions grasps my attention, as I know that, traditionally, I have the most trouble with that. That is why I am going to start with that one" (Participant P).

3.3.3. *Feedback consultation*

Two participants consulted the feedback on all subtests. Three participants did not consult any of the feedback information, as they did not notice the result button: "I really did not see the button; otherwise I would have clicked on it. I would really like to see it now" (Participant I). Though instructed about the button, apparently the button was not clear to all users.

Furthermore, three participants consulted the feedback only for some of the subtests. In those cases, feedback on social support and/or hours planned to study was neglected. These students did score relatively well on these tests, which was also mentioned as the main reason to skip the feedback: "Well, what else can I do? I ticked all the boxes (…) so I thought there is nothing to improve or do, it is fine like this and I feel comfortable with that" (Participant Y).

3.4. Reactions during Test Taking and Responses to Feedback (RQ4 and 5)

In this section, we discuss reactions during test taking (process aspect of validity) and how participants responded to their obtained scores and feedback (consequential aspect of validity) per subtest, before discussing these results for the SA in general.

3.4.1. *Numerical skills*

3.4.1.1. Process Aspect of Validity. For many participants the numerical skills test gave rise to feelings of insecurity (e.g., test-anxiety, feeling incompetent), both in advance and while taking the test. This became clear from actual statements uttered (e.g., "I will never manage this, I am so bad at mental arithmetic" (Participant L)), as well as other signals: repeatedly sighing, scrolling up and down, indicating that the test will take a long time or that by looking at how many questions still have to be filled out. For some, this test raised awareness that these

skills may be important, for many the test created feelings of frustration and/or doubts about the relevance of this test. For instance for participant P, stating, "I am surprised about the math exercises, it has little to do with the study I am interested in".

Feelings of insecurity bring forward different strategies for completing the test. One person mentioned to read extra carefully and write things down, because of finding it difficult (i.e., "Ok, fractions (…) I find that hard, so I'll have a closer look at it" (Participant I)). However, quite a few (n = 5), remarked that they just guessed some answers in order to complete the test. Furthermore, striking about this test was that, in contrast to the other tests, almost half of the participants felt ill at ease because the researcher was observing how they proceed through the test. Two participants even mentioned that, because of this, they filled it in at speed, at the expense of accuracy.

3.4.1.2. Consequential Aspect of Validity. Although the test tended to evoke frustration, insecurities, and invalid answering strategies (hurrying, guessing), the responses on the scores and feedback were rather positive. The most common reaction was relief regarding the obtained score: "I never took math classes or anything like that, so this is not so bad" (Participant Y). Two participants had expected to score better, while four had expected to score lower than they actually did. This appeared to raise their confidence regarding their own abilities: "That is interesting, I believe I can do this" (Participant J). The feedback also resulted in reflection on the relevance of numerical skills and two participants intended to consider the possibilities for further preparation (quote 15). As participant P stated, "Apparently there is a correlation between numerical skills and obtaining study credits, I did not know that. I clicked on a link to read more about that". One person maintained her opinion that the test was not relevant for the specific study direction she was interested in, and therefore did not recognize the added value.

One participant (L) scored in the 'high risk area' on the numerical skills test. When she read the feedback, she understood that her score related to lower chances for obtaining study credits, which she mentioned as the reason for feeling a bit discouraged. Her score did not surprise her, because she always experienced problems about arithmetic, which she also expressed when taking the subtest. While reflecting on the feedback she mentioned to feel scared, though generally hopeful, because she scored well on the other tests and would not have to do that much with numerical skills in her study direction of interest, i.e., law.

3.4.2. *Discipline*

3.4.2.1. Process Aspect of Validity. In general, during this test, participants verbalized their reasoning towards an answer, for instance how they based it on previous or similar (study) situations. They also indicate to be aware that it can be hard to stay disciplined when, for example, there are other, more enjoyable, things to do. One participant said she found it difficult to answer the questions, as she had no recent or similar experiences to draw from. This test was the only test in the SA in which a possible response flaw became apparent with one participant commenting that he did not fully trust his own answers. His score was sufficient and he indicated that he tried to answer as honestly as possible, but also knows that this might turn out to be a problem.

3.4.2.2. Consequential Aspect of Validity. One person scored lower than expected on the discipline test. This made her doubt her own answers on the test. After all, she did see herself as a disciplined person. In general, however, the discipline test results mainly reflected participants' self-views: "Yes, of course in dark green [visualization of the score], I knew that already" (Participant J). They went through this feedback faster, compared to the feedback on the numerical skills test. One person mentioned that he merely made a quick scan with the intention to read it more carefully if the feedback would mention something surprising.

3.4.3. *Social support*

3.4.3.1. Process Aspect of Validity. For five participants the test prompted adequate reflections in regard to social support. They summarized, for instance, which persons in their environment they had already discussed support with:

> My parents want to support me financially. Emotionally as well, there is lot of interest in what I do. Practically, I think so, I don't have children [*example given in the test*], but I think if I have to cancel things that people will understand that I have to study. (Participant Z)

3.4.3.2. Consequential Aspect of Validity. For one person this test was quite confronting, in the sense that it made her aware of the fact that she really has to do it on her own. For others the test was a confirmation of what they had already considered. Specifically in regard to social support, an interesting observation was that a maximum score triggered two opposite effects regarding feedback consultation. For one person, obtaining the maximum score was a reason to skip the feedback, as there is no room for improvement, whereas another person nevertheless wanted to see what the feedback said. In general, the feedback on this test evokes further reflection. For example, they think about previous studies they have done and what kind of support was helpful to them then. They also think about whether they have secured all types of support or whether they could do anything for further preparation:

> I see that I am prepared quite well, I have talked to people about this. This did not happen overnight, I have weighed things and I also see that especially my husband supports me in this and we will be able to do this. (Participant I)

One participant mentioned that she does not receive all of these sources of social support, but also does not feel a need for them. Thus, her score indicated room for improvement in social support, which was not in line with her personal needs. As a result, she was confused when receiving her obtained score; she began to wonder whether she completed the test correctly.

3.4.4. *Hours planned to study*

3.4.4.1. Process Aspect of Validity. Thoughts expressed by participants while filling out this test indicate that the hours planned to study had already quite extensively been considered prior to taking the test:

> I have already calculated that I have 15 hours to spend on studying. I work 2 days, so 3 days I am free and the children are at school for 5 hours then, so then I have 15 hours to study. (Participant I)

In addition, they did seem to think about the consequences of specific answers, yet that did not distract them from answering honestly: "I think I need to do more in the numbers of hours planned to study but I will stick to the 6-10 hours anyway" (Participant J).

3.4.4.2. Consequential Aspect of Validity. The obtained scores and feedback on this test mainly raised awareness of how long it will take to complete a study program, given the number of hours planned for studying. For this purpose, the feedback includes a calculation example that helps prospective students to gain insight into how long it will take them to complete a study program, based on the number of hours they plan to study (i.e., Participant P: "This is good, an open door really, but I did not calculate it like this yet"). Although for some this means that they will spend a considerable period of time studying, it does not demotivate them: "It was a confirmation. I do like studying, so I do not really care about the nine years. It did not demotivate me, the time indication" (Participant J). For one person, the feedback did not have added value, because she already made the calculation together with a study advisor.

3.4.5. *Overall*

3.4.5.1. Process Aspect of Validity. Even though all tests included in the SA are relevant in terms of 'study preparedness', it was not anticipated that prospective students would take all subtests. Still, participants in this study did take all subtests. Moreover – made overt by the think-aloud protocol – they seem to make an adequate translation of their personal situation and/or self-image into an answer to various test items. The numerical skills test, the only 'cognitive' test included, clearly evoked frustration and stress (i.e., "The stress level goes up for a little with those first questions" (Participant Y)), even though most of the participants scored well on it. To some extent, this is inherent to the content of the test, yet we will have to consider how to minimize this effect, as we do not want to discourage respondents unnecessarily.

3.4.5.2. Consequential Aspect of Validity. In general, it can be said that the SA provides food for thought (e.g., about social support, relevance of numerical skills) and feedback for action (e.g., calculating study time, intentions for further reading). Participants find the feedback clear and praise the headings and links, which makes it easier for them to read. However, some also indicate that they scanned through the feedback quickly and read more intently when seeing something striking.

3.5. Further Orientation and Preparation (RQ6)

Three participants reported that they are planning to take some steps for further orientation or preparation. One participant wanted to gain additional insight into the fit between her interests and a specific study direction, so she planned to discuss this with a study advisor. Two participants mentioned that they will make further inquiries regarding numerical skills, e.g. through links included in the feedback. Other participants indicated that they are not planning to take further steps in orientation. The main reason, mentioned by three participants, is that they do not think it is necessary, because they already took diverse orientation steps. Participants also indicated that it depends on the obtained score whether there is an intention to do something with the feedback:

> It depends 100% on the score to what extent I am inclined to do something with it, because you do want to make it a success and if you see that one success factor is a bit less than others, you want to work on it. (Participant Y)

And they do not feel like their obtained scores indicate that they should take further action:

> I would have, if something surprising resulted from that test. For instance, if discipline would have been low, should you even consider taking a study program focused on self-study? In that case, I would have liked to talk to a student, alumnus, or study advisor. (Participant E)

3.6. Study Choice Certainty (RQ7)

A summary of participants' study choice certainty is provided in Table 1. Most participants in the present study were rather certain already of enrolling in a course or study program at the OUNL. Study choice certainty changed only for the participant reporting a certainty of 5 prior to the SA. She was more certain of the decision to enroll afterwards (7), because her insecurity about numerical skills turned out to be unjustified and the SA raised awareness of the time it would take her to complete a study program: "It is higher than 5 now, because of the confirmation in arithmetic, that I don't have to be insecure about that, and the realization that if it takes me 9 years, I wouldn't mind so much" (Participant J).

In general, the SA did not seem to have an impact on study choice certainty. For some participants, fulfilling the SA took place after what they experienced as an elaborate orientation process. Participants stated that they believe the SA to be of more influence in the beginning of

the orientation process (e.g., Participant P: "If I were still at the beginning of my orientation, then it would still have an influence. Now it is like another drop in a bucket full of water") and that the SA in itself has an impact only on study choice (certainty) as a part of a broader pallet of orientation activities. Three participants indicated that their insecurity lies mainly in the choice of study direction and the SA does not provide any tests on that. It is also noteworthy that two participants (participant Y and I) mentioned that they were planning to just start and see how they experience and perform (in) the first half year.

Though their study choice certainty did not change, five participants (both very certain and not so certain) mentioned they felt affirmed after taking the SA. Participant P, for instance, said "The test could only have affected me negatively, but there were no big red flags to find that. Now it was more an affirmation". Participant Y stated the following:

> Before I started the test, I thought I was not prepared that well and that I had not thoughtvery well about the study I was going to do. Now I think that I actually did think well about it and I have not rushed into things. So this test may have made me even more certain that I have made the right choice.

And participant E stated "If you still have some doubts, the test can remove them and if you are almost certain, the test can give you confidence that you are making the right choice". Three participants mentioned that it did trigger reflection on how to start well-prepared:

> In general, it is a good test (…) It gives you a realistic picture of how much study time you have to put in and how long it will take you and also, that it is important that you think about the financial picture and personal support, so it gives you all kinds of facets to think about. (Participant I)

3.7. Other Validity Evidences

Though the present study was targeted at process and consequence validity, the think aloud and interview data also revealed results on the *content aspect of validity* – the relationship between a test's content and the construct it is intended to measure, referring to themes, wording, and format of items on an assessment instrument (Beckman et al., 2005). In regard to the content of the SA as a whole, participants find the content relevant and understand the choices for the current set of subtests. Nonetheless, they have reservations about specific tests. Regarding the numerical skills test some indicate that they assume that this test is chosen to (partly) measure their intelligence, which they do consider relevant content for the SA. However, several indicate that they would expect another test to measure intelligence (i.e., reasoning skills) instead of or in addition to the current numerical skills test.

The tests on discipline and social support, raised doubts with three participants who thought the number of items the tests relied on was too limited to draw sound conclusions from. In addition, they commented on the formulation of specific test items, e.g., they found it hard to interpret words like 'often' (I often do not finish what I planned, because I feel lazy or tired) 'hard' (I find it hard to stick to a (study) schedule), or receiving support 'to some extent'. Finally, some participants questioned the relevance of the social support test, since it does not take into account to what extent people experience a need for various kinds of support.

Please present the findings/results in this section. This section should give significant results obtained from the study clearly and concisely. Please present the findings/results in this section. This section should give significant results obtained from the study clearly and concisely.

4. DISCUSSION and CONCLUSION

The present study was a mixed method study aimed at investigating the process and consequential aspects of validity of a self-assessment for informed study decisions in higher online education.

Regarding the *process aspect of validity*, a general point of concern is that self-assessments, i.e. self-report measures, may be subject to all kinds of measurement errors, due to inaccurate self-perceptions (Dunning et al., 2004) or social desirable answering (Niessen et al., 2017; Viswesvaran & Ones, 1999). In the present study, one participant hinted at this stating that he did not fully trust his own answers on the discipline test. However, in general, our results demonstrate evidence in support of the process aspect of validity as the think-aloud protocol reveals that prospective students appear to base their answers on adequate (sensible) reflections. This evidence was most prominent in the non-cognitive tests (i.e., discipline, social support, and study intentions): participants brought to mind examples from their personal environment and current or previously experienced circumstances in order to decide which answer to select.

The numerical skills test specifically revealed two typical response processes, arising from feelings of uncertainty that are stirred up by the test. Most participants react on this, by adopting the strategy to fill in the test in a hurry and to guess the answers on questions they cannot answer immediately. Occasionally, this leads participants to the opposite approach: taking their time, writing down calculations and reading questions several times. Though the research context (read: the presence of an observer) may have played a role in this as well, these kind of responses are partly inherent to this type of test (Abbasi & Ghosh, 2020; Dowker et al., 2016; Liebert & Morris, 1967).

The limited number and shortness of tests in the SA appeared to motivate prospective students to take all subtests, even those that initially did not seem relevant to their study of interest. We consider this as an advantage to the process aspect of validity, as all the tests provide relevant insights independent of the study of interest (Delnoij et al., 2021).

An important threat that came to light in the current study is that some users missed the result button. Consequently, they missed important feedback information that can support them in choosing and preparing for a study in higher online education.

With respect to the *consequential aspect of validity* it appears that the SA feedback triggers reflection. The obtained scores and feedback on the numerical skills test were generally positive, in contrast to what some prospective students expected while taking the subtest. The feedback taught them that they could influence their skills by taking time and effort to practice. This resulted in enhanced self-efficacy – a person's sense of their own ability to accomplish something successfully (Bandura, 1977). We see this as an advantage for the consequences of the SA, as self-efficacy is an important determinant for students' motivation and success in higher online education (Harnett, 2016). The feedback on the other tests triggers reflection, in particular tests on social support and hours planned to study. Here, prospective students start to rethink their preparedness and intentions and whether they could do more.

However, the feedback hardly appears to influence further actions for orientation or preparation. The main reason appears to be that the prospective students in the present study had already undertaken many orientation activities. For example, they had already spoken with a study advisor (which is also recommended in the feedback on the SA), they attended an open day or orientation day of a specific study direction and consulted the information on the website. In addition, they indicated that, to them, their scores did not imply that further preparation was necessary and that they might have followed up on the feedback more if their scores had been lower.

Furthermore, the SA did not appear to have a big impact on study choice certainty. This finding must, again, be interpreted against the same background of a relatively well-prepared group of participants who felt already quite certain before completing the SA. None of the participants felt less certain or discouraged, but of course, their relatively high scores gave no reason for

this. In general, participants in the present study stated that the SA would have had a bigger impact with respect to following up on the feedback and/or study choice certainty if they had taken it earlier in their study orientation process. This explains why many of the participants indicated beforehand that they were mainly looking for affirmation. In that sense, the SA did meet their expectations. Overall, these results appear to be in line with other research. For instance, Soppe et al. (2020) have already shown that study choice certainty plays an overarching and important role in (the absence of) the effects of various study orientation activities. They also have demonstrated that the more certain prospective students are about their initial choice, the less impact an orientation activity has on their final choice and, thus, the less likely a change in choice certainty will take place. An interesting finding in their study was that some participants, who were 100% certain initially, nevertheless said that the orientation activity made them even more certain. So it seems that affirmation is an important consequence even for those who may not appear to need it.

4.1. Implications for the SA, Theory and Practice

4.1.1. *Implications for the SA*

For the current SA specifically, based on the present study, some refinements are proposed, before 'mainstream deployment'. First, recommendations are based on the evidence and threats in regard to the SA's *content*, despite the current study's focus on process and consequential aspect of validity. Results indicate that an addition of test items to the discipline and the social support test as well as an addition to the present set of subtests should be considered to reduce the threat of construct under-representation (Downing & Haladyna, 2004). Regarding additional items to existing subtests, further analyses should be carried out to secure the internal structure and predictive value of the tests. At the same time, when adding test items or subtests to the SA, parsimony should not be lost sight of, as the limited number and shortness of tests did motivate students to take all subtests, even those that did not seem relevant to them initially. In regard to adding new subtests, a broader range of knowledge and skills tests would be valuable (e.g., reasoning skills, study strategies) and a content sample test would be recommendable. After all, prospective students indicate they expect and desire some feedback regarding the fit with the subject of study they are considering to choose. A content sample subtest can offer them a hands-on experience prior to enrolment. Ideally, this would consist of for instance, studying course literature and/or watching video-lectures, followed by a short exam (Niessen et al., 2018).

Secondly, results in regard to the *process aspect of validity* showed that the numerical skills test seems to create a stressful state of mind regarding the SA that eases in the other tests with questions that merely require an answer realistically reflecting personal characteristics or circumstances rather than a correct answer. Since prospective students seem to fill out the SA from top to bottom, it is recommended either to change the linear presentation of the subtests or to change the order of the tests so that the numerical test is not the first test they encounter. In general, the SA should not frustrate or discourage students more than necessary. In that respect, we recommend to monitor test-anxiety and avoidant test-taking strategies in further evaluation as well.

A final refinement for the SA concerns the result button. To prevent prospective students from missing out on relevant feedback information, it is suggested to consider a push communication strategy (e.g., an automatic pop-up feedback window after taking a test) instead of the current pull strategy. In that way, no extra attention is required from users by which they are more likely to take the feedback in and perhaps act on it.

4.1.2. *Implications for theory*

More generally, this study adds to the literature by providing a distinctive and authentic example of collecting and interpreting process and consequential evidence with the aim to enhance assessment validity. Though validity literature provides a clear picture of the different sources of evidence and threats to validity, a flaw of many applied validation studies is that they tend to focus solely on content, internal structure and predictive aspects of validity (Cook et al., 2014). Moreover, regrettably these examples mainly involve so called high-stakes assessments (i.e., for selection, pass/fail, or grading decisions), standardized tests, predominantly in the context of health professions (Cook et al., 2014). As our results showed, a self-assessment can have an impact in prospective students' study decisions and progress. Access to higher education – even if (or especially when) it is open – requires the best possible decision making support. It is a call of duty to justify assessment procedures in this context, based on empirical arguments (Niessen & Meijer, 2017).

4.1.3. *Implications for practice*

The self-assessment is embedded in the existing practice of providing information and advice prior to enrolment. Combining orientation activities with expert advice has been shown to be relevant for the quality of study decisions and the study process (Borghans et al., 2015; Zhang et al., 2019). Hence, study advisors were closely involved in the development process of the SA and especially of the feedback provided aligned to the subtests, as this feedback refers to study advisors' services. Based on this feedback, prospective students, thus, might contact study advisors for further clarification or advice in following up the feedback. This assumes that study advisors are able to interpret the SA results with the necessary nuances. In that regard, recommended future steps involve additional training (e.g., a handout of how to interpret SA scores) and exchange of experiences, for quality assurance purposes.

The SA evokes reflection on study preparedness and offers concrete insights and suggestions regarding opportunities to improve chances of success, both prior to and after enrolment. The 'advice' category in the feedback links for example also to existing remedial tutorials and courses the educational institute provides to its students. Previous research has shown that such (early) remediation is a promising effective strategy for improving retention (Delnoij et al., 2020a; Muljana & Luo, 2019; Robbinson et al., 1996; Sage et al., 2018; Wachen et al., 2016).

4.2. Limitations and Implications for Future Research

Reflecting on the specific research method used for this study, an observer effect (i.e., the Hawthorne effect, see Sommer, 1968; or McCambridge et al., 2012 for a more recent review) might have played a role as the researcher was watching participants while taking the test. For instance, regarding the numerical skills test, some participants mentioned that they felt rushed or insecure, because of being observed. In general, however, there were only few indications of flawed answers. Some participants indicated the tendency to choose a specific answer option because that might lead to a higher score, but eventually selected their original answer. Still, the results have to be interpreted with some caution.

For future research, we recommend to expand the investigation of the consequential aspect of validity by evaluating the effects of the SA on enrolment and study success after enrolment (Downing, 2003). In that regard, the classification model (i.e. accuracy, false positives/negatives) set in an earlier stage of the design process (Delnoij et al., 2021) should be evaluated. In addition, the current sample involved a relatively large group already reasonably certain of their study decision while participating. In the present study, the sample consisted of prospective students who indicated their interest by, for instance, calling the student service office (see method section). It seems that students do so, in case they are already relatively certain of enrolling. Future research is needed to investigate the SA's impact on prospective

students who are less certain of their study decision (Cobern & Adams, 2020; Guba, 1981; Shenton, 2004). In that regard, we recommend utilizing an additional or different sampling method.

Nevertheless, relatively rapid and innocuous pilot tests like the present study are important in design-based research in general and for the SA in specific, to enable adjustments and refinements aligned to the intended effects prior to a 'full release'. In addition, small-scale qualitative studies provide in-depth insight into prospective students' response processes while taking the SA and the consequences of the SA on their study decision process, two aspects that are underreported in applied validation studies, yet tremendously important in determining assessment effectiveness.

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. Ethical approval was obtained by the Ethics Committee of the Open University (code: U20200563).

Authorship Contribution Statement

Laurie E. C. Delnoij: Investigation, Resources, Visualization, Formal Analysis, and Writing-original draft and revisions. **José P. W. Janssen**: Methodology, Formal Analysis, Supervision, Validation, and Writing-revision. **Kim J. H. Dirkx**: Methodology, Supervision, Validation, and Writing-revision. **Rob L. Martens**: Supervision, Validation, and Writing-revision.

Orcid

Laurie E. C. Delnoij https://orcid.org/0000-0001-6363-5714
José P. W. Janssen https://orcid.org/0000-0002-5104-7648
Kim J. H. Dirkx https://orcid.org/0000-0001-8014-0916
Rob L. Martens https://orcid.org/0000-0001-7193-8125

REFERENCES

Abbasi, N., & Ghosh, S. (2020). Construction and standardization of examination anxiety scale for adolescent students. *International Journal of Assessment Tools in Education*, *7*(4), 522-534. https://doi.org/10.21449/ijate.793084.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* American Psychological Association.

Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. *Psychologi cal Review, 84*(2), 191-215.

Beckman, T.J., Cook *et al.*, D.A., & Mandrekar, J.N. (2005). What is the validity evidence for assessments of clinical teaching? *Journal of General Internal Medicine*, *20*(12), 1159-1164. https://doi.org/10.1111/j.1525-1497.2005.0258.x

Borghans, L., Golsteyn, B., & Stenberg, A. (2015). Does expert advice improve educational choice? *PLoS One, 10*(12), Article e0145378. https://doi.org/10.1371/journal.pone.0145378

Broos, T., Verbert, K., Langie, G., Van Soom, C., & De Laet, T. (2018). Multi-institutional positioning test feedback dashboard for aspiring students: lessons learnt from a case study in Flanders. *Proceedings of the 8th International Conference on learning Analytics and Knowledge,* 51-55. https://doi.org/10.1145/3170358.3170419

Broos, T., Pinxten, M., Delporte, M., Verbert, K., & De Laet, T. (2019). Learning dashboards at scale: early warning and overall first year experience. *Assessment & Evaluation in Higher Education,* 45(6), 1–20. https://doi.org/10.1080/02602938.2019.1689546

Charters, E. (2003). The use of think-aloud methods in qualitative research an introduction to think-aloud methods. *Brock Education: A Journal of Educational Research and Practice*, *12*(2), 68-82. https://doi.org/10.26522/brocked.v12I2.38

Cobern, W.W., & Adams, B.A. (2020). When interviewing: how many is enough?. *International Journal of Assessment Tools in Education*, *7*(1), 73-79. https://dx.doi.org/10.21449/ijate.693217.

Cohen, A.D. (2006). The coming age of research on test-taking strategies. *Language Assessment Quarterly, 3*(4), 307-331. https://doi.org/10.1080/15434300701333129

Cook, D.A., Zendejas, B., Hamstra, S. J., Hatala, R., & Brydges, R. (2014). What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Advances in Health Sciences Education*, *19*(2), 233-250. https://doi.org/10.1007/s10459-013-9458-4

Creswell, J.W. (2014). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Edinburgh Gate: Pearson Education Limited.

Delnoij, L.E.C., Dirkx, K.J.H., Janssen, J.P.W., & Martens, R.L. (2020a). Predicting and resolving non-completion in higher (online) education – A literature review. *Educational Research Review,* 29, 100313. https://doi.org/10.1016/j.edurev.2020.100313

Delnoij, L.E.C., Janssen, J.P.W., Dirkx, K.J.H., Gijselaers, H.J.M., De Groot, R.H.M., Neroni, J., De Bie, M., & Martens, R.L. (2021). Predicting completion: The road to informed study decisions in higher online education. *Frontiers in Education, 6*. https://doi.org/10.3389/feduc.2021.668922

Delnoij, L.E.C., Janssen, J.P.W., Dirkx, K.J.H., Martens, R.L. (2020b) *Designing an online self-assessment for informed study decisions: The user perspective*. In C. Alario-Hoyos, M. J. Rodríguez-Triana, M. Scheffel, I. Arnedillo-Sánchez, & S.M. Dennerlein (Eds), Lecture Notes in Computer Science: Vol. 12315. Addressing Global Challenges and Quality Education. Springer.

Downing, S.M. (2003). Validity: on the meaningful interpretation of assessment data. *Medical Education*, *37*(9), 830-837. https://doi.org/10.1046/j.1365-2923.2003.01594.x

Downing, S.M., & Haladyna, T.M. (2004). Validity threats: overcoming interference with proposed interpretations of assessment data. *Medical Education*, *38*(3), 327-333. https://doi.org/10.1046/j.1365-2923.2004.01777.x

Dowker, A., Sarkar, A., & Looi, C.Y. (2016). Mathematics anxiety: What have we learned in 60 years?. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00508

Dunning, D., Heath, C., & Suls, J.M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest*, *5*(3), 69–106. https://doi.org/10.1111/j.1529-1006.2004.00018.x

Erlingsson, C., & Brysiewicz, P. (2017). A hands-on guide to doing content analysis. *African Journal of Emergency Medicine*, *7*(3), 93-99. https://doi.org/10.1016/j.afjem.2017.08.001

Fonteyne, L., & Duyck, W. (2015). Vraag het aan SIMON! [Ask SIMON!]. *Thema Hoger Onderwijs, 2*, 56-60.

Goodwin, L.D., & Leech, N.L. (2003). The meaning of validity in the new standards for educational and psychological testing: Implications for measurement courses. *Measurement & Evaluation in Counseling & Development, 36*(3), 181-191. https://doi.org/10.1080/07481756.2003.11909741

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

Hartnett, M. (2016). The importance of motivation in online learning. In Harnett, M. (Ed.) *Motivation in online education* (pp. 5-32). Springer. https://doi.org/10.1007/978-981-10-0700-2_2

Jivet, I., Scheffel, M., Schmitz, M., Robbers, S., Specht, M., & Drachsler, H. (2020). From students with love: An empirical study on learner goals, self-regulated learning and sense-making of learning analytics in higher education. *The Internet and Higher Education*, *47*. https://doi.org/10.1016/j.iheduc.2020.100758

Krefting, L. (1991). Rigor in qualitative research: The assessment of trustworthiness. *American Journal of Occupational Therapy*, *45*(3), 214-222.

Kubinger, K.D., Frebort, M., & Müller, C. (2012). Self-assessment im rahmen der studientberatung: Möglichkeiten und Grenzen. In: Kubinger, K.D., Frebort, M., Khorramdel, L., & Weitensfelder, L. (eds). *Self-Assessment: Theorie und Konzepte,* 9-24. [Self-Assessment: Theory and Concepts] Lengerich: Pabst Science Publishers.

Kutlu, O., & Yavuz, H. C. (2019). An Effective way to provide item validity: Examining student response processes. *International Journal of Assessment Tools in Education*, *6*(1), 9-24. https://doi.org/10.21449/ijate.447780

Liebert, R.M., & Morris, L.W. (1967). Cognitive and emotional components of test anxiety: a distinction and some initial data. *Psychological Reports, 20*(3)*,* 975-978. https://doi.org/10.2466/pr0.1967.20.3.975

McCambridge, J., Witton, J., & Elbourne, D.R. (2014). Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. *Journal of Clinical Epidemiology*, *67*(3), 267-277. https://doi.org/10.1016/j.jclinepi.2013.08.015

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5-11. https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2330-8516.1988.tb00303.x

Muljana, P.S., & Luo, T. (2019). Factors contributing to student retention in online learning and recommended strategies for improvement: A systematic literature review. *Journal of Information Technology Education: Research*, *18*. https://doi.org/10.28945/4182

Nolden, P., & Wosnitza, M. (2016). Webgestützte selbstreflexion von abbruchrisiken bei studierenden [Web-based self-reflection of drop-out risk among students]. *Empirische Pädagogik, 30*(3/4), 576-603.

Nolden, P., Wosnitza, M., Karabenick, S.A., Peixoto, F., Gonida, E., Ilic, I.S., Almeida, L.S., Stamovlasis, D., Do Céu Taveira, M., Toskoviç, O., Falanga, K., Aivazidis, C., Krstic, K., Videnovic, M., Gouveia, M., Castro Silva, J., Delzepich, R., Holder, L., & Clinton, E. (2019). Enhancing student self-reflection on the situation at university. The SRT scale inventory.

Nicol, D. (2009). Assessment for learner self-regulation: enhancing achievement in the first year using learning technologies. *Assessment & Evaluation in Higher Education*, *34*(3), 335-352. https://doi.org/10.1080/02602930802255139

Niessen, A.S.M., & Meijer, R.R. (2017). Voorspellen in het kader van de studiekeuzecheck: Tijd voor verbetering [Predicting in the context of study decisions: time for improvement]. *Onderzoek van Onderwijs, 46,* 5-7.

Niessen, A.S.M., Meijer, R.R., & Tendeiro, J.N. (2017). Measuring non-cognitive predictors in high-stakes contexts: The effect of self-presentation on self-report instruments used in admission to higher education. *Personality & Individual Differences 106*, 183-189. http://dx.doi.org/10.1016/j.paid.2016.11.014

Niessen, A.S.M., Meijer, R.R., & Tendeiro, J.N. (2018). Admission testing for higher education: A multi-cohort study on the validity of high-fidelity curriculum-sampling tests. *PloS One, 13*(6), e0198746. https://doi.org/10.1371/journal.pone/0198746

O'Regan, L., Brown, M., Harding, N., McDermott, G., & Ryan, S. (2016) *Technology-enabled feedback in the first year: A synthesis of the literature*. http://y1feedback.ie/wpcontent/uploads/2016/04/SynthesisoftheLiterature2016.pdf

Padilla, J.L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26*(1), 136-144.

Robinson, D.A. (1996). Orientation programs: A foundation for student learning and success. *New Directions for Student Services*, *75*, 55-68. https://doi.org/10.1002/ss.37119967507

Royal, K.D. (2017). Four tenets of modern validity theory for medical education assessment and evaluation. *Advances in Medical Education and Practice*, *8,* 567-570. https://doi.org/10.2147/AMEP.S139492

Sage, A.J., Cervato, C., Genschel, U., & Ogilvie, C.A. (2018). Combining academics and social engagement: a major-specific early alert method to counter student attrition in science, technology, engineering, and mathematics. *Journal of College Student Retention: Research, Theory & Practice, 22*(4), 611-626. https://doi.org/10.1177/1521025118780502

Shenton, A.K. (2004). Strategies for ensuring trustworthiness in qualitative research projects. *Education for Information*, *22*(2), 63-75.

Sommer, R. (1968). The Hawthorne dogma. *Psychological Bulletin*, *70*(6, Pt.1), 592. https://doi.org/10.1037/h0026728

Soppe, K.F.B., Wubbels, T., Leplaa, H.J., Klugkist, I., & Wijngaards-de Meij, L.D.N.V. (2019). Do they match? Prospective students' experiences with choosing university programmes. *European Journal of Higher Education*, *9*(4), 359-376. https://doi.org/10.1080/21568235.2019.1650088

St-Onge, C., Young, M., Eva, K.W., & Hodges, B. (2017). Validity: one word with a plurality of meanings. *Advances in Health Sciences Education, 22*, 853-867. https://doi.org/10.1007/s10459-016-9716-3

Tinto, V. (2005) Taking student success seriously: Rethinking the first year of college. IN: Taking student success seriously: Rethinking the first year of college. *Paper Presented at the Ninth Annual Intersession Academic Affairs Forum, California State University, Fullerton.*

Van den Akker, J., Bannan, B., Kelly, A.E., Nieveen, N., & Plomp, T (2013). *Educational design research: An introduction. Educational design research.* Enschede.

Van Den Haak, M., De Jong, M., & Jan Schellens, P. (2003). Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behaviour & Information Technology*, *22*(5), 339-351. https://doi.org/10.1080/0044929031000

Viswesvaran, C., & Ones, D.S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*, 197–210. http://dx.doi.org/10.1177/00131649921969802.

Vossensteyn H., Kottmann, A., Jongbloed, B.W.A., Kaiser, F., Cremonini, L, Stensaker, B., Hovdhaugen, E., & Wollscheid, S. (2015). Dropout and completion in higher education in Europe: Main report. https://research.utwente.nl/en/publications/dropout-and-completion-in-higher-education-in-europe-main-report.

Wachen, J., Pretlow, J., & Dixon, K.G. (2016). Building college readiness: exploring the effectiveness of the unc academic summer bridge program. *Journal of College Student Retention: Research, Theory & Practice, 20*(1), 116-138. https://doi.org/10.1177/1521025116649739

Zhang, X., Gossett, C., Simpson, J., & Davis, R. (2019). Advising students for success in higher education: An all-out effort. *Journal of College Student Retention: Research, Theory & Practice*, *21*(1), 53-77. https://doi.org/10.1177/1521025116689097

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 2, 397–409
https://doi.org/10.21449/ijate.1101295
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Comparison of Normality Tests in Terms of Sample Sizes under Different Skewness and Kurtosis Coefficients

Suleyman Demir[1,*]

[1]Sakarya University, Faculty of Education, Department of Educational Sciences, Sakarya, Türkiye.

ARTICLE HISTORY

Received: Apr. 06, 2021
Revised: Feb. 14, 2022
Accepted: Apr. 05, 2022

Keywords:
Normal distribution,
Skewness,
Kurtosis,
Normality tests.

Abstract: This study aims to compare normality tests in different sample sizes in data with normal distribution under different kurtosis and skewness coefficients obtained simulatively. To this end, firstly, simulative data were produced using the MATLAB program for different skewness/kurtosis coefficients and different sample sizes. The normality analysis of each data type was conducted using the MATLAB program and ten different normality tests; namely, (Kolmogorov Smirnov (KS) Test, KS Stephens Modification, KS Marsaglia, KS Lilliefors Modification, Anderson-Darling Test, Cramer- Von Mises Test, Shapiro-Wilk Test, Shapiro-Francia Test, Jarque-Bera Test, and D'Agostino & Pearson Test). As a result of the analyses conducted according to ten different normality tests, it was found that normality tests were not affected by the sample size when the skewness and kurtosis coefficients were equal to or close to zero; however, in cases where the skewness and kurtosis coefficients moved away from zero, it was found that normality tests are affected by the sample size, and such tests tend to give significant results. Therefore, in large samples, it may be suggested that critical values for skewness and kurtosis coefficients' z-scores as proposed by Kim (2013) and Mayers (2013) or the histogram graphs be used.

1. INTRODUCTION

Parametric methods in data analysis (t-tests, ANOVA, etc.) are used in cases where the data obtained from the sample have a normal distribution. If the data do not have a normal distribution, non-parametric methods (Mann Whitney U, Kruskal Wallis, Wilcoxon, etc.) are used to analyze data. Parametric methods, based on a specific distribution such as normal distribution, can be used in conditions where the normality assumption is provided. Non-parametric methods are implemented independently of the distribution, converting data to ordinal data type (Field, 2013).

The definition of normal distribution was first introduced by Abraham de Moivre in 1667 and defined by a mathematical formula (Howell, 2013; Martin & Bridgmon, 2012). Also called Gaussian distribution as given in Formula 1.1. If it is accepted as ($\mu = 0$) and ($\sigma = 1$) in the formula in 1.1, the standard normal distribution function is expressed as (1.2).

*CONTACT: Suleyman DEMIR ✉ suleymand@sakarya.edu.tr Sakarya University, Faculty of Education, Department of Educational Sciences, Sakarya, Türkiye

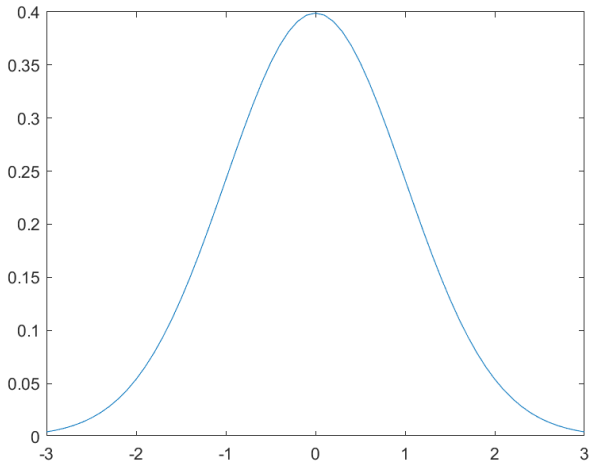
e-ISSN: 2148-7456 /© **IJATE** 2022

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad (1.1)$$

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}x^2} \quad (1.2)$$

The standard normal distribution chart given in Figure 1 has standard deviation (σ) values on the X-axis, while probability values are on the Y-axis. For a sample with a standard normal distribution; 68.2% of the sample fall within the ±1 standard deviation range, 95.4% within ±2 standard deviations, 99.7% within ±3 standard deviations, and 0.3% outside of the ±3 standard deviation range (Field, 2013; Howell, 2013; Martin & Bridgmon, 2012; Tabachnick & Fidell, 2013).

Figure 1. *Standard normal distribution chart*



The normal distribution has two components, namely, skewness and kurtosis. Skewness is related to the status of the data's mode median and mean relative to each other. There is symmetric distribution when the mean is in the middle of the distribution; thus, there is no skewness. When the mean is not in the middle of the distribution, there is a non-symmetric distribution (skewed distribution) (Tabachnick & Fidell, 2007). The kurtosis is related to how far the data move away from the mean or how close they get to the mean. In other words, it is related to the standard deviation of data. When the standard deviation is small, there is a pointed distribution (leptokurtic, short-tailed); whereas, when the standard deviation is large, there is a flattened distribution (platykurtic, long-tailed) (Baykul & Güzeller, 2013; Field, 2013; Tabachnick & Fidell, 2013).

Although there are different methods for the calculation of the coefficient of skewness, Baykul and Güzeller (2013) state that using the mean as the central tendency measure of the coefficient of skewness and the standard deviation as the measure of the central distribution, thereby calculating it according to the third moment around the mean gives better results. Accordingly, the formula calculates the skewness coefficient in (1.3) according to the third moment around the mean. When the skewness coefficient is equal to zero, as in a normal distribution, there is symmetric distribution since the majority of the data are around the average. The skewness coefficient being negative indicates that most of the data are located on the right side of the mean and the tail on the left side is longer, while the skewness coefficient being positive shows that the majority of the data are located on the left side of the average and the tail on the right side is longer (Tabachnick & Fidell, 2013).

$$skew = \frac{(n-1)(n-2)}{n} \sum_{j=1}^{n} \frac{\left(X_j - \overline{X}\right)^3}{\sigma_x^3} \tag{1.3}$$

The formula calculates the kurtosis coefficient in (1.4) according to the fourth moment around the mean (formula in 1.5 is mostly preferred). A positive kurtosis coefficient indicates that the distribution is more pointed than the normal distribution. In contrast, a negative kurtosis coefficient indicates that the distribution is more flattened than the normal distribution. That the kurtosis coefficient equals zero indicates the distribution neither too flattened nor too pointed, as in the normal distribution (Field, 2013; Howell, 2013; Martin & Bridgmon, 2012; Tabachnick & Fidell, 2013).

$$kurt = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{j=1}^{n} \frac{\left(X_j - \overline{X}\right)^4}{\sigma_x^4} \tag{1.4}$$

$$kurt = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{j=1}^{n} \frac{\left(X_j - \overline{X}\right)^4}{\sigma_x^4} - 3 \tag{1.5}$$

The fact that both the skewness and kurtosis coefficients using the formulas in (1.3) and (1.5) are zero shows that the data have normal distribution; however, since the skewness and kurtosis values are mostly different from zero, acceptable ranges are determined for these values. These ranges have been suggested to be ±2 in some sources (Field, 2013; Gravetter & Wallnau, 2014; George & Mallery, 2010; Trochim & Donnelly, 2006); ±1.5 in some (Tabachnick & Fidell, 2013); and ±1 in other sources (Bulmer, 1979). In addition, Hair et al. (2010) and Bryne (2010) state that the normality assumption is not fulfilled when the skewness coefficient is outside the range of ±2 and the kurtosis coefficient is outside the range of ±7; while according to Kline (2011), these ranges are ±3 for the skewness coefficient and ±10 for the kurtosis coefficient.

In addition to skewness and kurtosis coefficients, standard z-scores of skewness and kurtosis coefficients, histogram graphs, and normality tests are also used to test the normality assumption. The standard z-score of the skewness coefficient is calculated as follows (1.6): the skewness coefficient is divided by the standard error value of the skewness coefficient (1.7). The standard z-score of the kurtosis coefficient is calculated as follows (1.6): the kurtosis coefficient is divided by the standard error value of the kurtosis coefficient (1.7). Field (2013) and Tabachnick and Fidell (2013) recognize that these values in the range of ±1.96 show a normal distribution in small samples. The standard z-scores fall outside the range of ±1.96 because the standard error values of the skewness and kurtosis coefficients decrease as the sample size increases. In such cases, the histogram graph should be interpreted instead of the standard z-scores (Field, 2013; Tabachnick & Fidell, 2013).

$$z_{skew} = \frac{skew - 0}{SE_{skew}} \tag{1.6}$$

$$SE_{skew} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}} \tag{1.7}$$

$$z_{kurt} = \frac{kurt - 0}{SE_{kurt}} \tag{1.8}$$

$$SE_{kurt} = \sqrt{\frac{24n(n-1)^2}{(n-3)(\mathrm{n}-2)(\mathrm{n}+3)(n+5)}} \tag{1.9}$$

It was stated by Kim (2013) and Mayers (2013) that, for all sample sizes, the standard z-scores of the skewness and kurtosis coefficients which are within the range of ±1.96 are not sufficient for normal distribution. According to Kim (2013), the fact that when the sample number is n<50 and z-scores are in the range of ±1.96; and when the sample size is in the range of 50<n<300, and z-scores are in the range of ±3.29 means that the data have a normal distribution. When the sample size is n>300, the histogram graph should be interpreted. According to Mayers (2013), unlike Kim (2013), it is accepted that in cases where the sample size is in the range of 50<n<175, z-scores should be in the range of ±2.58.

Another method used to test the normality assumption is normality tests. The fact that normality tests are significant indicates that the data differ significantly from the normal distribution. In contrast, the fact that normality tests are not significant indicates that the data do not differ significantly from the normal distribution. In research studies, mostly Kolmogorov Smirnov (KS) Test (Smirnov, 1948) and Shapiro-Wilk Test (Shapiro & Wilk, 1965) are used. However, tests such as KS Stephens Modification, KS Marsaglia, KS Lilliefors Modification, Anderson-Darling Test, Cramer- Von Mises Test, Shapiro-Francia Test, Jarque-Bera Test, D'Agostino-Pearson Test (Detailed information is given in the Appendix) are also used (Lee et al., 2016; Marsaglia et al., 2003; Stephens, 1974).

Sample size affects the results of normality tests. When sample is small, normality tests tend to accept the null hypothesis. In large samples, even small deviations from the normal distribution cause the normality test to reject the null hypothesis (Öztuna et al., 2006). Some researchers (Lumley et al., 2002; Wilcox, 2010) state that in large samples, according to the central limit theorem, the data will approach normal distribution, therefore it can be assumed that the normal distribution assumption is achieved in large samples regardless of the normality determination methods However, some researchers (Micceri, 1989; Öztuna et al., 2006) state that this is not true: since the number of samples is large, the data will not always have a normal distribution. It seems more important to interpret the histogram graph with the skewness and kurtosis coefficients of the data.

Because of the confusion about the sample size, it is stated in the related literature that if the number of samples exceeds 200, the data should be considered to have a normal distribution due to the central limit theorem or that only the histogram graph should be interpreted. In the studies carried out, the power of the test was calculated for different sample sizes to determine the sensitivity of normality tests in data with normal and non-normal distribution (Douglas & Edith, 2002; Frain, 2007; Keskin, 2006; Nornadiah & Yap 2011; Nor-Aishah & Shamsul 2007; Öztuna et al., 2006; Rinnakorn & Kamon 2007; Stephens, 1974; Ukponmwan & Ajibade, 2017; Yap & Sim 2011). In cases where the sample size exceeds 200, it can be accepted that the data have a normal distribution or the histogram graph can be interpreted. However, the histogram graph will visually move away from the normal distribution as the kurtosis and skewness coefficients move away from zero. It is important to determine how much the histogram graph differs from the normal distribution according to the sample size, kurtosis coefficient, and skewness coefficient. However, when the relevant literature is reviewed, no study can be found to determine the sensitivity of normality tests under different sample sizes in the data that are considered to show the normal distribution in terms of different skewness and kurtosis coefficients. This study, therefore, aims to compare normality tests in terms of different sample

sizes in data with normal distribution in terms of different kurtosis and skewness coefficients obtained simulatively.

2. METHOD

2.1. Obtaining Data

In order to realize the purpose of the study, firstly conditions were created depending on the kurtosis and skewness coefficients and the sample size. If the sample is less than 30, it is expressed as a small sample, and if the sample size is larger than 400, it is expressed as a large sample (Abbott, 2011, Demir et al., 2016; Orcan, 2020). In the present study, 11 different sample sizes were determined to cover small and large samples (10, 20, 30, 40, 50, 100, 200, 300, 400, 500, 900). That the coefficient of kurtosis and skewness are zero means normal distribution. According to all the references in the related literature, the acceptable range for the skewness and kurtosis coefficients is ±1. In the current study, five different skewness and kurtosis values were determined, (-0.50, -0.25, 0, 0.25, 0.50) with a specific purpose to compare normality tests on data showing normal distribution (Bryne, 2010; Bulmer, 1979; Field, 2013; Gravetter & Wallnau, 2014; George & Mallery, 2010; Hair et al., 2010; Kline, 2011; Tabachnick & Fidell, 2013; Trochim & Donnelly, 2006). The sample sizes and skewness/kurtosis coefficients used in this study are shown in Table 1.

Table 1. *Sample sizes and skewness/kurtosis coefficients used for data generation*

	Sample Sizes	Skewness/Kurtosis Coeff.
Condition	10, 20, 30, 40, 50, 100, 200, 300, 400, 500, 900	-0.50, -0.25, 0, 0.25, 0.50

The current study examined normality tests under 55 different conditions, including 11 different sample sizes and five different skewness/kurtosis coefficients.

2.2. Analysis of the Data

The data obtained according to five different skewness and kurtosis coefficients and 11 different sample sizes were analyzed using the MATLAB program according to 10 different normality tests (Kolmogorov Smirnov, KS Stephens Modification, KS Marsaglia, KS Lilliefors Modification, Anderson-Darling Test, Cramer- Von Mises Test, Shapiro-Wilk Test, Shapiro-Francia Test, Jarque-Bera Test, and D'Agostino & Pearson Test). MATLAB codes created by Öner and Kocakoç (2017) were used in the analysis phase.

3. FINDINGS

This section gives ten different normality test results of the data obtained simulatively under 55 different conditions. In Table 2, significance values obtained from normality tests are given for 11 different sample sizes under the condition that the kurtosis and skewness coefficients are -0.50. Table 2 shows that all normality tests do not have a normal distribution of data under conditions where the sample size is 200 or more for the significance level of α=0.05. Methods of Anderson-Darling Test and Cramer-Von Mises Test conclude that data do not have a normal distribution under conditions where the sample size is 40 or more for the significance level of α=0.05. Kolmogorov Smirnov, KS Marsaglia, Jarque-Bera Test, and D'Agostino & Pearson Test also conclude that data do not have a normal distribution under the sample size of 200 or more for the significance level of α=0.05. As a result, it can be said that the Anderson-Darling Test and Cramer-Von Mises Test methods are the most affected ones by the sample size when the coefficients of kurtosis and skewness are -0.50. At the same time, Kolmogorov Smirnov, KS Marsaglia, Jarque-Bera Test, and D'Agostino & Pearson Test are relatively less affected. In Table 3, the significance values obtained from the normality tests are given for 11 different sample sizes under conditions where the kurtosis and skewness coefficients are -0.25.

Table 2. *Normality test results for different samples in cases where kurtosis and skewness coefficients are (-0.50, -0.50)*

Sample Size	10	20	30	40	50	100	200	300	400	500	900
Normality Tests	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*
Kolmogorov Smirnov	0.890	0.942	0.602	0.493	0.302	0.061	0.004	0.001	0.000	0.000	0.000
KS Stephens Modification	0.150	0.150	0.139	0.081	0.027	0.010	0.010	0.010	0.010	0.010	0.010
KS Marsaglia Method	0.832	0.911	0.556	0.455	0.276	0.056	0.004	0.001	0.000	0.000	0.000
KS Lilliefors Modification	0.200	0.200	0.140	0.079	0.019	0.000	0.000	0.000	0.000	0.000	0.000
Anderson-Darling Test	0.778	0.644	0.367	0.049	0.025	0.001	0.000	0.000	0.000	0.000	0.000
Cramer- Von Mises Test	0.751	0.735	0.327	0.045	0.020	0.000	0.000	0.000	0.000	0.000	0.000
Shapiro-Wilk Test	0.844	0.519	0.338	0.144	0.070	0.001	0.000	0.000	0.000	0.000	0.000
Shapiro-Francia Test	0.785	0.567	0.391	0.168	0.089	0.003	0.000	0.000	0.000	0.000	0.000
Jarque-Bera Test	0.771	0.594	0.458	0.353	0.272	0.074	0.005	0.000	0.000	0.000	0.000
D'Agostino & Pearson Test	0.652	0.545	0.427	0.329	0.253	0.065	0.004	0.000	0.000	0.000	0.000

Based on an analysis of Table 3, it can be concluded that all normality tests do not have a normal distribution of data under conditions where the sample size is 900 for the significance level of α=0.05. The KS Stephens Modification and KS Lilliefors Modification methods prove that data do not have a normal distribution under conditions where the sample size is 40 or more for the significance level of α=0.05. The Kolmogorov Smirnov and KS Marsaglia methods conclude that data do not have a normal distribution under conditions where the sample size is 900 for the significance level of α=0.05. In conclusion, it can be said that the KS Stephens Modification and KS Lilliefors Modification methods are affected most by the sample size when the coefficients of kurtosis and skewness are -0.25. At the same time, Kolmogorov Smirnov and KS Marsaglia are relatively less affected. In Table 4, the significance values obtained from the normality tests are given for 11 different sample sizes under conditions where the kurtosis and skewness coefficients are 0.00.

Table 3. *Normality test results for different samples in cases where kurtosis and skewness coefficients are (-0.25, -0.25)*

Sample Size	10	20	30	40	50	100	200	300	400	500	900
Normality Tests	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*
Kolmogorov Smirnov	0.892	0.998	0.970	0.707	0.659	0.719	0.388	0.392	0.144	0.081	0.009
KS Stephens Modification	0.150	0.150	0.150	0.150	0.150	0.150	0.047	0.048	0.010	0.010	0.010
KS Marsaglia Method	0.835	0.993	0.953	0.666	0.622	0.693	0.372	0.379	0.139	0.078	0.008
KS Lilliefors Modification	0.200	0.200	0.200	0.200	0.200	0.200	0.046	0.049	0.003	0.001	0.000
Anderson-Darling Test	0.612	0.887	0.846	0.512	0.504	0.270	0.053	0.034	0.012	0.004	0.000
Cramer- Von Mises Test	0.475	0.878	0.774	0.362	0.358	0.217	0.072	0.046	0.012	0.002	0.000
Shapiro-Wilk Test	0.911	0.923	0.903	0.789	0.771	0.497	0.196	0.101	0.039	0.013	0.000
Shapiro-Francia Test	0.722	0.869	0.866	0.710	0.699	0.438	0.203	0.127	0.055	0.020	0.000
Jarque-Bera Test	0.937	0.878	0.823	0.771	0.722	0.522	0.272	0.142	0.074	0.039	0.003
D'Agostino & Pearson Test	0.763	0.822	0.802	0.767	0.728	0.538	0.280	0.144	0.074	0.037	0.003

Table 4 shows that all normality tests have a normal distribution of data for the significance level of α=0.05, regardless of the number of samples for the significance level of α=0.05. In conditions where the kurtosis and skewness coefficients are 0.00, it can be said that no normality test is affected by the sample size. In Table 5, the significance values obtained from the normality tests are given for 11 different sample sizes under conditions where the kurtosis and skewness coefficients are 0.25.

Table 4. *Normality test results for different samples in cases where kurtosis and skewness coefficients are (0.00, 0.00)*

Sample Size	10	20	30	40	50	100	200	300	400	500	900
Normality Tests	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*
Kolmogorov Smirnov	0.846	0.968	0.953	0.838	0.858	0.922	0.770	0.733	0.716	0.777	0.527
KS Stephens Modification	0.150	0.150	0.150	0.150	0.150	0.150	0.150	0.150	0.150	0.150	0.109
KS Marsaglia Method	0.780	0.947	0.930	0.803	0.827	0.905	0.752	0.717	0.702	0.766	0.519
KS Lilliefors Modification	0.200	0.200	0.200	0.200	0.200	0.200	0.200	0.200	0.200	0.200	0.120
Anderson-Darling Test	0.384	0.564	0.687	0.654	0.764	0.730	0.582	0.757	0.674	0.612	0.246
Cramer- Von Mises Test	0.246	0.494	0.599	0.578	0.714	0.730	0.594	0.736	0.603	0.521	0.212
Shapiro-Wilk Test	0.781	0.844	0.898	0.899	0.930	0.932	0.842	0.873	0.876	0.850	0.606
Shapiro-Francia Test	0.475	0.666	0.772	0.758	0.811	0.799	0.737	0.825	0.841	0.797	0.558
Jarque-Bera Test	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
D'Agostino & Pearson Test	0.685	0.832	0.876	0.900	0.915	0.950	0.972	0.980	0.985	0.987	0.993

Based on data in Table 5, the Jarque-Bera Test and D'Agostino & Pearson test methods conclude that data do not have a normal distribution under conditions where the sample size is 500 or more for the significance level of α=0.05. Kolmogorov Smirnov, KS Stephens Modification, KS Marsaglia, KS Lilliefors Modification, Anderson-Darling Test and Cramer-Von Mises Test show that regardless of the number of samples, data have a normal distribution for the significance level of α=0.05. In conclusion, it can be said that the Jarque-Bera Test and D'Agostino & Pearson Test methods are affected most by the sample size when the coefficients of kurtosis and skewness are 0.25. At the same time, the Kolmogorov Smirnov, KS Stephens Modification, KS Marsaglia, KS Lilliefors Modification, Anderson-Darling Test, and Cramer-Von Mises methods are not affected by the sample size. In Table 6, the significance values obtained from the normality tests are given for 11 different sample sizes under conditions where the kurtosis and skewness coefficients are 0.50.

Table 5. *Normality test results for different samples in cases where kurtosis and skewness coefficients are (0.25, 0.25)*

Sample Size	10	20	30	40	50	100	200	300	400	500	900
Normality Tests	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*
Kolmogorov Smirnov	0.732	0.792	0.803	0.936	0.965	0.932	0.936	0.957	0.957	0.931	0.862
KS Stephens Modification	0.150	0.150	0.150	0.150	0.150	0.150	0.150	0.150	0.150	0.150	0.150
KS Marsaglia Method	0.657	0.739	0.760	0.913	0.950	0.917	0.925	0.950	0.951	0.924	0.855
KS Lilliefors Modification	0.198	0.200	0.200	0.200	0.200	0.200	0.200	0.200	0.200	0.200	0.200
Anderson-Darling Test	0.208	0.246	0.410	0.545	0.684	0.581	0.464	0.576	0.506	0.484	0.136
Cramer- Von Mises Test	0.125	0.180	0.344	0.574	0.767	0.738	0.612	0.693	0.573	0.576	0.353
Shapiro-Wilk Test	0.541	0.547	0.672	0.732	0.767	0.666	0.332	0.230	0.143	0.081	0.005
Shapiro-Francia Test	0.262	0.342	0.477	0.528	0.569	0.455	0.242	0.184	0.119	0.066	0.006
Jarque-Bera Test	0.937	0.878	0.823	0.771	0.722	0.522	0.272	0.142	0.074	0.039	0.003
D'Agostino & Pearson Test	0.482	0.577	0.574	0.554	0.529	0.401	0.221	0.121	0.066	0.036	0.003

In Table 6, the Shapiro-Wilk Test, Shapiro-Francia Test, Jarque-Bera Test, and D'Agostino & Pearson Test methods prove that data do not have a normal distribution under the sample size of 200 or more for the significance level of α=0.05. The Kolmogorov Smirnov and KS Marsaglia methods, on the other hand, give the result that the data have a normal distribution for the significance level of α=0.05, regardless of the sample size. As a result, it can be said that the Shapiro-Wilk Test, Shapiro-Francia Test, Jarque-Bera Test, and D'Agostino & Pearson Test

methods are affected most by the sample size when the coefficients of kurtosis and skewness are 0.50. At the same time, the Kolmogorov Smirnov and KS Marsaglia methods are not affected by the sample size.

Table 6. *Normality test results for different samples in cases where kurtosis and skewness coefficients are (0.50, 0.50)*

Sample Size	10	20	30	40	50	100	200	300	400	500	900
Normality Tests	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*
Kolmogorov Smirnov	0.672	0.562	0.644	0.910	0.973	0.893	0.806	0.665	0.369	0.247	0.105
KS Stephens Modification	0.142	0.099	0.150	0.150	0.150	0.150	0.150	0.150	0.044	0.014	0.010
KS Marsaglia Method	0.596	0.507	0.597	0.882	0.961	0.873	0.789	0.649	0.358	0.239	0.102
KS Lilliefors Modification	0.148	0.104	0.171	0.200	0.200	0.200	0.200	0.200	0.042	0.014	0.001
Anderson-Darling Test	0.125	0.059	0.211	0.379	0.485	0.260	0.055	0.036	0.014	0.006	0.000
Cramer- Von Mises Test	0.074	0.062	0.191	0.474	0.663	0.442	0.187	0.098	0.035	0.014	0.001
Shapiro-Wilk Test	0.349	0.261	0.352	0.380	0.361	0.139	0.011	0.002	0.000	0.000	0.000
Shapiro-Francia Test	0.152	0.147	0.226	0.251	0.243	0.093	0.011	0.002	0.000	0.000	0.000
Jarque-Bera Test	0.771	0.594	0.458	0.353	0.272	0.074	0.005	0.000	0.000	0.000	0.000
D'Agostino & Pearson Test	0.271	0.288	0.243	0.199	0.161	0.054	0.006	0.001	0.000	0.000	0.000

4. CONCLUSION and DISCUSSION

This study aims to compare normality tests in terms of different sample sizes in data with normal distribution in terms of different kurtosis and skewness coefficients obtained simulatively. For this purpose, 55 data sets with different skewness and kurtosis coefficients and different sample sizes were produced simulatively. The analysis results according to 10 different normality tests showed that that normality tests are not affected by the sample size when the skewness and kurtosis coefficients are equal to or close to zero. However, in cases where the skewness and kurtosis coefficients moved away from zero, it was found that normality tests are affected by the sample size, and normality tests tend to give significant results, especially for n>200. It can also be said that the Kolmogorov Smirnov and KS Marsaglia tests are relatively less affected by sample size than other normality tests under all conditions. In other words, the Kolmogorov Smirnov and KS Marsaglia methods tend to accept the H_0 hypothesis. In studies conducted with data that did not show a normal distribution, it was seen that larger samples were needed for the Kolmogorov Smirnov method to reject the H_0 hypothesis; however, smaller samples were sufficient for the Anderson-Darling and Shapiro-Wilk methods (Ahad et al., 2011; Kundu et al., 2011). It can be said that the Kolmogorov Smirnov and KS Marsaglia methods tend to accept the H_0 hypothesis in normal and non-normal data. In cases where skewness and kurtosis coefficients are close to zero, researchers may be advised to use the Kolmogorov Smirnov and KS Marsaglia methods for small samples. However, instead of normality tests in large samples, histogram graphs or critical values as suggested by Kim (2013) and Mayer (2013) for z-scores may be used.

Declaration of Conflicting Interests and Ethics

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

Orcid

Suleyman Demir https://orcid.org/0000-0003-3136-0423

REFERENCES

Abbott, M.L. (2011). *Understanding educational statistics using Microsoft Excel and SPSS*. Wiley & Sons, Inc.

Ahad, N.A., Yin, T.S., Othman, A.R., & Yaacob, C.R. (2011). Sensitivity of normality tests to non-normal data. *Sains Malaysiana, 40*(6), 637-641. https://core.ac.uk/download/pdf/11491563.pdf

Anderson, T.W., & Darling, D.A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics, 23*(2), 193-212. https://doi.org/10.1214/aoms/1177729437

Anderson, T.W., & Darling, D.A. (1954). A test of goodness of fit. *Journal of the American Statistical Association, 49*(268), 765-769. https://doi.org/10.1080/01621459.1954.10501232

Baykul, Y., & Güzeller, C.O. (2013). *Sosyal bilimler için istatistik: SPSS uygulamalı [Statistics for social sciences: SPSS applied]. Pegem Akademi.*

Bulmer, M.G. (1979). *Principles of Statistics.* Dover.

Byrne, B.M. (2010). *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming.* Taylor and Francis Group Publication.

Csörgö, S., & Faraway, J.J. (1996). The exact and asymptotic distributions of Cramer-von Mises statistics. *Journal of Royal Statistical Society. Series B (Methodological), 58*(1), 221-234.

D'Agostino, R.B., & Pearson, E.S. (1973). Tests for departures from normality. Empirical results for the distribution of b2 and √b1. *Biometrika, 60*(3), 613-622. https://doi.org/10.1093/biomet/60.3.613

Dellal, G.E., & Wilkinson, L. (1986). An analytic approximation to the distribution of Lilliefors's test statistic for normality. *The American Statistician, 40*(4), 294-296. https://doi.org/10.1080/00031305.1986.10475419

Demir, E., Saatcioğlu, Ö., & İmrol, F. (2016). Uluslararası dergilerde yayımlanan eğitim araştırmalarının normallik varsayımları açısından incelenmesi [Examination of educational researches published in international journals in terms of normality assumptions]. *Current Research in Education, 2*(3), 130-148.

Douglas G.B., & Edith, S. (2002). A test of normality with high uniform power. *Journal of Computational Statistics and Data Analysis 40*(3), 435-445. https://doi.org/10.1016/S0167-9473(02)00074-9

Facchinetti, S. (2009). A procedure to find exact critical values of Kolmogorov-Smirnov test. *Statistica Applicata – Italian Journal of Applied Statistics, 21*(3-4), 337-359.

Field, A. (2013). *Discovering statistics using SPSS.* Sage Publications.

Frain, J.C. (2007). Small sample power of tests of normality when the alternative is an α-stable distribution. *Trinity Economics Papers TEP-0207,* Trinity College Dublin, Department of Economics. http://www.tcd.ie/Economics/TEP/2007/TEP0207.pdf

George, D., & Mallery, M. (2010). *SPSS for Windows Step by Step: A Simple Guide and Reference, 17.0*. Pearson.

Gravetter, F., & Wallnau, L. (2014). *Essentials of statistics for the behavioral sciences.* Wadsworth.

Hair, J.F., Black, W.C., Babin, B.J., & Anderson, R.E. (2010). *Multivariate data analysis: A global perspective.* Prentice Hall.

Harter, H.L. (1961). Expected values of normal order statistics, *Biometrika, 48,* 151-65.

Howell, D.C. (2013). *Statistical methods for psychology*. Belmont, Wadsworth/Cengage Learning.

Jarque, C.M., & Bera, A.K. (1987). A test for normality of observations and regression residuals. *International Statistical Review, 55*(2), 163-172. https://doi.org/10.2307/1403192

Keskin, S. (2006). Comparison of several univariate normality tests regarding type I error rate and power of the test in simulation based small samples. *Journal of Applied Science Research 2*(5), 296-300.

Kim, H.Y. (2013). Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restor Dent Endod, 38*(1), 52-4. https://doi.org/10.5395/rde.2013.38.1.52

Kline, R.B. (2011). *Methodology in the Social Sciences: Principles and practice of structural equation modeling.* Guilford Press.

Kundu, M.G., Mishra, S., & Khare, D. (2011). Specificity and Sensitivity of Normality Tests. *In Proceedings of VI International Symposium on Optimisation and Statistics.* Anamaya Publisher.

Lee, C., Park, S., & Jeong, J. (2016). Comprehensive Comparison of Normality Tests: Empirical Study Using Many Different Types of Data. *Journal of the Korean Data and Information Science Society, 27*(5), 1399-1412. https://doi.org/10.7465/jkdi.2016.27.5.1399

Lilliefors, H.W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association, 62*(318), 399-402. https://doi.org/10.1080/01621459.1967.10482916

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health, 23,* 151–169. https://doi.org/10.1146/annurev.publhealth.23.100901.140546

Marsaglia, G., Tsang, W.W., & Wang, J. (2003). Evaluating Kolmogorov's distribution. *Journal of Statistical Software, 8*(18). https://doi.org/10.18637/jss.v008.i18

Martin, W., & Bridgmon, K. (2012). *Quantitative and statistical research methods: from hypothesis to results.* Jossey-Bass.

Mayers, A. (2013). *Introduction to statistics and SPSS in psychology.* Pearson Education Limited.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156–166. https://doi.org/10.1037/0033-2909.105.1.156

Nor-Aishah H., & Shamsul R. A (2007, 12-14 December). *Robust Jacque-Bera Test of Normality.* The 9th Islamic Countries Conference on Statistical Sciences, University Malaya, Malaysia.

Nornadiah, M.R., & Yap, B.W. (2011). Power comparison of Shapiro-Wilk, Kolmogorov-Smirnov, Lillieforsand Anderson-Darling tests. *Journal of Statistical Modeling and Analytics, 2*(1), 21-33.

Orcan, F. (2020). Parametric or non-parametric: skewness to test normality for mean comparison. *International Journal of Assessment Tools in Education, 7*(2), 255-265. https://doi.org/10.21449/ijate.656077

Öztuna, D., Elhan, A.H., & Tuccar, E. (2006). Investigation of four different normality tests in terms of Type I error rate and power under different distributions, *Turk. J. Med. Sci. 36*(3), 171–176.

Öner, M., & Kocakoç, İ.D. (2017). JMASM 49: A compilation of some popular goodness of fit tests for normal distribution: their algorithms and MATLAB codes (MATLAB). *Journal of Modern Applied Statistical Methods, 16*(2), 547-575. https://doi.org/10.22237/jmasm/1509496200

Rinnakorn, C., & Kamon, B. (2007). A power comparison of goodness-of-fit tests for normality based on the likelihood ratio and the non-likelihood ratio. *Thailand Statistician, 5,* 57-68.

Shapiro, S.S., & Francia, R.S. (1972). An approximate analysis of variance test for normality. *Journal of the American Statistical Association, 67*(337), 215-216. https://doi.org/10.1080/01621459.1972.10481232

Stephens, M.A. (1986). Tests based on EDF statistics. In R. B. D'Agostino & M. A. Stephens (Eds.), *Goodness-of-fit techniques* (pp. 97-194). Marcel Dekker.

Stephens, M.A. (1974), EDF Statistics for Goodness of Fit and Some Comparisons. *Journal of the American Statistical Association, 69*, 730-737. https://doi.org/10.2307/2286009

Shapiro, S.S., & Wilk, M.B. (1965). An analysis of variance test for normality. *Biometrika, 52,* 591-611. https://doi.org/10.2307/2333709

Smirnov, N.V. (1948). Table for estimating the goodness of t of empirical distributions. *The Annals of Mathematical Statistics, 19,* 279-281.

Tabachnick, B.G., & Fidell, L.S. (2013). *Using Multivariate Statistics.* Pearson.

Trochim, W.M., & Donnelly, J.P. (2006). *The research methods knowledge base.* Atomic Dog.

Ukponmwan, H.N., & Ajibade, F.B. (2017). Evaluation of techniques for univariate normality test using monte carlo simulation. *American Journal of Theoretical and Applied Statistics*, *6*(5), 51-61. https://doi.org/10.11648/j.ajtas.s.2017060501.18

Wilcox, R.R. (2010). *Fundamentals of modern statistical methods: Substantially improving power and accuracy.* Springer-Verlag.

Yap, B.W., & Sim, C.H. (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation, 81*(12), 2141-2155. https://doi.org/10.1080/00949655.2010.520163

APPENDIX

Description of Normality Tests

Kolmogorov Smirnov Test	$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \le x}$ $D_n = \max_x \left\lvert F_n(x) - F(x) \right\rvert$	$F_n(x)$= the empirical distribution function n= number of observations $I_{X_i \le x}$ = the indicator function If the value of D_n is greater than the critical value, the null hypothesis is rejected.
KS Stephens Modification	$D^* = D_n\left(\sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}}\right)$ $p-value = \begin{cases} D^* < 0.775 & p > 0.15 \\ 0.775 < D^* < 0.819 & 0.10 < p < 0.15 \\ 0.819 < D^* < 0.895 & 0.05 < p < 0.10 \\ 0.895 < D^* < 0.995 & 0.025 < p < 0.05 \\ 0.995 < D^* < 1.035 & 0.01 < p < 0.025 \\ D^* \ge 1.035 & p < 0.01 \end{cases}$	n= number of observations If p-value is smaller than 0.05, the null hypothesis is rejected.

KS Marsaglia Method	$\Pr(D_n \le d) = \frac{n!}{n^n} H^n,\ d = \frac{k-h}{n}$ $p-value = \Pr(D_n > d) = 1 - \Pr(D_n \le d) = 1 - \frac{n!}{n^n} H^n$	n= number of observations H is a *m x m* matrix that depends on h only. k is a positive integer $0 \le h < 1$ If p-value is smaller than 0.05, the null hypothesis is rejected.
KS Lilliefors Modification	$p-value = \begin{cases} D_n = D10 & 0.10 \\ D_n > D10 & \exp\left(aD_n^{\,2} + bD_n + c - 2.3025851\right) \\ D_n \ge D15 & 0.15 + (D_n - D15)\left[(0.10-0.15)/(D10-D15)\right] \\ D_n \ge D20 & 0.20 + (D_n - D20)\left[(0.15-0.20)/(D15-D20)\right] \\ D_n \le D10 & p > 0.20 \end{cases}$	D10, D15 and D20 corresponding to n from the table given by Dellal and Wilkinson (1986). If p-value is smaller than 0.05, the null hypothesis is rejected.

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 2, 410–429
https://doi.org/10.21449/ijate.904456
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Estimation of the Academic Performance of Students in Distance Education Using Data Mining Methods

Resul Butuner[1,*], **M. Hanefi Calp**[2]

[1]Ankara Beypazarı Fatih Vocational and Technical Anatolian High School University, Faculty of Education, Department of Computer, Ankara, Türkiye
[2]Ankara Hacı Bayram Veli University, Faculty of Economics & Administrative Sciences, Department of Management Information Systems, Ankara, Türkiye

ARTICLE HISTORY
Received: Mar. 27, 2021
Revised: Feb. 03, 2022
Accepted: Mar. 15, 2022

Keywords:
Distance Education,
Academic Performance,
Estimation,
Data Mining,
Artificial Intelligence.

Abstract: Many institutions in the field of education have been involved in distance education with the learning management system. In this context, there has been a rapid increase in data in the e-learning process as a result of the development of technology and the widespread use of the internet. This increase is in the size of large data. Today, big data can be primarily processed, the relationships between data can be discovered, a meaningful conclusion can be drawn, and predictions about the future using big data can be made. However, these data are generally not used in a way to contribute to the people and institutions (educators, education administrators, ministries, etc.) involved in the education process. Therefore, this study aims to estimate the academic success of students who receive education in the distance education process using data mining methods. The reason why data mining is used is that these methods are particularly effective and powerful tools in classification and prediction processes. The methods used in the study are Random Forest, Artificial Neural Networks, Naive Bayes, Support Vector Machines, Logistic Regression, and Deep Learning algorithms, respectively. The dataset includes primary, secondary, and high school students' data, which were obtained from the learning management system used in the distance education process. As a result, the study findings showed that Deep Learning, Random Forest, and Support Vector Machines algorithms provide prediction success at higher performance than others.

1. INTRODUCTION

Coronavirus (COVID-19) is an epidemic disease that spreads all over the world in a very short time and has fatal consequences. This epidemic has adversely affected many institutions. Therefore, working methods in transportation, industry, health, and education have started to be carried out remotely (Karakaya et al., 2020; Savas et al., 2021). In the field of education, educators continue this process by providing students with education through the method of distance education (Yamamoto & Altun, 2020; Yilmaz & Buyrukoglu, 2021). Some tools are used in distance education. One of these tools is the Learning Management System (LMS). LMS is software created to manage e-learning processes more efficiently and effectively. Using

*CONTACT: M. Hanefi CALP ✉ hanefi.calp@hbv.edu.tr Ankara Hacı Bayram Veli University, Faculty of Economics & Administrative Sciences, Department of Management Information Systems, Ankara, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

this web tool, an organization can systematically carry out, manage, and measure training activities. LMS provides teachers and students with an online classroom that reinforces learning processes. In online classroom environments, LMS reinforces teachers and students in the learning process. A standard LMS supports an inclusive learning environment for academic progress with interceding structures that promote online collaborative groupings, professional training, discussions, and communication among other LMS users (Dias & Dinis, 2014; Jung & Huh, 2019; Oakes, 2002). It is important to understand the function and responses of the human brain, which is one of the current research topics, and the events of solving the working principle of the brain. It is because it will help to understand the level of learning and changes, especially in individuals (Dogan, 2012). It is considered an academic achievement to state the status of learning with various measurement methods and values such as grades or points (Turgut & Baykul, 2013). It is an important issue in increasing success to know the academic achievement levels of students in advance as it will greatly benefit both students and educational institutions (Luan, 2002). Data such as grades and scores are transformed into information by processing with computer software for a specific purpose (Kurt & Erdem, 2012; Savas, 2021). Therefore, techniques that process data and make them usable have become of great importance today. In this context, the conversion process of raw data into information or meaningful results is realized by data mining methods (Kiray et al., 2015). Statistical methods may not always yield meaningful results in analyzing the data collected in the LMS and revealing meaningful information. Data mining methods are used to process and analyze data on these issues (Beitel, 2005; Luan, 2002; Siemens & Baker, 2012).

Data mining is defined as the manual or automatic processing of large amounts of data to obtain meaningful results by reaching meaningful data. The field of data mining is gradually developing as a result of the processing of data warehouses with the electronic storage and development of data and the development of various analysis tools. It can be seen as obtaining new information by processing large-scale data belonging to students through data mining in the field of education. It is necessary to store the data more effectively and efficiently, and the analysis to be made on the data should work effectively in the background since a lot of data is kept for students, managers, and parents in education. In this process, data mining produces descriptive and predictive models for processing recorded data and generating information. The interest in this method is increasing, and new methods are emerging for more successful decisions due to the high benefit of the prediction method in the decision-making process (Akpinar, 2000; Uzut & Buyrukoglu, 2020; Yurtoglu, 2005).

The studies using data mining in traditional and distance education, the prediction of academic success for students, and the determination of the reasons that affect their failures have started to be seen frequently in the field of education (Ozbay, 2015). Data mining for education is an understanding of a discipline that develops methods to examine large-scale data and it uses these methods to understand students' performance and learning environments more comprehensively (Algami, 2016). Data acquisition from LMS, social network analysis, and visualization are used in this area. The process of transforming raw data collected from students, teachers, and parents into meaningful results by educators, researchers, and software in the education system is defined as educational data mining (García et al., 2011). The meaningful results obtained from educational data mining can be used as information that management will need to increase the competence and productivity of educational institutions. It is beneficial to identify problems in educational institutions and to create a more efficient and productive educational environment in terms of analyzing the data of students with the method of data mining in education, predicting students' success, and determining the reasons for their failure (Ozbay, 2015). Predictions can be created, and models for students' academic performance, and these predictive models can be used to advise students for their academic studies by using data mining methods in the field of education. In addition, new approaches to learning analytics can

be created, and students' profiles can be modeled through data mining methods in education. Individualized education environment, curriculum, and new learning styles can be created by classifying similar students (Bienkowski et al., 2012). Currently, the methods used in the field of education cause a lot of time and great effort. However, less time will be spent, which will lead business processes to be made automatically using the models developed with data mining (Lopez et al., 2012).

Many studies using different methods, techniques, and applications are encountered when the field is searched on the subject. Subbanarasimha et al. (2000) compared two different datasets by using artificial neural networks (ANN) and regression techniques to predict the academic performance of MBA students. As a result, a high success rate was obtained with the ANN model (Subbanarasimha et al., 2000). Guneri and Apaydin (2004) classified student achievement status using ANN and Logistic Regression (LR) methods. The classification success rate was found to be 95% in the methods used in the classification of success situations. Ibrahim and Rusli (2007) compared algorithms by predicting student success using ANN, Decision Trees (DT), and linear regression methods in data mining. It was found that ANN analysis gave better results in the prediction of overall academic achievement.

Bresfelean et al. (2008) used classification through learning, and data clustering methods to determine the academic success or failure of students in the study on the "Farthest First" algorithm and "Weka J48". They achieved quite successful results and stated that the methods used could be used effectively in education. Sembiring et al. (2011) developed a model using data mining methods to analyze student behavior and achievements and predict student performance. They revealed that the model they developed was quite effective. Sengur and Tekin (2013) estimated the graduation grades of students at the Department of Computer and Instructional Technologies in the Faculty of Education at Fırat University by using ANN and DT data mining methods. According to the results, it was observed that ANN provided better prediction performance compared to DT. Buyrukoglu and Yilmaz (2021) proposed a semi-automated data mining method for answering questions asked by students for online learning. This case-based reasoning model provided 84% time saving for instructors.

Akcapinar (2014) developed a model by using data mining methods on the data from an online learning environment for 76 students. He clustered the academic achievements of the course as “passed-fail” and categorized similar student profiles. Aydin (2015) predicted students' success in courses using variables such as course name, time passed through the e-service, the number of times students took the course, an average of exams taken, and students' age, using different classification algorithms. Kiray et al. (2015) analyzed the success of Turkish students in science and mathematics with the data mining method based on the data obtained from the international TIMSS and PISA exam results of Turkish students in their study. At the end of the study, they revealed the variables that affect science and mathematics achievement. Amrieh et al. (2016) developed a model to estimate the academic performance of the students with ANN, Naive Bayes (NB), and DT data mining methods using the data from 480 students through E-Learning. The reliability of the model was proven by providing 80% accuracy in the study.

Ozbay and Ersoy (2016) examined the relationship between the mobility of undergraduate students on the LMS and their academic achievement using data mining methods. It has been revealed that there is a significant relationship between mobility on LMS and academic achievement levels using log records containing the mobility of 40 students on LMS and year-end academic achievement scores. In another study, a model was developed by estimating 80% of when the students will complete the course based on the first day of the online course (Cunningham, 2017). Alsuwaiket (2018) proposed a model that predicts the academic performance of students in mathematics lessons that completed the 4th grade for students, teachers, and school principals. Altun et al. (2019) created models to estimate the academic

graduation average of the students, using the data such as gender, marital status, age of enrollment, and midterm exam scores in the 1st semester of the 1st grade. In the study, 94.30% success was obtained from the model made with regression analysis and 94.43% from the ANN model. Aydemir (2019) used 3794 students' data taking the Foreign Language-II course in the study on data mining at a university in Turkey. In the study, prediction models developed by ANN, M5P, Decision Stump, M5Rules, Decision Table, and Bagging methods were created and compared with each other. It was concluded that the model established with the bagging method produced estimates with the best result, 1.22 mean absolute error and 0.80 correlation coefficient. It was concluded that the students would learn passing grades of the course in advance and take precautions.

In summary, data mining methods can enable us to determine the field in which the students are successful, to obtain their level of success, and the factors that affect their success and cause failure. In this context, the aim of the study is to predict the academic success of students with data mining methods using the data obtained from the distance education system. For this purpose, all the details of the material and method were given in the second part of the study. The third part contains the experimental results and the discussion of these results. Finally, the conclusion and recommendations from the study were included in the fourth part.

2. MATERIAL and METHOD

In this section, all the details about the method and technique of the study were given. First of all, the data of the study were obtained from the official website of Kaggle (Aljarah, 2016), which is an open platform "https://www.kaggle.com/aljarah/xAPI-Edu-Data", and it includes fields such as gender, place of birth, nationality, education stages. Data from 304 students were used which have different country codes (from the student questionnaire in the CSV data file format) as a data source for the analysis. Data mining methods were used to extract meaningful information from the used dataset. E-Learning data and personal data obtained from LMS belonging to preschool, secondary school, and high school students were used in the study. These data produce 3 class outputs using data mining methods. These are Low Level (0-69 point range-L), Medium Level (70-89 point range-M), and High Level (90-100 point range-H). In this framework, the study was explained in detail with the process of the study, the dataset, data definition, the algorithms used, and the creation of the model.

2.1. The Process of the Study

The process of the study consisted of three stages. First, the data were organized by pre-processing and distinguishing from the noise before being analyzed. Later, the data were analyzed and visualized. Rapid Miner Studio and Orange applications, which are platforms for data science, were used to analyze data and determine validity. A comparison of the results obtained from these two platforms was made. The purpose of using this software was to analyze data, use various data mining techniques with pre-processing on data and evaluate these models by creating new models. Finally, the prediction results were obtained by training the data utilizing the seven algorithms in the study, and the success levels of these prediction results were listed in detail. The steps of data cleaning, data integration, data selection, data transformation, data mining, and presentation of results were carried out in the study, in general, to obtain meaningful information from the data.

2.2. Dataset

The dataset is an educational dataset collected from the LMS called Kalboard 360. Kalboard 360 is a multi-agent LMS designed to facilitate learning using the latest technology. This system provides users with simultaneous access to educational resources from any device with an internet connection. The data were collected using a student activity-tracking tool called the

experience API (xAPI). When the data set for the academic performance of the students is examined, it has a multivariate structure with 115 female students, 189 male students, 16 features, and classifications. It was determined that there were noisy data for the features used in the study when the data were analyzed, and these data were removed from the dataset. As a result of the extracted dataset, there were data belonging to 304 students in total, and this number represents 63.33% of the whole dataset. The dataset included students from different countries including Kuwait, Jordan, Palestine, Iraq, Lebanon, Tunisia, Saudi Arabia, Egypt, Syria, USA, Iran, and Libya. The properties of the dataset were given in Table 1 in detail (Aljarah, 2016).

Table 1. *Properties of the dataset(Aljarah, 2016).*

Nu	Attribute Name (Input)	Type	Value Range
1	Gender	Binominal	Male (1), Female (0)
2	Country	Polynomial	Kuwait, Jordan, Palestine, Iraq, Lebanon, Tunisia, Saudi Arabia, Egypt, Syria, USA, Iran, Libya
3	Place of birth	Polynomial	Kuwait, Jordan, Palestine, Iraq, Lebanon, Tunisia, Saudi Arabia, Egypt, Syria, USA, Iran, Libya
4	School Level	Polynomial	Primary School, Secondary School, High School
5	Grade Level	Polynomial	G-01, G-02, G-03, G-04, G-05, G-06, G-07, G-08, G-09, G-10, G-11, G-12
6	Section ID	Polynomial	A, B, C
7	Lesson Topic	Polynomial	English, Spanish, French, Arabic, IT, Mathematics, Chemistry, Biology, Science, History, Quran, Geology
8	Semester	Binominal	First, Second
9	Parent of Student	Binominal	Mother, Father
10	Number of Hand Raises	Number	0-100
11	Number of Resources Visited	Number	0-100
12	Number of Views Announcements	Number	0-100
13	Discussion Numbers	Number	0-100
14	Parent Responding Questionnaire	Number	0-100
15	Parent School Satisfaction	Binominal	Yes, No
16	Student Attendance Day	Binominal	Above 7, Below 7
17	Class	Polynomial	L (Low Level), M (Medium Level), H (High Level)

2.3. Data Cleaning

The following processes were carried out in the data cleaning: completing missing data, removing inconsistencies, detecting outliers, and removing noise. Data cleaning must be realized before data analysis. The following methods can be used to complete missing data (Oguzlar, 2003).

- Records with missing values can be deleted.
- Average value can be used instead of missing values.
- Median can be used instead of missing values.
- Instead of missing values, the average of the class in which it is located can be taken.
- The most appropriate value can be used instead of missing values by using methods such as regression.

2.4. Algorithms Used in the Study

In this section, each of the algorithms used in the study was briefly explained.

2.4.1. *Random forest*

The random forest model is a tree-based learning algorithm. The algorithm aims to combine tree decisions trained in different training sets and present them to the user instead of a single decision tree. While determining the attributes of each level, some calculations are made in all trees, and the attribute is determined. Then, the attributes in other trees are combined, and the most used attribute is selected. The selected attribute is included in the tree, and the process is repeated at all levels. In order to start the algorithm, the number of variables and the number of trees to be used in each node should be determined by the user (Breiman, 2001; Resende,2018).

2.4.2. *Naive bayes*

Naive Bayes is a kind of probabilistic classification mechanism based on Bayes' theorem. It aims to create a simple and effective statistical forecast showing a high level of success rate in application areas. The basic logic of the algorithm is based on the classification process according to the most appropriate label and the dependence of the attributes in the data only on a certain class (Bayes, 1763; Yildiz et al., 2007).

2.4.3. *Support vector machine*

Support vector machines are machine learning algorithms based on convex optimization aimed at minimizing risk. It can learn independently since it does not need distribution information. It aims to obtain the most suitable hyperplane to separate the classes in the support vector machines technique. That is, it predicts maximizing the distance between different classes (Ayhan & Erdogmus, 2014; Cortes & Vapnik, 1995).

2.4.4. *K-Nearest neighborhood*

It is a supervised learning algorithm that performs learning based on the data in the training set. It is used for both classification and regression problems. It performs the classification process by comparing the new data in the group with the data in the training set. Each sample in the training set is kept to represent a point in space. The k samples in training are set closest to the new sample, and the class of the new sample is determined when a new sample joins space (Fix, & Hodges, 1951; Han et al., 2011; Kilinc et al., 2016).

2.4.5. *Logistic regression*

It is used to match an item of data to a real-valued prediction variable. The main goal is to fit the data to a known type of function. In the regression technique, it is tried to find the function that best models the data given in the process. It is used to determine which function is the best by determining the difference between actual values and predictions (Tolles et al., 2003).

2.4.6. *Artificial neural networks*

ANNs are an information processing system used in solving nonlinear and complex problems, performing the learning function, which is the most basic feature of the human brain, and they can use experiments, to generate, create, discover, and predict new information without any help. ANN has a structure that builds a different calculation technique compared to the traditional calculation techniques, and it can adapt to its environment. It can also make decisions in uncertain situations. ANN is used effectively in many different areas such as forecasting, optical character handling, fingerprint recognition, pattern recognition, robot technology, job scheduling and quality control, power systems, system modeling, finance applications, image processing, industrial applications, and defense applications (Butuner & Yuksel, 2021; Calp, 2019; Calp & Kose, 2020).

2.4.7. *Deep learning*

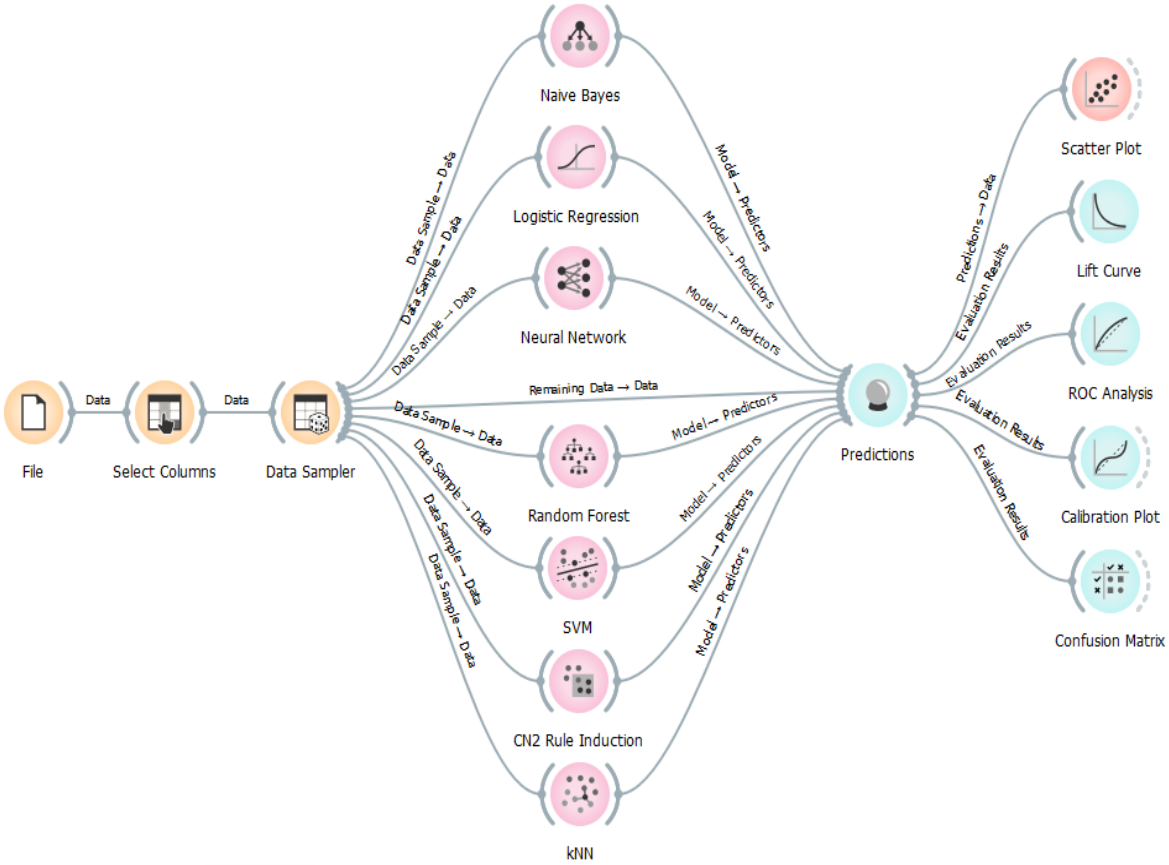
It is a technique that is used for machines to perceive and understand the world and provides solutions for nonlinear problems by using more than one layer. Generally, it is seen that it is

used more in areas such as data analysis, image classification, video analysis, speech recognition, information retrieval, object recognition, and natural language learning (Salman et al., 2020). In the deep learning technique, data is based on learning from the representation of data by learning more than one feature level. Deep learning methods are generally developed on ANNs, but they have more hidden neurons and layers. Deep learning methods have yielded very successful results in processing many types of data, such as video, audio, and text (Butuner, 2020; Calp, 2021).

2.5. Creating the Models

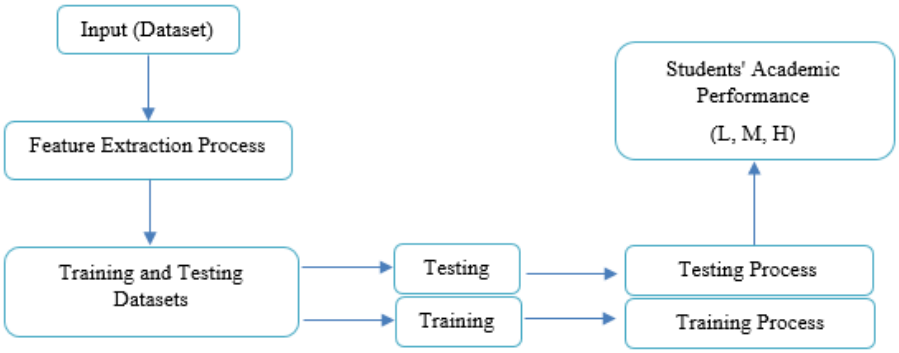
Rapid Miner and Orange programs were used to create the models. First of all, the methods to be used and their properties were determined. Then, the dataset containing the inputs and outputs of the study was loaded into the system. The models were created. Finally, the results and graphics of the created models were obtained (Figure 1).

Figure 1. *The process of creating models and obtaining results.*



There are two separate processes in the created models training and testing. First of all, the training of the model was carried out, then the testing process was applied. The performance of the models created was measured in the test part, and then the academic performance of the students was determined (Figure 2).

Figure 2. *Training, Testing, and Output Process.*



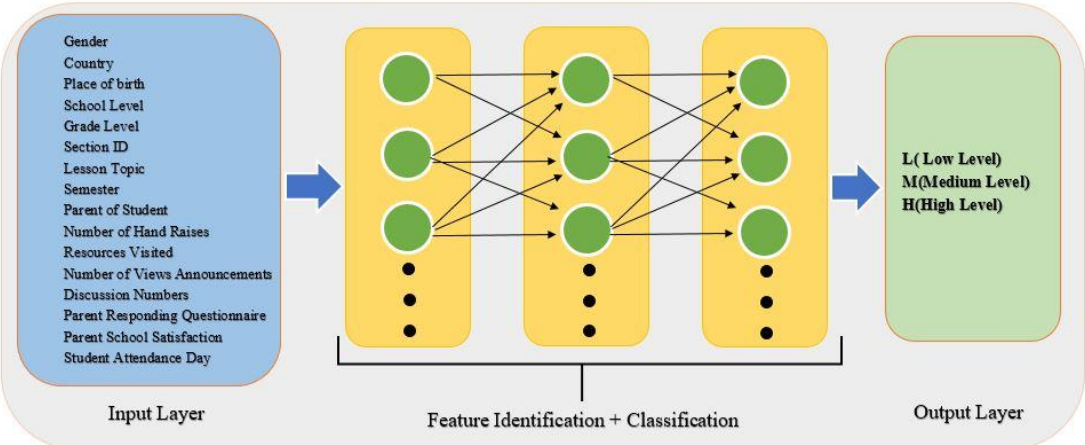
In the creation process of the models, the data were divided into two parts 75% to be used in the training of models, and 25% to be used in the testing process (Table 2). At this point, 228 of a total of 304 datasets belonging to preschool, secondary school, and high school students contained the training set, and 76 of them contained the testing set. These ratios and numbers may change to obtain the best model based on trial and error. The factor in determining these ratios and numbers used in the study was to obtain the training level of the model that gave the best result with these values.

Table 2. *The dataset used in the creation of the model.*

Purpose of usage	Number	%
Training	228	75
Testing	76	25

The mixed sampling "Shuffle Sampling" technique was used in the running of the models. RMSE (Root Mean Squared Error), CE (Classification Error), R2 (SC) Squared Correlation, and RE (Relative Error) performance measurement tools were used to evaluate the models. The general structure of the system used in all models was given in Figure 3. According to the general structure, 16 data received from the input layer were processed by using algorithms in feature determination and classification blocks, and meaningful results were produced for the output layer. These results were grouped as L (Low Level), M (Medium Level), and H (High Level).

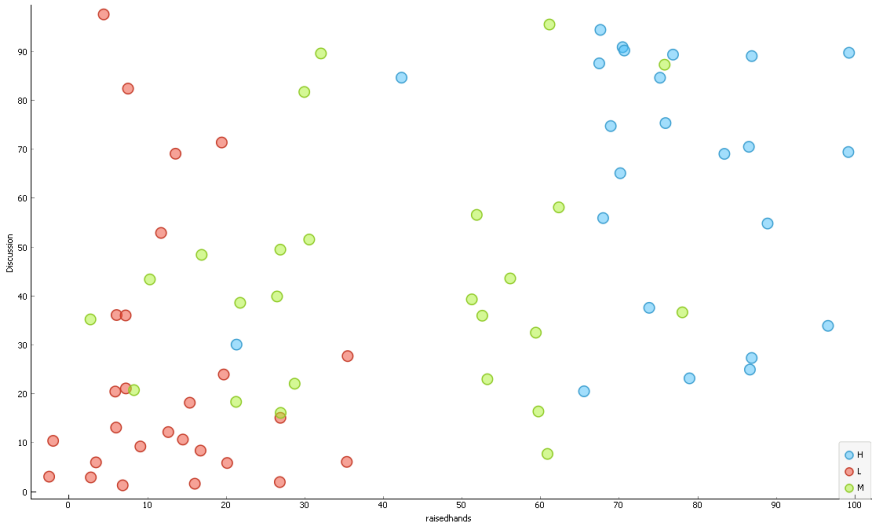
Figure 3. *The general structure of the system.*



3. EXPERIMENTAL RESULTS and DISCUSSION

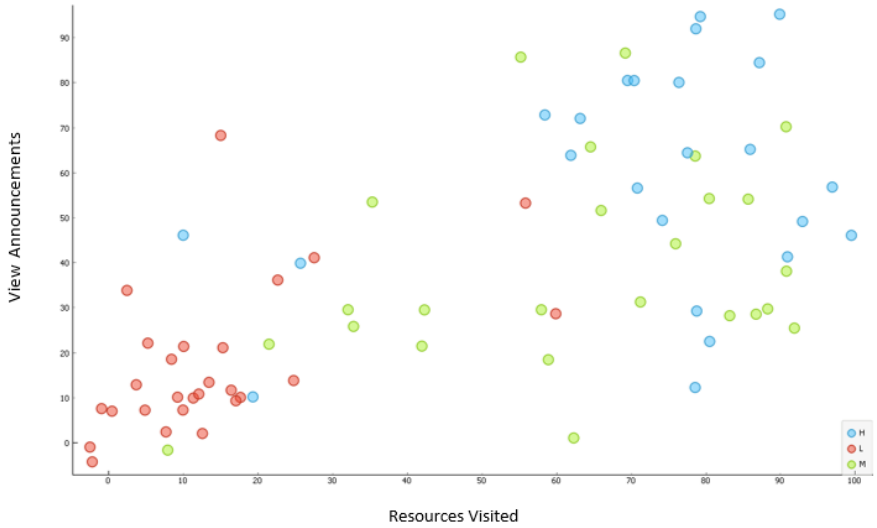
In this section, the experimental results obtained from the study were given, and these results were discussed in detail. In this context, data mining methods that consist of 16 column data entries belonging to students and produce meaningful results at three levels Low (L), Medium (M), and High (H) level were compared. In Figure 4, the correlation graph of the success levels in the Raisedhands and Discussion areas was shown. According to the density of the students in the Raisedhands and Discussion groups, it was seen that the success levels are High Level (H), but only the student density in the discussion groups was at Low Level (L). The students who gathered in the Raisedhands and Discussion groups in medium intensity were gathered at the Medium Level (M).

Figure 4. *The correlation graph between Raisedhands and Discussion.*



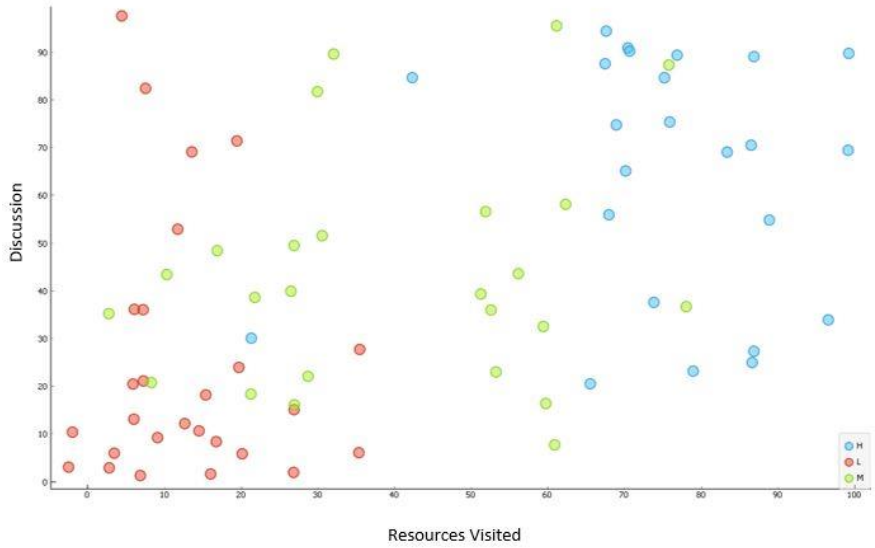
In Figure 5, the correlation graph between the Resources Visited and View Announcements by the students on the distance education platform was presented. According to the graph, student levels were also at (H) level and (M) level in direct proportion to the intensity of both elements (Resources Visited by students and View Announcements). It was observed that there was a student density at the (L) level in the region where the density of the resources visited by the students was low.

Figure 5. *The correlation graph between Resources Visited and View Announcements.*



In Figure 6, the correlation graph between the Resources Visited and Discussion areas by the students on the distance education platform was given. It was seen that the success levels were High Level (H) according to the density of the students in the discussion groups, but only the student density in the discussion groups were collected at the Low Level (L). It was observed that there was a student density at the (M) level in the region where the density of the resources visited by the students was medium in the Resources Visited and Discussion groups.

Figure 6. *The correlation graph between the Resources Visited and Discussion.*



The comparison of the academic achievement performances of the students according to the models used was given in Table 3. When Table 3 was examined, the success rates of over 96% were achieved with Deep Learning (DL), Support Vector Machines (SVM), LR, and Random Forest (RF) algorithms. In addition, the validation percentages of these algorithms obtained a 99% success rate.

Table 3. *Comparison of the models.*

Model	Validation	F1 Score	Accuracy	Recall
NB	0.965	0.896	0.904	0.895
LR	0.998	0.961	0.961	0,961
ANN	0.975	0.934	0.935	0.934
RF	0.994	0.960	0.961	0.961
SVM	0.996	0.974	0.960	0.974
DL	0.988	0.969	0.978	0.962
K-NN	0.929	0.839	0.850	0.842

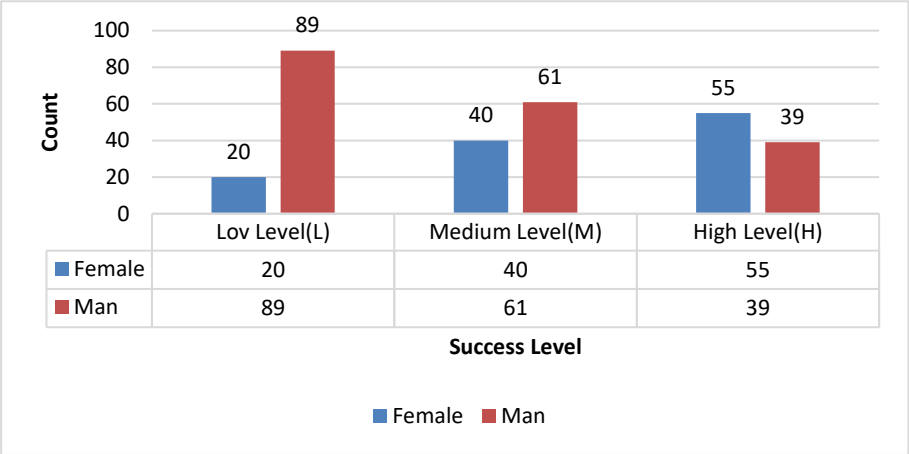
The information about the performance of the models prepared for the use of parameters according to the RF and DL algorithms was compared and given in Table 4.

Table 4. *Comparison of the results according to the models created.*

Model	Dataset	Testing	CE	RMSE	R^2 (SC)	RE
RF	304	0.961	0.049 +/- 0.084	0.222	5.33%	16.40% +/- 14.96%
DL		0.978	2.20%	0.141 +/- 0.000	0.969	6.53% +/- 12.46%

Figure 7 shows the success levels in the dataset according to gender distribution. It was observed that the number of male students in the (L) level exceeds four times the number of female students, and the number of female students in the (M) level and (H) level is 1.5 times the number of male students. As a result, it was seen that the success level of female students from online platforms was higher than male students in the distance education process.

Figure 7. *Success graph according to gender.*



In Figure 8, the success classification in the dataset according to school types was given. It was observed that the density of middle school students whose success classification was (H) level was high, and the density of high school students was lower. At the (L) level, it was understood that the density of both primary school and secondary school students was high and close to each other. At the (M) level, it was seen that the density of secondary school students was high, and the density of high school students was low.

Figure 8. *Student success intensity graph according to school types.*

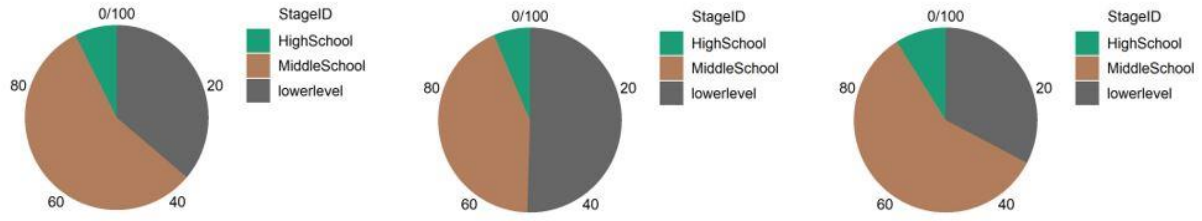


Figure 9 shows the success classification in the dataset according to absenteeism. Absenteeism was grouped as less than 7 days and more. It was seen that absenteeism of students with high levels of success was less than 7 days. On the other hand, it was seen that the students with low achievement levels have more than 7 days of absenteeism and the number of absenteeism of the students was concentrated in this number. In terms of success, the absenteeism of middle-level students under 7 days was approximately 1.5 times those above 7 days. In addition, it was understood that success classification according to the graph was inversely proportional to absenteeism and student numbers.

Figure 9. *Student success graph according to absenteeism.*

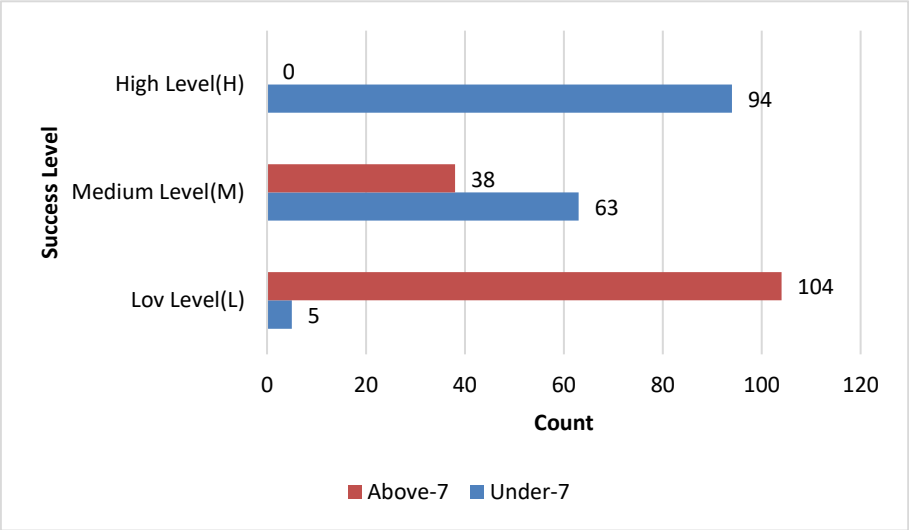
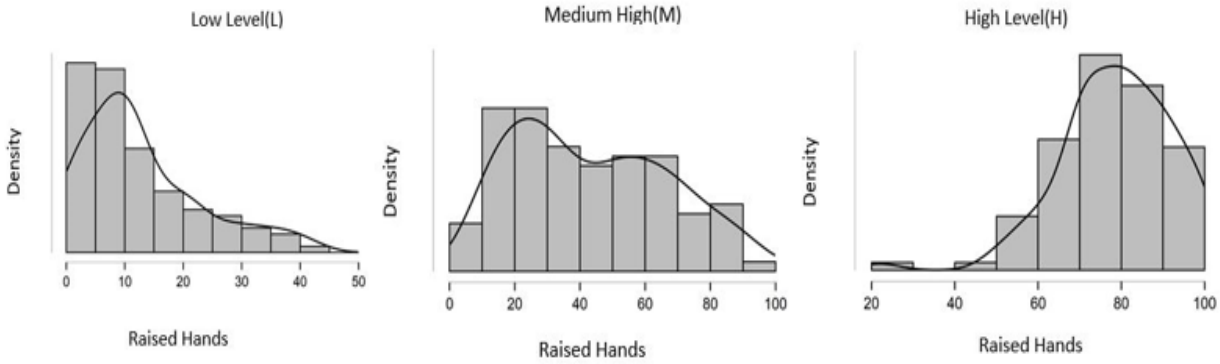


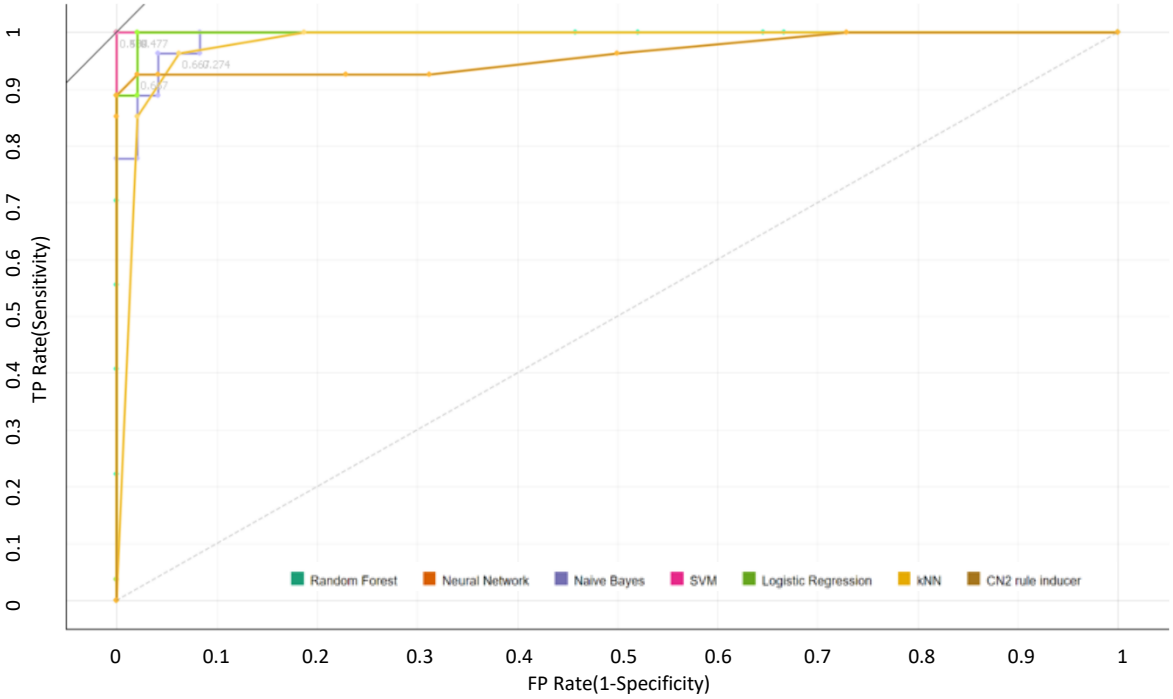
Figure 10 shows the graphs showing the success levels of students for asking questions by raising their fingers during the distance education process. As the number of Raisedhands of the students in the (L) level decreased in online education, the intensity of student achievement decreased. It revealed that while the Raisedhands intensity of the students was 20%, the (M) level had a peak value, and this value gradually decreased. It was understood that when the level of Raisedhands at (H) level made the peak value, the student success density was the highest.

Figure 10. *Relationship graph between Raised Hands and student achievement.*



In Figure 11, true positive and false positive ROC graphs of students with low achievement levels (L) according to the estimated percentages of the algorithms were given. According to the graph, the best accuracy rate was obtained from RF and DL algorithms.

Figure 11. *The accuracy graphs of the models (Low Level) (ROC Analysis).*



In Figure 12, the correct positive and false-positive ROC graphs of students whose achievement level was medium level (M) according to the estimation percentages of algorithms was given. According to the graph, the best accuracy rate was obtained from RF and DL algorithms.

Figure 12. *The accuracy graphs of the models (Medium Level) (ROC Analysis).*

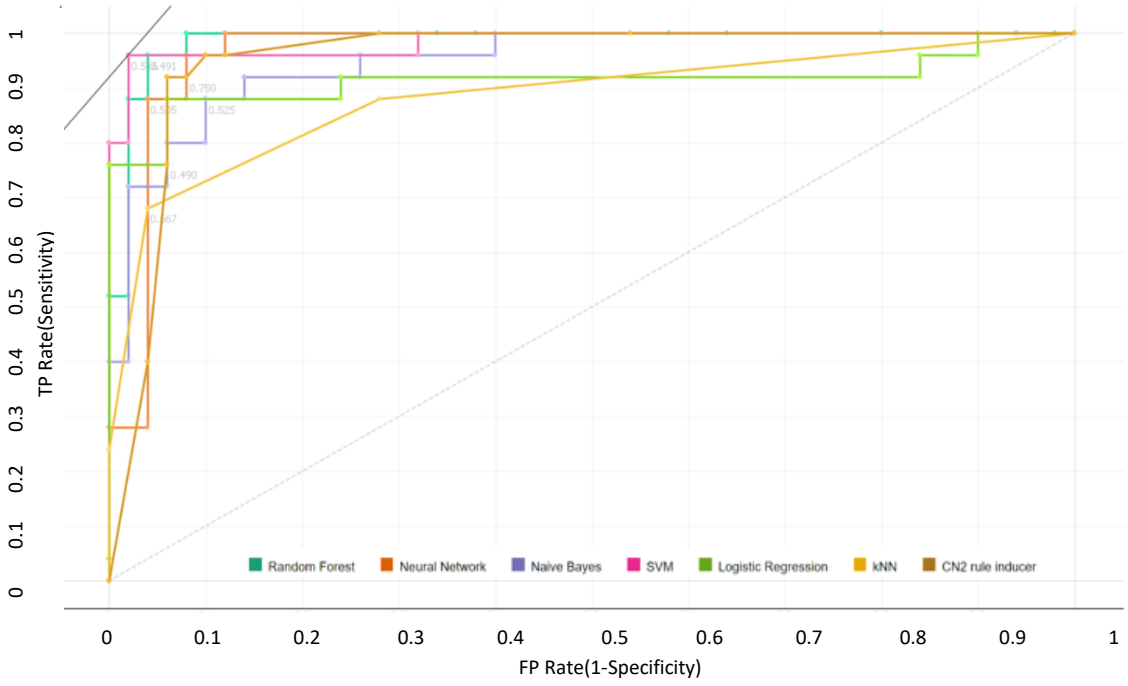


Figure 13 shows the correct positive and false-positive ROC graphs of the students with a high level of success (H) according to the estimated percentages of the algorithms. According to the graph, the best accuracy rate was again obtained from RF and DL algorithms.

Figure 13. *The accuracy graphs of the models (High Level) (ROC Analysis).*

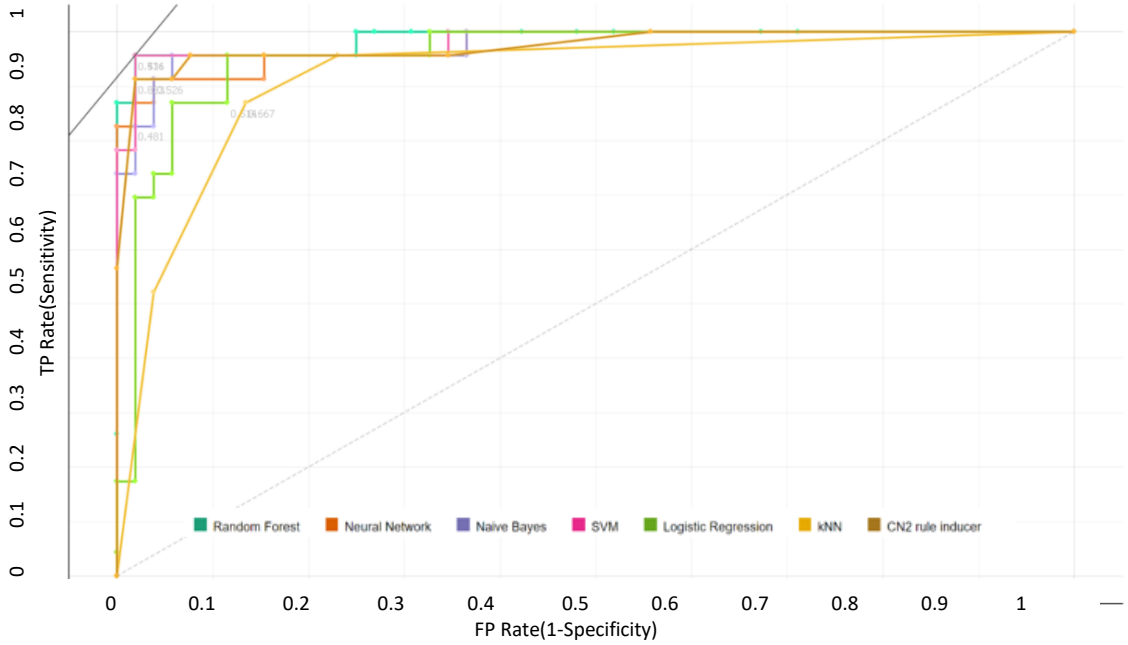


Table 5. *The confusion matrix results of the algorithms used in the study (respectively).*

Naive Bayes				
	H	L	M	∑
H	91.67%	0.00%	8.33%	24
L	7.41%	88.89%	3.70%	27
M	8.00%	0.00%	92.00%	25
∑	24	26	26	76

Artificial Neural Networks				
	H	L	M	∑
H	91.67%	4.17%	4.17%	24
L	3.70%	96.30%	0.00%	27
M	8.00%	0.00%	92.00%	25
∑	24	26	26	76

Logistic Regression				
	H	L	M	∑
H	95.83%	0.00%	4.17%	24
L	3.70%	96.30%	0.00%	27
M	4.00%	0.00%	96.00%	25
∑	24	26	26	76

Random Forest				
	H	L	M	∑
H	100.00%	0.00%	0.00%	24
L	3.70%	96.30%	0.00%	27
M	0.00%	8.00%	92.00%	25
∑	24	26	26	76

Support Vector Machines				
	H	L	M	∑
H	100.00%	0.00%	0.00%	24
L	3.70%	96.30%	0.00%	27
M	4.00%	4.00%	92.00%	25
∑	24	26	26	76

Deep Learning				
	H	L	M	∑
H	100.00%	0.00%	0.00%	24
L	7.41%	92.59%	0.00%	27
M	0.00%	0.00%	100.00%	25
∑	24	26	26	76

K-Nearest Neighbor				
	H	L	M	∑
H	87.50%	8.33%	4.17%	24
L	7.41%	81.48%	11.11%	27
M	8.00%	4.00%	88.00%	25
∑	24	26	26	76

The Confusion Matrix values for the algorithms used in the study were given in Table 5. It was seen that the Deep Learning model gave the best results according to (L) level, (M) level, and (H) level categories. In this DL model, the ratio of (H) level was 100%, (L) level was 92.59%, and (M) level was 100%. On the other hand, the KNN model was obtained with (L) level 81.48%, (M) level 88%, and (H) level 87.50%, with the lowest results in the study.

Finally, 15 (fifteen) different tests were carried out using real data with the proposed RF, ANN, KNN, NB, SVM, DL, and LR models. Experimental results obtained from these tests were given in Table 6.

Table 6. *Experimental results obtained from the models.*

Nu	RF	RF-Error	ANN	ANN-Error	KNN	KNN-Error	NB	NB- Error	SVM	SVM- Error	DL	DL- Error	LR	LR- Error	Output
1	0.86	0.14	0.86	0.14	0.75	0.25	0.92	0.08	0.86	0.14	0.90	0.10	0.88	0.12	M
2	0.94	0.09	0.99	0.01	0.58	0.42	0.93	0.07	0.98	0.02	0.99	0.01	0.94	0.06	H
3	0.94	0.06	0.94	0.06	0.83	0.17	0.90	0.10	0.97	0.03	0.99	0.01	0.93	0.07	H
4	0.95	0.05	0.99	0.01	0.75	0.25	0.94	0.06	0.95	0.05	0.97	0.03	0.97	0.03	M
5	0.95	0.06	0.94	0.06	0.50	0.50	0.98	0.02	0.96	0.04	0.98	0.02	0.95	0.05	M
6	0.97	0.05	0.93	0.07	0.83	0.17	0.92	0.08	0.98	0.02	0.97	0.03	0.97	0.03	M
7	0.97	0.03	0.93	0.07	0.92	0.08	0.91	0.09	0.94	0.06	0.98	0.02	0.99	0.01	H
8	0.96	0.08	0.94	0.06	0.42	0.58	0.85	0.15	0.97	0.03	0.93	0.07	0.96	0.04	H
9	0.97	0.03	0.97	0.03	0.92	0.08	0.92	0.08	0.99	0.01	0.98	0.02	0.94	0.06	L
10	0.96	0.06	0.90	0.10	0.92	0.08	0.91	0.09	0.97	0.03	0.96	0.04	0.97	0.03	L
11	0.97	0.03	0.90	0.10	0.92	0.08	0.95	0.05	0.96	0.04	0.99	0.01	0.96	0.04	M
12	0.99	0.06	0.99	0.01	0.92	0.08	0.89	0.11	0.98	0.02	0.97	0.03	0.96	0.04	L
13	0.96	0.06	0.89	0.11	0.83	0.17	0.76	0.24	0.99	0.01	0.99	0.01	0.98	0.02	L
14	0.98	0.02	0.95	0.05	0.75	0.25	0.95	0.05	0.98	0.02	0.99	0.01	0.97	0.03	L
15	0.98	0.02	0.95	0.05	0.58	0.42	0.88	0.12	0.97	0.03	0.98	0.02	0.98	0.02	M
Ort.	0.96	0.06	0.94	0.06	0.76	0.24	0.91	0.09	0.96	0.04	0.97	0.03	0.96	0.04	

To understand the success of the study more clearly, the obtained findings and test results were compared with other studies in the literature (Table 7). When Table 7 was examined, it was seen that the performance of the study was quite successful compared to other studies.

Table 7. *Comparison of the proposed study with the studies in the literature.*

Nu	Year	Authors	Number of Data (Student)	Method	Success Rate
1	2004	Guneri, N., & Apaydin, A. (2004).	352	LR	%95.17
				ANN	%97.14
2	2013	Sengur, D., & Tekin, A. (2013).	127	ANN	%93
				DT	%76
3	2013	Cokluk, O. T. D., & Cırak, G. Y. (2012).	419	LR	%66.10
				ANN	%70.16
4	2013	Turhan, K., Kurt, B., & Engin, Y. Z. (2013).	111	LR	%85
				ANN	%93
5	2016	Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016).	500	DT	%77.8
				NB	%72.4
				RF	%76.6
6	2016	Ozbay, O., & Ersoy, H. (2017).	40	CART	%85
				C5.0	%82.5
				CHAID	%65
				QUEST	%65
7	2019	Altun, M., Kayikci, K., & Irmak, S. (2019).	578	ANN	%94.43
				LR	%94.30
8	2019	Aydemir, E. (2019).	3794	M5P	%12
9	2018	Aydogan, I., & Zırhlıoglu, G. (2018).	1049	ANN	%97,2
10	2022	Butuner, R. & Calp, M. H. (2022) -The results of the models used in this study-	304	ANN	%97
				SVM	%99
				DL	%98
				NB	%96.5
				LR	%99.8
				RF	%99.4
				K-NN	%92.9

4. CONCLUSION and RECOMMENDATIONS

Education is the most important element for the future of society. It is inevitable that data mining methods, which have successful applications in many fields in the 21st century, will also be applied in the field of education and create new concrete outputs. This study proposed a model that predicts the academic achievements of primary, secondary, and high school students who receive education with the learning management system in the distance education process using data mining methods. 7 different models were created with data mining methods. Predictions were made in 3 classes Low (L), Medium (M), and High (H) according to the students' grades using the created models.

It can be said that predicting the academic performance of students is meaningful in increasing academic success when the models created in the study are evaluated. In addition, it was seen that results with high accuracy were obtained in predicting the academic success of students when the study was compared with similar studies using data mining techniques.

Finally, today, big and various data are stored in the systems of many institutions in Turkey such as e-government services, MEBSIS, and e-school which depend on the Ministry of National Education, OSYM, Universities, and Ministry of Health. However, it is seen that data mining or artificial intelligence techniques have not been used enough to draw meaningful

results from these data. It is obvious that the quality and production success of the institutions will increase by processing these data by each institution and obtaining meaningful and concrete outputs. As a result, it is expected that this study will set an example for other studies to increase student success.

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

Resul BUTUNER: Investigation, Resources, Visualization, Software, - original draf original draft. **M. Hanefi CALP**: Introduction, Methodology and Validation, Formal Analysis, and Writing -original draf.

Orcid

Resul BUTUNER https://orcid.org/0000-0002-9778-2349
M. Hanefi CALP https://orcid.org/0000-0001-7991-438X

REFERENCES

Akcapinar, G., Altun, A., & Aşkar, P. (2015). Modeling students' academic performance based on their interactions in an online learning environment. *Primary education Online*, *14*(3), 815-824.

Akpinar, H. (2000). Information discovery and data mining in databases. *Istanbul University Journal of the School of Business*, 1-22.

Aljarah, I. (2017). *Students' academic performance dataset*. Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/aljarah/xAPI-Edu-Data

Alsuwaiket, M. (2018). *Measuring academic performance of students in higher education using data mining techniques* [Doctoral dissertation, Loughborough University].

Altun, M., Kayikci, K., & Irmak, S. (2019). Estimation of Graduation Grades of Primary Education Students by Using Regression Analysis and Artificial Neural Networks. *E-International Journal of Educational Research*, *10*(3), 29-43. https://doi.org/10.19160/ijer.624839

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. https://doi.org/10.14257/ijdta.2016.9.8.13

Aydemir, E. (2019). Forecasting of the Course Learning Notes by Data Mining Methods. *European Journal of Science and Technology*, 70-76. https://doi.org/10.31590/ejosat.518899

Aydin, S. (2015). Data Mining and an Application in Anadolu University Open Education System. *Journal of Research in Education and Teaching*, *4*(3), 36-44.

Aydogan, I., & Zirhlioglu, G. (2018). Estimation of Student Successes by Artificial Neural Networks. *YYU Journal of Education Faculty*, *15*(1), 577-610. http://dx.doi.org/10.23891/efdyyu.2018.80

Ayhan, S., & Erdogmus, S. (2014). Kernel Function Selection for the Solution of Classification Problems via Support Vector Machines. *Eskisehir Osmangazi University Journal of Economics and Administrative Sciences*, *9*(1), 175-201.

Bayes, T. (1763). LII. *An essay towards solving a problem in the doctrine of chances*. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. Philosophical transactions of the Royal Society of London, 370-418. https://doi.org/10.1098/rstl.1763.0053

Beitel, S. (2005). *Applying Artificial Intelligence Data Mining Tools to the Challenges of Program Evaluation*. Connecticut.

Bienkowski, M., Feng, M., & Means, B. (2012). *Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics: An Issue Brief*. Office of Educational Technology, US Department of Education.

Breiman, L. (2001). Random Forests, *Machine Learning, 45(1)*, 5–32.

Bresfelean, V.P., Bresfelean, M., Ghisoiu, N., & Comes, C.A. (2008, June 23-26). *Determining students' academic failure profile founded on data mining methods*. Proceedings of the ITI 2008 30th International Conference on Information Technology Interfaces, 317-322, https://doi.org/10.1109/ITI.2008.4588366

Butuner, R. (2020). *Sentiment Analysis with Deep Learning Methods and Its Use in School Guidance Services*, [Master's Thesis, Necmettin Erbakan University]. Coincil of Higher Education Libraries: https://tez.yok.gov.tr/UlusalTezMerkezi/TezGoster?key=fl0Kw4p1rmMDotyKRdYv1BKdBnLg10dCC3PJQ2laOIvx6m-b832uTqLlcfv5bVHP

Butuner, R., & Yuksel, H. (2021). Diagnosis and Severity of Depression Disease in Individuals with Artificial Neural Networks Method. *International Journal of Intelligent Systems and Applications in Engineering*, *9*(2), 55-63. https://doi.org/10.18201/ijisae.2021.234

Buyrukoglu, S., & Yilmaz, Y., (2021). A Novel Semi-Automated Chatbot Model: Providing Consistent Response of Students' Email in Higher Education based on Case-Based Reasoning and Latent Semantic Analysis, *International Journal of Multidisciplinary Studies and Innovative Technologies, 5*(1), 6-12.

Calp, M.H. (2019). An estimation of personnel food demand quantity for businesses by using artificial neural networks. *Journal of Polytechnic*, *22*(3), 675-686.

Calp, M.H. (2021). Use of Deep Learning Approaches in Cancer Diagnosis. In: Kose U., Alzubi J. (eds) *Deep Learning for Cancer Diagnosis. Studies in Computational Intelligence, vol 908.* Springer. https://doi.org/10.1007/978-981-15-6321-8_15

Calp, M.H., & Kose, U. (2020). Estimation of burned areas in forest fires using artificial neural networks. *Ingeniería Solidaria*, *16*(3), 1-22.

Cokluk, O.T.D., & Cirak, G.Y. (2013). The Usage of Artifical Neural Network and Logistic Regresssion Methods in the Classification of Student Achievement in Higher Education. *Mediterranean Journal of Humanities*, *3*(2), 71-79. https://doi.org/10.13114/MJH/201322471

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20 (3)*, 273-297. https://doi.org/10.1007/BF00994018. S2CID 206787478.

Cunningham, J. (2017). *Predicting student success in a self-paced mathematics MOOC* (Order No. 10272808). Available from *Pro Quest Dissertations & Theses Global*, (1900990574).

Dias, S.B., & Dinis, J.A. (2014). Towards an enhanced learning in higher education incorporating distinct learner's profiles. *Educational Technology & Society*, *17*(1), 307–319.

Dogan, A. (2012). *Yapay Zeka [Artificial intelligence]*. Kariyer Publishing.

Dunham, M.H. (2003). *Data mining introductory and advanced topics*. Prentice Hall.

Fix, E., & Hodges, Joseph L. (1951). Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. USAF School of Aviation Medicine, Randolph Field, Texas.

García, E., Romero, C., Ventura, S., & De Castro, C. (2011). A collaborative educational association rule mining tool. *The Internet and Higher Education*, *14*(2), 77-88.

Guneri, N., & Apaydin, A. (2004). Logistic Regression Analysis and Neural Networks Approach in the Classification of Students Achievement. *Gazi University Journal of Commerce & Tourism Education Faculty*, *1*, 170-188.

Han, J., Pei J. & Kamber, M., (2011). *Data Mining: Concepts and Techniques*. Elsevier.

Ibrahim, Z., & Rusli, D. (September, 2007). *Predicting Students' Academic Performance: Comparıng Artificial Neural Network, Decision Tree and Linear Regression*. 21st Annual SAS Malaysia Forum, (s. 5). Shangri-La Hotel, Kuala Lumpur.

Algarni, A. (2016). Data mining in education. *International Journal of Advanced Computer Science and Applications*, *7*(6), 456-461.

Jung, S., & Huh, J.H. (2019). An Efficient LMS Platform and Its Test Bed. *Electronics*, *8*(2), 154.

Karakaya, F., Arik, S., Cimen, O., & Yilmaz, M. (2020). Investigation of the views of biology teachers on distance education during the COVID-19 pandemic. *Journal of Education in Science, Environment and Health (JESEH), 6*(4), 246-258. https://doi.org/10.21891/jeseh.792984

Kilinc, D., Borandag, E., Yucalar, F., Tunali, V., Simsek, M., & Ozcift, A. (2016), Classification of Scientific Articles Using Text Mining with KNN Algorithm and R Language. *Marmara Journal of Pure and Applied Sciences, 3*, 89-94. https://doi.org/10.7240/mufbed.69674.

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 (JESEH), 1*(1), 28-48.

Kurt, C., & Erdem, O. (2012). Discovering the Factors Effect Student Success via Data Mining Techniques. *Journal of Polytechnic*, *15*(2), 111-116.

Lopez, M.I., Luna, J.M., Romero, C., & Ventura, S. (2012). Classification via clustering for predicting final marks based on student participation in forums. International Educational Data Mining Society.

Luan, J. (2002). Data Mining and Knowledge Management in Higher Education-Potential Applications. *42nd Associate of Institutional Research International Conference (s. 1-20). Toronto*, Canada: ERIC.

Oakes, K. (2002). E-learning: LCMS, LMS—They're not just acronyms but powerful systems for learning. Training & Development, *56*(3), 73–75.

Oguzlar, A. (2003). Data Preprocessing. *Erciyes University Journal of Faculty of Economics and Administrative Sciences*, *21*, 67-76.

Ozbay, O. (2015). Data Mining Concept and Data Mining Applications in Education. *The Journal of International Education Science, 5*, 262-272.

Ozbay, O., & Ersoy, H. (2017). Analysis of Student Dynamism into Learning Management.

Resende, P.A.A., & Drummond, A.C. (2018). A survey of random forest based methods for intrusion detection systems. ACM Computing Surveys (CSUR), *51*(3), 1-36.

Salman, F.M., Abu-Naser, S.S., Alajrami, E., Abu-Nasser, B.S., & Ashqar, B.A. (2020). COVID-19 Detection using Artificial Intelligence, *International Journal of Academic Engineering Research*, 18-25.

Savas, S., (2021). Artificial Intelligence and Innovative Applications in Education: The Case of Turkey, *Journal of Information Systems and Management Research, 3*(1), 14-26.

Savas, S., Guler, O., Kaya, K., Coban, G., & Guzel, M.S., (2021). Digital Games in Education and Learning through Games, *International Journal of Active Learning*, *6*(2), 117-140.

Sembiring, S., Zarlis, M., Hartama, D., Ramliana, S., & Wani, E. (2011). Prediction of Student Academic Performance by an Application of Data Mining Techniques. *International Conference on Management and Artificial Intelligence, 6*(1), 110-114). IACSIT Press.

Sengur, D., & Tekin, A. (2013). Prediction of Student's Grade Point Average by Using the Data Mining Methods. *Journal of Information Technologies, 6*(3), 7-16.

Siemens, G., & Baker, R. (2012). Prediction of student academic performance by an application of k-means clustering algorithm. *Towards Communication and Collaboration. 2nd international conference on learning analytics and knowledge*. Vancouver, Canada.

Subbanarasimha, P., Arinzeb, B., & Anandarajanb, M. (2000). The Predictive Accuracy of Artificial Neural Networks and Multiple Regression in the Case of Skewed Data. Exploration of Some Issues. *Expert Systems with Applications*, 117-123.

System through Data Mining Methods. *Journal of Gazi University, Faculty of Education, 37*(2), 523-558.

Tolles, J., & Meurer, W.J. (2016). Logistic regression: relating patient characteristics to outcomes. *Jama, 316*(5), 533-534.

Turgut, M., & Baykul, Y. (2013). *Eğitimde Ölçme ve Değerlendirme [Measurement and Evaluation in Education]*. Pegem Yayıncılık.

Turhan, K., Kurt, B., & Engin, Y.Z. (2013). Estimation of Student Success with Artificial Neural Networks. Education and Science, *38(170)*, 112-120.

Uzut, O.G., & Buyrukoglu, S., (2020). Prediction of real estate prices with data mining algorithms, *Euroasia Journal of Mathematics, Engineering, Natural and Medical Sciences*, *8*(9), 77-84.

Yamamoto, G. T., & Altun, D. (2020). The Coronavirus and the Rising of Online Education. *Journal of University Research*, 25-34. https://doi.org/10.32329/uad.711110

Yildiz, H.K., Genctav, M., Usta, N., Diri, B., & Amasyalı, M.F. (2007). *A New Feature Extraction Method for Text Classification*. 2007 IEEE 15th Signal Processing and Communications Applications.

Yurtoglu, H. (2005). *Yapay Sinir Ağları Modellemesi ile Öngörü Modellemesi: Bazı Makroekonomik Değişkenler için Türkiye Örneği [Predictive Modeling with Artificial Neural Network Modeling: The Case of Turkey for Some Macroeconomic Variables]*. [Expertise Thesis, DPT]. https://www.sbb.gov.tr/wp-content/uploads/2018/11/HasanYurtoglu.pdf

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 2, 430–450
https://doi.org/10.21449/ijate.958871
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Adaptation of the Children's Perceived Academic Self-Efficacy Scale: Validity and Reliability Study

Neslihan Tugce Ozyeter[1,*], **Omer Kutlu**[2]

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

ARTICLE HISTORY
Received: June 28, 2021
Revised: Jan. 12, 2022
Accepted: Apr. 20, 2022

Keywords:
Academic Self-Efficacy,
Secondary School Students,
Adaptation,
Validity,
Reliability.

Abstract: Academic self-efficacy, which is the belief that the student can achieve an academic task, has a highly strong impact on the academic performance of students. It is known that students with high academic self-efficacy show high academic performance, see the academic difficulties they encounter as areas of development and continue to strive for success. The review of the related literature has identified no scale whose validity and reliability analyses have been carried out by following the necessary scientific procedures that can be used to measure this quality for the 9-13 age group. Therefore, in the current study, an adaptation study of the scale developed by Jinks and Morgan into the Turkish language and culture was performed. The study group consisted of secondary school students, and the data were collected in two separate sessions. Upon completion of the adaptation-based translation of the scale, its degree of validity was calculated based on the linguistic, content, construct and criterion approaches and its degree of reliability was calculated by Cronbach Alpha and Composite Reliability coefficient. The findings show that the adapted scale can be used to obtain valid and reliable results for the 9-13 age group.

1. INTRODUCTION

Various explanations and theories have been put forward to answer the questions about how the organism learns since the 1900s. Due to the inadequacy of the Behavioral Theory to explain all learning of organisms (Tolman, 1948), theorists have started to argue that there are some cognitive processes determining the relationship between stimulus and behavior of an organism (Özdel, 2015). Based on this view, a number of scientists have developed cognitive theories, which advocate the role of cognitive structures in the learning processes. One of these theories is the Social Cognitive Theory, developed by Albert Bandura.

The Social Cognitive Theory, which postulates that the organism learns from its social environment, was first put forward by Rotter. Rotter (1990) stated that the individuals' reactions are not merely instantaneous responses to stimuli, but are shaped by their previous learning, observations and experiences and the results they draw. Social Cognitive Theory was developed by Albert Bandura. Bandura (1976) stated that not only observations and information obtained from others, but also reward and punishment play a role in learning. Accordingly, the first of

*CONTACT: Neslihan Tugce OZYETER simsekneslihantugce@gmail.com Kocaeli University, Faculty of Education, Department of Educational Sciences, Measurement and Evaluation Department, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

the foundations of social learning is that learning is not just behavioral. Learning is a cognitive process that mostly develops in social settings. Learning involves not only observing a behavior, but also observing the consequences of this behavior. This is called a vicarious reinforcer. Based on these first two foundations, observation of a behavior may not be necessary for the learning to take place. In other words, when an organism observes that the result of a certain behavior is negative it does not perform that behavior, and this is a learning process. However, it cannot be observed because there is no behavior. The environment in which the organism is located, its cognition and behavior interact with one another in the learning process.

In addition to mentioned discussions regarding learning and behavior, Bandura brought the concept of self-efficacy, which has been viewed as critical for observation-based learning. Self-efficacy, which is defined as the self-belief of the organism's ability to do a task or job (Bandura, 1977), has an essential role in determining how an individual tackles the difficulties he or she encounters in life, in achieving his/her goals, attempting to perform an action, and having a new experience. According to this theory, the social and cognitive processes of the organism are affected by the observations and experiences of individuals in their social environment. This is closely related to self-efficacy, which is the perception on oneself of the social experiences perceived externally by the person. It is the belief that s/he can have these experiences him/herself. Accordingly, individuals with high self-efficacy in a field are those who believe that they will perform well in that field. They have a strong belief that they can cope with tasks in this particular field but not with those in other fields. When this belief turns into action, the individual will probably achieve it, and this belief will be reinforced (Türkçapar, 2008). Similarly, individuals with low self-efficacy about entering unfamiliar environments will avoid this task due to their low belief that they will fulfill this task, and will not achieve their best performance in this field. Wood and Bandura (1989) define self-efficacy as an individual's belief in his or her capacity in the qualities (motivation, cognitive features, etc.) necessary to achieve certain situational goals. Self-efficacy levels of individuals determine how they feel, think, act and self-motivate. Bandura (2002) states that individuals' self-efficacy beliefs determine their ways of thinking, how they motivate themselves when they are faced with a challenge, and how they make their choices. Individuals with high self-efficacy in a specific area see the difficulties they encounter as obstacles to be overcome. They have a very high motivation and interest in getting over these impediments. People with low self-efficacy see the same obstacles as threats. They are more likely to give up whenever they encounter some. Their interest and motivation are low (Bandura, 1997; Driscoll, 2000).

Social learning theorists argue that processes such as thinking, planning, decision making, and believing have an important role in the learning of the organism (Bayrakçı, 2007). As such, the person's belief that s/he can complete a task becomes important. Yıldız (2014) emphasizes that the individual's perception of self-efficacy is critical in the learning process. In their research, Doğan et al., (2012) found that individuals with positive self-efficacy perceptions want to achieve higher-standard goals, and thus they make much more effort. According to Bandura (1997), whether individuals will be successful in a task or not is not only related to their cognitive capacities. In other words, cognitive skills are indispensable but not sufficient for high academic performance. Students usually know what to do, but they do not make the effort to cope with the difficult processes required by the task (Digiunta et al., 2013). The studies highlight the fact that students who learn to organize their own learning are more effective in making this necessary effort (Zimmerman & Martinez-Pons, 1988), which requires self-efficacy (Digiunta et al., 2013).

Academic self-efficacy is defined as an individual's belief in his or her own capacity to learn or perform an academic task at the targeted level (Schunk & Pajares, 2002). According to Pajares (1996), this belief has a wide range of manifestations. In other words, while a person's belief in

his or her general performance at school reflects his/her academic self-efficacy, his/her belief in his/her capacity to perform only four actions is a part of his/her academic self-efficacy. Students' beliefs about their academic performance stand out in every moment of their academic life. Their belief in themselves (self-efficacy) plays an effective and major role in many areas such as thinking effectively, thinking positively or negatively, how they motivate themselves or show determination when they encounter academic obstacles, and how they regulate their own ideas and behaviors. Consequently, all these processes contribute to the student's performance at school.

There are four sources from which students derive their academic self-efficacy (Bandura, 1997; Pajares, 1996). These are their own past experiences, those of others (indirect experiences), social persuasion, and physical and emotional states. The past experiences of the individual are related to his/her previous performance for a similar academic task. If the individual has achieved a result that s/he thought to be successful in the past, this will increase his/her self-efficacy in that field or task. Similarly, if s/he has had a result that s/he has described as unsuccessful in the past, s/he will exhibit a lower self-efficacy. However, the experience of the individual may be limited or the individual may doubt his/her self-efficacy. In this case, s/he uses the experiences of others as a reference to build his/her own self-efficacy. If the individual develops a similarity between him/herself and the person s/he observes, s/he will be more affected by the results obtained by the person based on this similarity. Social persuasion includes words of encouragement or discouragement that students hear from others. While the student receiving encouraging verbal stimuli develops positive self-efficacy, discouraging stimuli can even weaken his/her strong self-efficacy. Finally, the student's physical and emotional state also play a role in shaping his/her academic self-efficacy. Experiencing depression, one may feel less confident about his/her own skills or feeling physically poor may impact one’s way of thinking of how to deal with the issues. On the other side, the confidence and sense of achievement that s/he feels when his/her task is completed can also strengthen the student's self-efficacy. In summary, the process of creating and using students' academic self-efficacy beliefs is intuitive. They participate in an academic task, interpret their results. By using their interpretations based on these results for similar contexts and tasks at a later time, they form a belief that they can perform a task themselves.

The review of the related literature clearly shows that academic self-efficacy is an important determinant of student performance at all educational levels (Bassi et al., 2006; Doğan, 2005; Ferla et al., 2009; Khan, 2014; Mercer et al., 2011; Zajacova et al., 2005). Studies reveal that there is a positive and significant relationship between academic self-efficacy and academic performance of students from almost every age group.

As far as the available scales of academic self-efficacy are concerned, one concludes that these are mostly adaptations. The adaptation studies measuring the academic self-efficacy were conducted on preservice teachers by Yılmaz, Gürçay and Ekici (2007), on high school students by Kemer (2006), and on university students by Ekici (2012). A scale was adapted by Telef and Karaca (2012) to measure social and emotional self-efficacy, which make up students’ general self-efficacy perceptions. The target group of this scale is adolescents aged 12-19. When all these studies are examined, no scale has been found in Turkey, measuring academic self-efficacy at the secondary school level. However, an international scale adapted to Turkish was identified. When the adaptation study of the scale was reviewed, it can be concluded that the scientific adaptation procedures had not been followed. it is clear that the adaptation process did not come up to standards necessary for a valid and reliable scale as the report provided almost no information about how the study group for adaptation process was selected and what characteristics it covered. Furthermore, the researcher (the adaptor) conducted an explanatory process to re-examine the factor structure of the scale whose factor structure was already

determined. Explanatory studies are suggested in scale development studies. Adapting a scale means if the original scale functions in another culture as well, that's why, researchers are held responsible for the equity and the meaning of the construct between cultures. Another point is the removal of 9 items according to the results of EFA and researcher mentioned nothing about the probable causes and discussions behind it; s/he only presented statistical issues to justify the exclusion of 9 items. Also, there are no information regarding the exclusion process (eg. Which criteria were taken into account, which item was removed in the first place and what the reason was, etc.). The last problem is the very low reliability of the third subscale for the adapted form (0.51). All these reasons lead the researcher to go through the adaptation process by following the standard adaptation steps offered by International Test Commission (ICT) (2017) and Hernandes et al. (2020). Therefore, the current study aimed to adapt the Children's Perceived Academic Self-Efficacy Scale originally developed by Jinks and Morgan (1999) into Turkish, by following the requisite scientific steps indicated in the literature and to introduce this scale, which can measure academic self-efficacy at the secondary school level, to the national literature. The reason why the adaptation of this scale is considered crucial is its vivid relationships with the academic performance. It is thought that by making this scale usable in the national literature, measuring the target trait in younger age groups, and applying support and guidance strategies, if necessary, would be possible, and this can help to increase the academic self-efficacy and therefore academic performance of secondary school students.

2. METHOD

2.1. Study Design

This study is designed as a cross-sectional survey. Survey studies are generally carried out to describe the characteristics (belief, knowledge, attitude, etc.) of a community, and cross-sectionality means one-time data collection on the same group (Fraenkel & Wallen, 2006).

2.2. Study Group

The study group was chosen with maximum variation sampling and consisted of the students from urban and suburban cities, with both lower and higher economic levels as well as various ethnic identities, in accordance with the original study group. The data were collected at two different times using online platforms. The first step aimed to gather evidence for the construct validity of the scale, which was translated based on adaptation. Thus, the scale given to the students in the online environment was filled by a total of 313 students. Table 1 shows the distribution of the pre-tested group by grade levels.

Table 1. *Perceived academic self-efficacy scale study group 1.*

Grade Level	N
5th grade	79
6th grade	80
7th grade	74
8th grade	80
Total	313

As shown in Table 1, the distribution of grade levels is similar. The second step was performed for criterion validity and validity studies based on group differences. Questions were added to the online form about what the students' grades were in the Turkish, Mathematics, Science and Social Studies courses in 2020-2021 Fall semester, and what they think their grades will be in these courses at the end of 2020-2021 Spring term. This tool was administered to a total of 173 secondary school students. 14 students with missing data were excluded from the dataset. In Table 2, the distribution of the data collected in the second application for validity study by

grade levels is given. As Table 2 shows, the distribution of the data collected in the second step by grade levels is close to each other.

Table 2. *Perceived academic self-efficacy scale study group 2.*

Grade Level	N
5th grade	41
6th grade	21
7th grade	57
8th grade	40
Total	159

2.3. Data Collection Instruments

The instrument used in this study to collect data is the adapted form of the Perceived Academic Self-Efficacy Scale.

2.3.1. *Perceived academic self-efficacy scale*

The scale development process suggested by DeVellis (1991) was followed in the development of the Perceived Academic Self-Efficacy Scale developed by Jinks and Morgan (1999), and 53 items were created by the researchers to measure academic self-efficacy. These items were subjected to content validity study in three separate sessions. In the first of the sessions, 5 instructors (academics), in the second, 4 secondary school teachers, and in the last panel, 15 secondary school students gave their assistance. The 53 items were written under four sub-dimensions called ability, effort, task difficulty, and context. In the first and second panels, teachers and instructors (teacher trainers) were asked to place all the items in the item pool into the four predicted sub-dimensions of the scale. After placing each item, they were asked to mark on a scale from 1 (not sure) to 5 (very sure) how sure they were about the decision they made about the sub-dimension. Items with ambiguity in their narration were either rewritten or removed. If the level of agreement of the experts about the sub-dimension in which the item is located was low, it was decided to remove the item even if the sub-dimension in which the item was included was consistent. The third panel was held with a group of 15 secondary school students. Here, students were asked to think aloud about the ease of reading and intelligibility of the items.

The response category of the scale is a 4-point Likert scale. The ranking is performed between the highest level of agreement (1) and the lowest level of agreement (4). In the piloting of the scale, students were asked to write down the grades they received the last semester on the Turkish, Mathematics, Science and Social Studies courses. In the piloting of the scale, 900 usable observations from 3 different schools were obtained.

Items with an item-total correlation of less than .30 were excluded from the exploratory factor analysis results of the scale. A 3-dimensional structure for a total of 30 items was adopted for the scale. The correlations of the total score of the scale and the subscale scores with the grades of the students reported in the Turkish, Mathematics, Science and Social Studies courses were examined. These correlations were found to be moderate and high and statistically significant. Thus, it was concluded that valid results would be obtained with the developed scale.

The Cronbach Alpha reliability coefficients of the scale, which consists of a total of 30 items in three sub-dimensions, are 0.78 for 13 items in the talent sub-dimension, 0.70 for 13 items in the context sub-dimension and 0.66 for 4 items in the effort sub-dimension (Jinks & Morgan, 1999).

2.4. Data Collection

The translation of the scale, which was completed based on the principles of adaptation, was transferred to the online environment by taking into account the layout of the original scale. An instruction on how to complete the form was added to the form, which was sent to be filled by 5th, 6th, 7th and 8th grade students using Google form. The IRB research ethics permission required for the study was obtained from Ankara University. At the end of the scale, a section is reserved for students to note the situations that they have difficulty or do not understand during the administration phase. The piloting step was completed, and the scale was revised with the feedback of the students before the validity analysis of the scale began.

In order to sustain data quality, the online forms were sent to the teachers directly by the authors. The teachers were selected based on the school they work (in terms of the financial status of the families, school neighborhood and type of residence). The authors received information about the general atmosphere of the online classroom for each of the classroom. All teachers were provided with the exactly same instruction for the test to be read loud just before the test to make the data collection process standardized as possible. As teachers had students fill out the forms during instruction hours, it was made sure that students completed them in person. The link of the scale was activated before the courses and deactivated afterwards to restrain multiple entries by the same students. The form was carried out in each classroom only for once to avoid the repetitive entries.

2.5. Data Analysis

Before engaging in the adaptation work, the literature was reviewed. No scale could be identified measuring academic self-efficacy for the 9-13 age group in Turkey or adapted to Turkish by following scientific processes. In addition, the scale developed by Jinks and Morgan (1999) for the targeted age group was examined. Adapting this scale was decided for reasons such as the eliminating the costs of developing a new scale and the difficulty of collecting data during the COVID-19 outbreak. The adaptation of the scale to Turkish culture had been carried out by Öncü in 2002.

Scale adaptation steps are collected under five main headings by ICT (2017) and Hernandez et al. (2020). These are Pre-Adaptation, development process (Test Development), verification process [Confirmation (Empirical Analyses)], implementation (Administration), scale scores and comments (Score Scales and Interpretation), and reporting (Documentation). Pre-adaptation steps are about decisions that need to be made before starting the translation/adaptation process. The development process is the main part of the adaptation work and includes explanations and suggestions about the adaptation process. The verification phase is about collecting empirical evidence about the validity, reliability and comparability of the scale. The next two steps refer to the implementation of the scale and the scoring processes, and reporting refers to the writing of the actions taken. This adaptation study has been reported in detail by selecting the relevant adaptation standards under these headings.

2.5.1. *Pre-adaptation*

The first step in this process is to obtain permission from the developers of the original scale for the scale adaptation study, which was also taken for the current study. The next stage involves a review of the extent to which the structure measured by the scale is compatible (overlaps) with its counterpart in the target culture, and how appropriate the items in the original scale are for the group in the target culture. For this, the expert opinions were obtained. The last step of this phase is to review the details that will make a difference in the measured structure related to the physical characteristics of the scale such as the suitability of the item type to the target culture and age group, the application period of the scale, and the materials used in the

implementation. It was observed that there was no feature in the scale that might cause problems in the target culture.

2.5.2. *Adaptation*

The translation process, one of the most important steps in the adaptation process, is under this heading. This process also stands for the content validation proofs of the scale. The first aim in the adaptation-based translation process is to ensure linguistic, psychological and cultural equivalence. Performing this step meticulously is the prerequisite for the scale to produce valid and reliable results in the target culture. Therefore, choosing the experts who will carry out the translation and adaptation very well is strongly advised. An expert is defined as a person who has sufficient knowledge in the target and source language, target and source culture, the scope of the test and the test process (ITC, 2017). Due to the difficulty of finding experts in all of the four fields specified above, the experts who met the most of the four criteria were selected for this study.

In the first step of the adaptation process, the scale was translated into Turkish by translators who are proficient at the C1 level in both English and Turkish translated the scale into Turkish. 21 items in the article translated and published by Öncü (2012) were added to the translation form. The expert group was asked to examine the translation by considering the structure and age group for items with only one translation; and for the items with two translations, they were asked to choose the better translation, and/or write their own suggestions. The expert group for the translation consisted of an English teacher, a British citizen who has been living in Turkey for 4 years, and an assessment and evaluation specialist with expertise in the source language.

Another step in the test adaptation process is the process of collecting evidence that the target group is familiar with the item type, response category, administration process, and other test-related processes. The original scale is structurally suitable for the target group. The item structure used in the scale is one of the most frequently used item structures. Since the administration of the scale was to be performed online, this was predicted to pose some challenges for the students. Therefore, some trials were run with the students to identify any problems. The last step here was conducting a small-group piloting before proceeding to the actual implementation. At this point, the scale was sent to 10 students before the actual pre-trial. At the end of the implementation, these students were asked about whether the process, the instructions and the expressions in the items were clear and unambiguous, and the necessary adjustments were made accordingly. A sample item for each dimension included in the original scale and the adaptation-based translation is given in Table 3.

Table 3. *Sample items from original scale and adapted scale for each sub-scale.*

Sub-Scale	Original Scale	Adapted Scale
Talent	It is not hard for me to get good grades in school.	Okulda iyi notlar almak benim için zor değil.
Context	No one cares if I do well in school.	Okulda başarılı olmam kimsenin umurunda değil.
Effort	I always get good grades when I try hard.	Çok çalıştığımda her zaman iyi notlar alırım.

An important change made in the adaptation process concerns response categories. On the original scale, 1 represents the highest agreement and 4 the lowest, which drew the attention of the experts. The review of the Likert-type scales developed for Turkish secondary school students (Bakırtaş & Tonga, 2016; Can & Topçuoğlu Ünal, 2017; Karakuş Tayşi & Özbay, 2016; Karakuş Tayşi & Taşkın, 2018; Kukul et al., 2017; Mete, 2021;) revealed that the response categories were generally arranged from negative to positive. Thus, based on the literature review and the expert opinions indicating that Turkish students were accustomed to this format, the response categories of the scale were revised. In other words, the scale was revised to begin with 1 “Strongly disagree” and end with 4 “Strongly agree” (See Appendix).

2.5.3. *Verification*

This stage involves the pre-trial administration of the translated scale and the calculation of some of its psychometric properties. Empirical evidence at this stage can prove that the scale is usable in the target culture. As already mentioned in the study group, the socio-economic levels and the location they live were taken into consideration. The size of the study group was deemed important because only then can the evidence of proofs of validity and reliability be obtained when the group is large enough.

Linguistics validity, content validity, construct validity and the criterion validity evidences of the scale were investigated for verification process. Linguistic validity evidences were gathered based on experts' opinions for whether the translated items cover the original meaning. After the item editing was completed, the percentage of agreement calculation process developed by Davis (1992) was planned to calculate the agreement among experts. The process, whose main function is to determine the content validity index, was used in this study to calculate the expert agreement between the English items and the adapted Turkish items. While calculating the rate of agreement among experts, the agreed items are summed up and divided by the total number of items. Whereas inter-expert agreement 1 (translation not appropriate) and 2 (translation should be seriously reviewed) denote to non-use, 3 (translation should be slightly revised) and 4 (translation appropriate) indicate that the translation is usable. The ratio of the items with observed agreement to the total number of items yields the measurement of agreement among experts.

Content validity is studied if the adapted items are thought to measure the academic self-efficacy in Turkish language and culture. For this aim, an expert group was formed to evalate the appropriateness of the items with the structure in Turkish culture. In other words, they assessed if the items could measure the academic self-efficacy in Turkish context. The expert group consisted of an academician who has a doctoral degree in psychology and lived in England for 6 years, and a Turkish teacher working in a secondary school.

To conduct construct validity, confirmatory factor analysis and testing group differences were carried out. The confirmatory factor analyses are carried out to test whether the structure intended to be measured with the adapted scale works in the target culture (Kline, 2005) in adaptation studies.

Before proceeding to the analysis, nine items that needed to be reverse coded were done so. There was no missing data in the dataset. Some assumptions had to be met to be able to perform the confirmatory factor analysis. Harrington (2009) and Kline (2005) suggest examining the dataset in terms of univariate normality, univariate outlier, multivariate normality and multivariate outlier. The univariate normality and univariate outlier analysis were performed based on the kurtosis and skewness values, and the z standard scores. The calculations based on the Mahalanobis distance and residuals revealed multivariate outlier and multivariate normality. As a result, univariate and multivariate outliers were found in the dataset. The clusters in certain categories in the student responses were identified as the reason for these outliers. These are the natural reactions of the students. Therefore, this distribution of the target trait was considered as the reflection of the natural distribution of the target trait. The univariate normality values are given in Table 4.

Table 4. *Descriptive statistics of items for perceivedacademic self-efficacy scale.*

Items	Kurtosis (s.e.)	Skewness (s.e.)	Items	Kurtosis (s.e.)	Skewness (s.e.)
m1	1.33(0.27)	-1.28(0.14)	m16	0.08	-0.88
m2	4.79	-2.09	m17	22.70	-4.65
m3	-1.13	-0.23	m18	0.39	-0.98
m4	-1.29	-0.48	m19	-0.98	-0.22
m5	-0.53	-0.69	m20	8.19	-3.07
m6	1.89	-1.57	m21	2.67	-1.71
m7	0.86	-1.60	m22	-0.97	-0.64
m8	3.65	-2.04	m23	4.26	-2.32
m9	5.99	-2.47	m24	-0.51	-1.02
m10	0.43	-0.99	m25	1.18	-1.31
m11	0.63	-1.04	m26	0.15	-0.87
m12	.071	-1.31	m27	1.70	-1.45
m13	7.92	-2.95	m28	13.60	-3.84
m14	0.13	-0.88	m29	0.85	-1.42
m15	-0.97	-0.73	m30	-.45	-0.45

When the kurtosis and skewness values given in Table 4 were examined, it was found that some items had sharp and skewed item score distributions, which was taken into consideration while determining the estimation method.

The confirmatory factor analysis was carried out using the Mplus 7 program. Before the confirmatory factor analysis began, the related literature was reviewed to identify the estimation method that would be the best fit for the descriptive features and scale levels of the items. Rhemtulla, Brosseau-Liard, and Savalei (2012) demonstrated the importance of the estimation method to be used in the analysis to ensure model-data fit and to estimate standard errors and factor loads in an unbiased way. Accordingly, the WLSMV estimation method produced more unbiased results in the non-normally distributed simulative dataset with fewer than 5 response categories. Newsom (2017), on the other hand, states that with the Likert scale, measurements with 7 or more response categories can be estimated continuously, and the others should be analyzed sequentially. Thus, the estimation method used in the CFA process of the scale, which was developed as a 4-point Likert scale, was decided to be the WLSMV.

Another evidence for construct validity is the studies based on group differences (Crocker & Algina, 2008). In studies based on group differences, the significance of the difference between the mean scores obtained from the scale of the groups that are expected to differ in terms of the measured feature is tested. Finding a statistically significant difference indicates that the scores obtained from the scale can identify individuals with and without the trait of interest.

Considering the relationships between students' academic achievement and self-efficacy and the theoretical structure, students with high self-efficacy are expected to have high grades on their report card and a high perception of the grades they will get, whereas the students with low self-efficacy are expected to have low grades and similarly low expectations for the grades they will get. First of all, the students were ranked from the lowest to the highest according to their total scale scores. The listed observations were examined in terms of the assumptions of the analysis technique. The SPSS 22.0 program was used in the analysis. At this point, the data set was analyzed in terms of univariate normality and univariate outliers. Univariate normality analysis was performed based on the kurtosis and skewness coefficient, and univariate extreme value (outlier) analysis was performed using standard z scores. Univariate outliers were

removed from the data set. The analysis was performed with 149 observations. Two groups with low and high academic self-efficacy were formed as lower and upper 27% (n_{lower}=40). Here, it is expected that the average score of the students in the lower and upper groups, and the average of the report card grades they hope to get at the end of the current semester will differ significantly. The kurtosis and skewness coefficients for examining the distributions of the dependent variables of these analyses in the categories of the independent variable are given in Table 5.

Table 5. *Kurtosis and skewness values of lower and upper 27% groups.*

Lessons	Groups	Kurtosis	Skewness
1.Turkish (1)	Lower	-1.04	-0.46
	Upper	10.22	-3.02
2.Mathematics (1)	Lower	-0.89	-0.50
	Upper	7.36	-2.67
3.Science Studies (1)	Lower	-1.22	-0.30
	Upper	14.88	-3.47
4. Social Studies (1)	Lower	-0.77	-0.78
	Upper	7.68	-2.69
5.Turkish (2)	Lower	-0.67	-0.66
	Upper	8.85	-2.88
6.Mathematics (2)	Lower	-1.16	-0.22
	Upper	5.80	-2.31
7. Science Studies (2)	Lower	-0.49	-0.74
	Upper	6.51	-2.36
8. Social Studies (2)	Lower	-1.22	-0.43
	Upper	3.84	-1.84

(1) 2020-2021 Fall
(2) 2020-2021 Spring

Based on the kurtosis and skewness values, it can be said that the levels of the independent variable do not have a normal distribution. The Levene test results regarding the homogeneity of variances are given in Table 6.

Table 6. *Levene test results.*

Lessons	F	*p*
1. Turkish (1)	27.44	0.00
2. Mathematics (1)	28.98	0.00
3. Science Studies (1)	38.07	0.00
4. Social Studies (1)	49.05	0.00
5. Turkish (2)	17.99	0.00
6. Mathematics (2)	19.71	0.00
7. Science Studies (2)	4.30	0.04
8. Social Studies (2)	42.95	0.00

As shown in Table 6, the assumption of homogeneity of variances is violated. The Mann Whitney U test, which is used to test the significance of the difference between the two means from the nonparametric tests, was used.

Criterion validity- another critical evidence for validating a scale- was performed. Criterion-based validity is achieved by examining whether the measurement tool has predictive power for the feature of interest (Bilican Demir, 2017). Thanks to this process, the compatibility of the developed/adapted scale with a number of other related measures is determined. High correlation between variables means high criterion-based validity, and low correlation means

low validity. The most important issue regarding criterion-based validity is the determination of the criterion measure. The relations of the selected criterion with the structure whose validity is tested must be theoretically and empirically proven (Crocker & Algina, 2008).

For their own criterion validity analysis of the original perceived academic self-efficacy scale, Jinks and Morgan (1999) asked students to indicate their last grades in four core subjects (reading, mathematics, science and social studies) because the students with high academic self-efficacy were expected to have high grades in these courses. Accordingly, in this adaptation study, the original form was adhered to. The theoretical and empirical relationships between self-efficacy and achievement and perception of success are known as well (Ayotola & Adedeji, 2009; Lee et al., 2014; Sebaee et al., 2017; Tenaw, 2013; Wilcox & Nordtokke, 2019). The scale, whose verification analysis was completed, was revised online to obtain information about the grades of the students and a total of 8 questions were added. For concurrent validity, the validity of the criterion measure needs to be obtained recently through the scale under examination. These questions are about the students' grades in Turkish, Mathematics, Science and Social Studies courses in 2020-2021 Fall semester and what they think their scores will be in these courses at the end of the 2020-2021 Spring term. The questions are presented below:

What was your grade for the Turkish class on your report card last semester?
What was your grade for the Mathematic class on your report card last semester?
What was your grade for the Social Studies class on your report card last semester?
What was your grade for the Science class on your report card last semester?
What do you think your grade for the Turkish class on your report card will be at the end of this semester?
What do you think your grade for the Mathematic class on your report card will be at the end of this semester?
What do you think your grade for the Social Studies class on your report card will be at the end of this semester?
What do you think your grade for the Science class on your report card will be at the end of this semester?

Differing from the original study, the students participating in the study were asked about their final report card scores as well as those which they would receive at the end of the current semester. The reason for this is that the data collection phase of the adaptation study was carried out during the COVID-19 outbreak. It was thought that during the pandemic, school education could not be systematically continued in either face-to-face or online environment, so the report card grades given to the students may not have been the scores that reflect their actual performances. In addition, the report cards they expect at the end of the current term reveal how confident they are about their success in school lessons. For these reasons, the students were asked to write both the most recent report card grades and the grade they will receive in the four basic learning areas. The correlations between the grades reported by the students and the total score, ability subscale score, context subscale score, and effort sub-score on the perceived academic self-efficacy scale under study were analyzed. A total of 173 students were included in the study group. The number of students excluding missing data is 159. The analyses were performed with the SPSS 22.0 program based on 159 participants. The normal distributions of the variables were tested before proceeding to the correlation calculation. Outliers were detected in some variables. When these values were examined, it was concluded that these values existed due to the nature of the collected data. For example, students stated that their last report card grade in mathematics was between 40 and 50 points. Since this value was far from the distribution, it was determined as an outlier. Since removing these values from the observation set would damage the nature of the collected data, no observations were excluded

from the dataset. Spearman Brown Rank Differences Correlation Coefficient, which is used in cases where the distribution does not show a normal distribution, was calculated.

For the reliability studies, Cronbach Alpha Coefficient and composite reliability were calculated for the total score as well as subscale scores. The study of the original scale was taken into consideration and calculations were made based on the original scale.

3. RESULT

The findings of the validity studies for the Perceived Academic Self-Efficacy Scale are given below. First, the confirmatory factor analysis was performed for the construct validity of the scale.

3.1. Linguistic Validity

In the adaptation-based translation, the adapted form was sent to three experts in English, who lived in the United States and whose native language was Turkish. Based on the feedback from the experts, three experts were cross-paired, and the percentage of agreement was calculated for each expert group of two. Accordingly, the percentage of agreement among experts varied between 77% and 97%. If this ratio is above 0.70, it is an indication of agreement between experts (Davis, 1992).

3.2. Content Validity

To collect proof of content validity of the scale, the translated form was sent to the experts, who were asked to evaluate it in terms of the presentability of the construct in Turkish language and culture as well as items' power to measure the academic self-efficacy. One Turkish language teacher who works at the secondary school level was intentionally invited to the expert group to see if the language is appropriate for those/students between 9-13; whether there are any statements or words students might not know or misunderstand. Besides, one academician whose study field covers the self-efficacy construct and who has doctoral degree in the field of psychology was asked to assess whether the items can measure the target trait or not.

The suggestions from the experts were examined by the researcher, it was observed that most of the feedback was regarding the wording of the items. Expert group concluded that the translated form can successfully measure and evaluate the academic self-efficacy of secondary school children, and the expressions are appropriate for the age group. The revision of the adapted items was completed in line with expert opinions.

3.3. The Confirmatory Factor Analysis

The very first construct validity evidence of the adapted scale was collected with confirmatory factor analysis. When confirmatory factor analysis was run, the model worked even though the data fit with the model was low. The modification indices suggested by the program were taken into consideration. While making modifications, as suggested by Hayduk (1989), improving the model was a priority; however, maintaining compatibility with the theoretical model is still a must. Thus, before the covariances between the observed variables suggested by the model were drawn, the meanings of the items and their connotations for the students were considered. The first pair of items that were related to each other were the items in the second sub-dimension, namely the context sub-dimension. Both items can be accepted as an indicator of the importance that the student attaches to high school education, and whether he/she wants to complete this level of education. The other items in the first sub-dimension, the talent sub-dimension, reflect the students' beliefs that they are good students, even in different courses.

The model that was run after the modifications is presented in Figure 1. The recommendations made by Hooper et al. (2008) were taken into account when deciding which indices to report on model fit, as the authors recommend reporting at least one index from each of the absolute fit indices, incremental fit indices and parsimony-adjusted indices categories. The absolute fit

indices are the most basic indicators that show how well the proposed theoretical structure fits the data. Thus, the absolute fit indices χ^2/df and RMSEA were reported. The incremental fit indices show if the established model fits better than the null model. The CFI and TLI were reported from indices in this group. The final group, the parsimony-adjusted fit indices, are used to decide which model is more useful when different models are built on the same dataset (Hooper et al., 2008). Reporting this final group of fit indices is irrelevant here, as the current study aims to confirm a predicted theoretical structure. The goodness-of-fit indices for the model are presented in Table 7.

Figure 1. *Confirmatory factor analysis model of the perceived academic self-efficacy scale.*

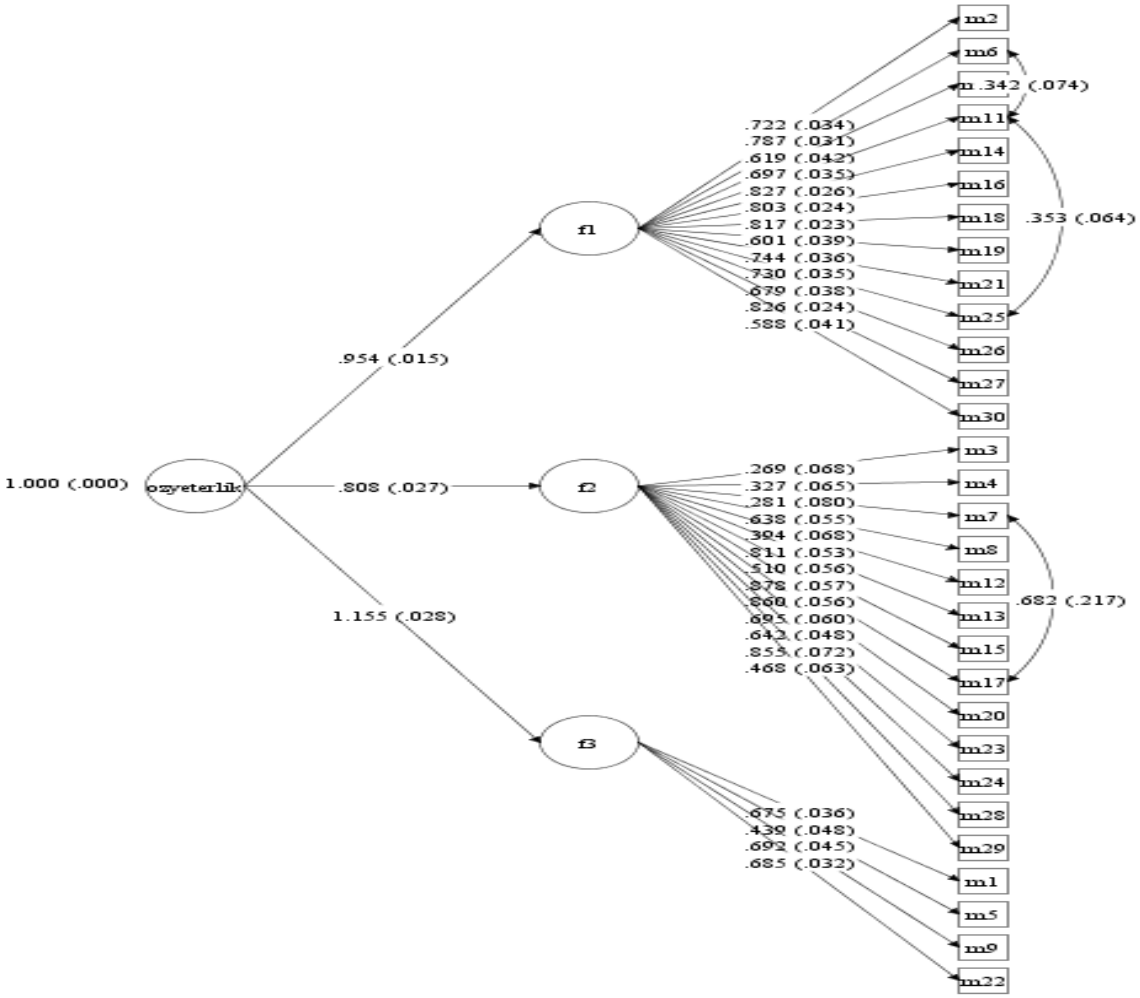


Table 7. *Goodness of fit indices.*

Goodness of Fit	Criteria	Model Value
x^2/df	<2; <5	2.33
RMSEA	<0.05 0.05-0.08	0.06
RMSEA 90% CI	-	0.06-0.07
CFI	>0.95 >0.90	0.90
TLI	>0.95 >0.90	0.90

The relationships between the sub-dimensions are presented in Table 8.

Table 8. *Correlations between sub-dimensions of perceived academic self-efficacy scale.*

	1-2	1-3	2-3
Correlations Between Sub-Dimensions	0.63	0.80	0.65

In Table 9, the item-total correlations and R^2 values related to the model are given.

Table 9. *Item-total correlations and R^2 values.*

Items	Item-Total Correlations	R^2	Items	Item-Total Correlations	R^2
1	0.62**	0.45	16	0.65**	0.65
2	0.50**	0.52	17	0.37**	0.77
3	0.28**	0.07	18	0.70**	0.67
4	0.37**	0.11	19	0.58**	0.36
5	0.46**	0.19	20	0.44**	0.74
6	0.61**	0.62	21	0.55**	0.55
7	0.21**	0.08	22	0.69**	0.47
8	0.41**	0.41	23	0.42**	0.48
9	0.50**	0.48	24	0.51**	0.41
10	0.52**	0.38	25	0.58**	0.53
11	0.59**	0.49	26	0.59**	0.46
12	0.32**	0.15	27	0.65**	0.68
13	0.43**	0.66	28	0.33**	0.73
14	0.72**	0.69	29	0.36**	0.22
15	0.43**	0.26	30	0.50**	0.35

** 0.01

Table 8 reveals that there is a moderate and high level of correlation between the subscales. According to Figure 1, most of the factor loading were around 0.50 and above, as high as recommended by Hair et al. (2014). Item-total correlations and R^2 values given in Table 9 reveal that items are mostly moderately correlated with the total score.

3.4. Criterion Validity Study

For the criterion validity, Spearman Brown Rank Differences Correlation Coefficient between the students' scores of perceived academic self-efficacy, and their reported grades were calculated. Table 10 presents the correlations between students' scores and the scale total and subscale scores.

Table 10 demonstrates that all correlation coefficients are significant at the level of .05 or .01. The context subscale and the predicted report card grades for social studies and science classes are lowly correlated. All other relationships are moderate or high.

Table 10. *Correlations between students' grades and perceived self-efficacy scale/sub-scales scores.*

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
1. Turkish (1)	1											
2. Mathematics (1)	0.74**	1										
3. Science Studies (1)	0.76**	0.74**	1									
4. Social Studies (1)	0.80**	0.76**	0.74**	1								
5. Turkish (2)	0.50**	0.51**	0.38**	0.45**	1							
6. Mathematics (2)	0.46**	0.63**	0.50**	0.49**	0.77**	1						
7. Science Studies (2)	0.30**	0.38**	0.35**	0.27**	0.72**	0.73**	1					
8. Social Studies (2)	0.36**	0.38**	0.29**	0.47**	0.68**	0.67**	0.67**	1				
9. Academic Self-efficacy total score	0.50**	0.53**	0.53**	0.46**	0.50**	0.57**	0.44**	0.42**	1			
10. Talent Sub-Dimension	0.51**	0.54**	0.50**	0.46**	0.51**	0.60**	0.51**	0.49**	0.92**	1		
11. Context Sub-Dimension	0.32**	0.31**	0.35**	0.29**	0.30**	0.31**	0.17**	0.17**	0.75**	0.48**	1	
12. Effort Sub-Dimension	0.40**	0.42**	0.41**	0.33**	0.43**	0.53**	0.42**	0.37**	0.84**	0.76**	0.53**	1

*0.05
**0.01
(1) 2020-2021 Fall
(2) 2020-2021 Spring

3.5. Validity Analysis Based on Group Differences

To test the group differences, upper and lower 27% of the students were determined. These groups were used for the analysis. Mann Whitney U test was carried out, and the results are given in Table 11.

Table 11. *Mann Whitney U results.*

Lessons	Groups	N	Mean Rank	Sum of Ranks	U	*p*
1. Turkish (1)	Lower	40	27.53	1101.00	281.00	0.00
	Upper	40	53.48	2139.00		
2. Mathematics (1)	Lower	40	26.44	1057.50	237.50	0.00
	Upper	40	54.56	2182.50		
3. Science Studies (1)	Lower	40	26.25	1050.00	230.00	0.00
	Upper	40	54.75	2190.00		
4. Social Studies (1)	Lower	40	29.85	1194.00	374.00	0.00
	Upper	40	51.15	2046.00		
5. Turkish (2)	Lower	40	26.63	1065.00	245.00	0.00
	Upper	40	54.38	2175.00		
6. Mathematics (2)	Lower	40	25.10	1004.00	184.00	0.00
	Upper	40	55.90	2236.00		
7. Science Studies (2)	Lower	40	27.10	1084.00	264.00	0.00
	Upper	40	53.90	2156.00		
8. Social Studies (2)	Lower	40	29.90	1196.00	376.00	0.00
	Upper	40	51.10	2044.00		

(1) 2020-2021 Fall
(2) 2020-2021 Spring

A closer look at Table 11 suggests a significant difference between high and low academic self-efficacy in all learning areas covered in the study (p<0.01). This significant difference in all

learning domains favors students with high academic self-efficacy. In other words, both the report card grades and the end-of-semester grades they think they will get in the Turkish course are higher for students with high academic self-efficacy and lower for students with low academic self-efficacy. This result is also valid for the Mathematics, Science and Social Studies lessons.

3.6. Reliability Analysis of the Perceived Academic Self-Efficacy Scale

Cronbach Alpha and composite reliability coefficients were calculated after determining the validity of the scale. The Cronbach Alpha reliability coefficients of the subscales of the perceived academic self-efficacy scale were 0.91,0.71 and 0.61, respectively. The Cronbach Alpha calculated for the total scale score was 0.91 and the composite reliability was 0.96.

Similar results were obtained for the application-based reliability estimations made for the criterion validity. The Cronbach alpha reliability coefficient was 0.91 for the talent sub-dimension, 0.82 for the context sub-dimension, and 0.58 for the effort sub-dimension. The Cronbach Alpha calculated for the whole scale was 0.92. Reliability estimations of the adapted scale were compatible with the original scale study. The reliability of the results obtained by implementing the scale was high.

4. DISCUSSION and CONCLUSION

In the present study of adapting the perceived academic self-efficacy scale to the Turkish language and culture, the relevant international standards were followed. The adaptation procedure was completed in accordance with these standards and by adhering to the procedure applied for the development of the original scale. Evidence of validity of the scale was collected using linguistic validity, content validity, construct validity, and criterion validity. In the adaptation-based translation, for linguistic validity, the percentage of agreement between experts was calculated to determine the consensus of experts on the translations. The agreement among experts on linguistic quality of the original and the adapted form is excellent, ranging from 77% to 97%. As regards to the content validity, Turkish form was sent to the expert group to be evaluated in terms of its capacity to measure the academic-self efficacy and appropriateness of the expressions of the items to the age group. Feedback was received and based on it changes were made; it was concluded that the translated items could measure academic self-efficacy. The construct validity of the adapted scale was analyzed based on confirmatory factor analysis and group mean differences.

The confirmatory factor analysis for the construct validity proved that the model and the data fit well (χ^2/df=2.33; RMSEA=0.06; CFI=0.90; TLI=0.90). The model presented in Figure 1 highlights that a path coefficient takes a value higher than 1. Jöreskog (1999) states that this is normal, that it is a regression coefficient, not a correlation. He states that the established model and the data should be checked for multicollinearity or negative residuals, most likely the researcher will notice something in these reviews, but it does not mean that there is a definite problem. The proposed checks were performed by the researcher, revealing no problems. Examining the factor loads revealed that most of them were above the 0.50 and R^2 values were higher than 0.40 mostly recommended by Hair et al. (2014). Items with lower R^2 mean they contribute to the common variance less. When investigated, the lowest R2 values are for item 3 and item 7. When the items were reviewed, it was observed that item 3 consisted of two statements and there were doubts about item 7, which might have a different meaning in Turkey than it has in the original form. For item 3, students may be confused to decide which statement to rank. For item 7, graduating from college may not match with Turkey context as it is not a corner stone as it is in Turkey. These are why the variance these items contribute to may be less than other items. However, due to the fact that this is an adaptation process, there may be items that do not work as well as they do in the original scale. This does not necessarily mean that the

contruct has not been validated for the Turkish culture. The fact that the standard estimations are statistically significant and apply in the adapted culture indicates that the scale is validated for the target culture. The correlations between subscales are moderate to high. This is also a sign that these subscales can measure the target trait. The item-total score correlations are observed to be positive, mostly moderate and statistically significant. The original structure was also confirmed for the Turkish culture.

The grade point averages of individuals with low self-efficacy in four basic learning areas were significantly lower than the averages of individuals with high self-efficacy. This finding shows that students who are more successful in their courses and those who think that they are more successful than others have higher academic self-efficacy. It was concluded that the scores obtained from the adapted scale indicated success at determining individuals with low and high academic self-efficacy. This finding is also in line with the available findings of the literature (Bassi et al., 2006; Ferla et al., 2009; Khan, 2014; Mercer et al., 2011).

For the analysis of the criterion validity, the correlations between the students' final report card grades and the expected report card grades and their scores on the adapted academic self-efficacy scale were taken into consideration and all the correlations were found to be significant. When the original scale study was examined, it was concluded that all the relationships were moderately or highly correlated. Therefore, it can be concluded that the adaptation study yielded similar results with the one including the original scale. The low level of correlation may be due to the different sample sizes in the studies. In any case, the relationships are significant, and the criterion validity of the adapted scale is high.

Cronbach Alpha and Composite Reliability were calculated to determine the reliability of the scale and the reliability estimations were found to be moderate to high. The reason for the low reliability coefficient calculated for the third subscale is the number of the items in the subscale. It is evident that reliability coefficients are closely linked to the number of the items. When the original scale was examined, it was detected that the reliability coefficient of the third subscale was calculated as low as well (Jinks & Morgan, 1999). The scale can be safely used for determining the perceived academic self-efficacy of students between 9 and 13.

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/ Social Science Institution, 30-3-2020/54.

Authorship Contribution Statement

Neslihan Tugce Ozyeter: Investigation, Resources, Methodology, Visualization, Software, Formal Analysis, and Writing -original draft. **Omer Kutlu**: Investigation, Resources, Methodology, Formal Analysis, and Writing -original draft.

Orcid

Neslihan Tugce OZYETER https://orcid.org/0000-0003-1558-1293
Omer KUTLU https://orcid.org/0000-0003-4364-5629

REFERENCES

Ayotola, A., & Adedeji, T. (2009). The relationship between mathematics self-efficacy and achievement in mathematics. *Procedia-Social and Behavioral Sciences, 1*(1), 953-957. https://doi:10.1016/j.sbspro.2009.01.169

Bakirtas, H., & Tonga, D. (2016). Development of want-need perception scale for primary and secondary school students. *Online Submission, 15*(4), 1436-1449. https://doi.org/10.17051/io.2016.08346

Bandura A. (1976). *Social learning theory.* Prentice-Hall, Inc.
Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychology Review 84(*2), 191-215. https://doi.org/10.1037/0033-295X.84.2.191
Bandura, A. (1997). *Self-efficacy: The exercise of control.* Freeman.
Bandura, A. (2002). Social cognitive theory in cultural context. *Applied Psychology: An International Review, 51*(2), 269-290. https://doi.org/10.1111/1464-0597.0009277
Bassi, M., Steca, P., Delle Fave, A., & Caprara, G.V. (2007). Academic self-efficacy beliefs and quality of experience in learning. *Journal of Youth and Adolescence, 36*(3), 301-312. https://doi.org/10.1007/s10964-006-9069-y
Bayrakçı, M. (2007). Social learning theory and its educational applications. *Sakarya University Journal of Education Faculty, 14,* 198-210.
Bilican Demir, S. (2017). Ölçmede geçerlik [Validity in measurement process]. In R.N. Demirtaşlı (Ed) *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education]* (4. Ed, pp 57-76). Anı Yayıncılık.
Can, E., & Topçuoğlu Ünal, F. (2017). Attitude scale towards writing for secondary school students: the study of validity and reliability. *International Journal of Languages' Education and Teaching, 5*(3), 203-212. https://doi.org/10.18298/ijlet.2026
Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.
Davis, K.A. (1992). Validity and reliability in qualitative research on second language acquisition and teaching: Another researcher comments. *Tesol Quarterly, 26*(3), 605-608. https://doi.org/10.2307/3587190
DeVellis, R.F. (1991). *Scale development: Theory and applications*. Sage.
Di Giunta, L., Alessandri, G., Gerbino, M., Kanacri, P.L., Zuffiano, A., & Caprara, G.V. (2013). The determinants of scholastic achievement: The contribution of personality traits, self-esteem, and academic self-efficacy. *Learning and Individual Differences, 27*, 102-108. https://doi.org/10.1016/j.lindif.2013.07.006
Dogan, U. (2015). Student engagement, academic self-efficacy, and academic motivation as predictors of academic performance. *The Anthropologist, 20*(3), 553-561. https://doi.org/10.1080/09720073.2015.11891759
Doğan, N., Beyaztaş, D.İ., & Koçak, Z. (2012). The Effect of self-efficacy level of students on their achievement in terms of their grade levels and gender in a social studies class: The case of Erzurum. *Education and Science, 37*(165). https://doi.org/10.1080/09720073.2015.11891759
Driscoll, M.P. (2000). *Psychology of learning for instruction*. Allyn and Bacon.
Ekici, G. (2012). Academic self-efficacy scale: The study of adaptation to Turkish, validity and reliability, *Hacettepe University Journal of Education, 43*, 174-185. http://www.efdergi.hacettepe.edu.tr/yonetim/icerik/makaleler/360-published.pdf
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
Fraenkel, J.R. & Wallen, N.E. (2006). *How to design and evaluate research in education* (6. Ed). McGraw-Hill International.
Hair, J.F., Black, W.C., Babin, B.J., & Anderson, R.E. (2014). *Multivariate data analysis* (7th edition). Pearson Education Limited.
Harrington, D. (2009). *Confirmatory factor analysis.* Oxford University Press.
Hayduk, L.A. (1989). *Structural equation modeling: Essentials and advances.* The John Hopkins University Press.

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

Hooper, D, Coughlan, J. & Mullen, M (2008) Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods, 6*(1), 53-60. https://doi.org/10.21427/D7CF7R

International Test Commission (2017). The ITC guidelines for translating and adapting tests. *International Journal of Testing, 18*(2), 101-134. https://doi:10.1080/15305058.2017.1398166

Jinks, J., & Morgan, V. (1999). Children's perceived academic self-efficacy: An inventory scale. *The Clearing House, 72*(4), 224-230. https://doi.org/10.1080/00098659909599398

Jöreskog, K.G. (1999). How large can a standardized coefficient be? http://www.statmodel.com/download/Joreskog.pdf.

Karakuş Tayşi, E.K., & Taşkın, Y. (2018). Development of the writing anxiety scale for secondary school students: Reliability and validity study. *International Journal of Turkish Literature Culture Education, 7*(2), 1172-1189. http://dx.doi.org/10.7884/teke.4169

Karakuş Tayşi, E., & Özbay, M. (2016). The development of listening attitude scale for secondary school students: Study on the validity and reliability. *Journal of Mother Tongue Education, 4*(2), 187-199.

Kemer, G. (2006). The role of self-efficacy, hope, and anxiety in predicting university entrance examination scores of eleventh grade students. [Unpublished master thesis, Ortadoğu Technical University]. https://open.metu.edu.tr/handle/11511/16273

Khan, M. (2013). Academic self-efficacy, coping, and academic performance in college. *International Journal of Undergraduate Research and Creative Activities, 5*(1), 4. https://doi.org/10.7710/2168-0620.1006

Kline, R.B. (2005). *Principles and practice of structural equation modeling.* The Guilford Press.

Kukul, V., Gökçearslan, Ş., & Günbatar, M.S. (2017). Computer programming self-efficacy scale (CPSES) for secondary school students: Development, validation and reliability. Educational *Technology Theory and Practice, 7*(1), 158-179. https://doi.org/10.17943/etku.288493

Lee, W., Lee, M.J., & Bong, M. (2014). Testing interest and self-efficacy as predictors of academic self-regulation and achievement. *Contemporary Educational Psychology, 39*(2), 86-99. https://doi.org/10.1016/j.cedpsych.2014.02.002

Mercer, S.H., Nellis, L.M., Martínez, R.S., & Kirk, M. (2011). Supporting the students most in need: Academic self-efficacy and perceived teacher support in relation to within-year academic growth. *Journal of School Psychology, 49*(3), 323-338. https://doi.org/10.1016/j.jsp.2011.03.006

Mete, G. (2021). Developing the 21st century skills scale for secondary school students: A validity and reliability study. *The Journal of Social Sciences, 51*, 196-208. http://doi.org/10.29228/SOBIDER.50754

Newsom, J.T. (2017). Structural models for binary repeated measures: Linking modern longitudinal structural equation models to conventional categorical data analysis for matched pairs. *Structural equation modeling: a multidisciplinary journal, 24*(4), 626-635. https://doi.org/10.1080/10705511.2016.1276837

Öncü, H. (2012). Adaptation of the academic self-efficacy scale to Turkish. *Journal of Kirsehir Education Faculty, 13*(1), 183-206. https://www.academia.edu/5815000/Adaptation_of_Academic_Self_Efficacy_Scale_into_Turkish

Özdel, K. (2015). From past to present cognitive behavioral therapies: Theory and practice. *Türkiye Clinics Journal Psychiatry-Special Topics, 8*(2), 10– 20.

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research, 66,* 543-578. https://doi.org/10.3102/00346543066004543

Rhemtulla, M., Brosseau-Liard, P.É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354-373. https://doi.org/10.1037/a0029315

Rotter, J.B. (1990). Internal versus external control of reinforcement: A case history of a variable. *American Psychologist, 45*(4), 489-493. https://doi.org/10.1037/0003-066X.45.4.489

Schunk, D.H., & Pajares, F. (2002). The development of academic self-efficacy. In A. Wigfield & J.S. Eccles (Eds) *Development of achievement motivation* (pp. 15-31). Academic Press.

Sebaee, H.A.A., Aziz, E.A.A., & Aziz, N.T. (2017). Relationship between nursing students' clinical placement satisfaction, academic self-efficacy and achievement. *IOSR Journal of Nursing and Health Science, 6*(2), 101-11.

Telef, B.B., & Karaca, R. (2012). The self-efficacy scale for children: A validity and reliability study. *Buca Faculty of Education Journal, 32*, 169-187.

Tenaw, Y.A. (2013). Relationship between self-efficacy, academic achievement and gender in analytical chemistry at Debre Markos College of teacher education. *African Journal of Chemical Education, 3*(1). https://www.ajol.info/index.php/ajce/article/view/84850

Tolman, E.C. (1948). Cognitive maps in rats and men. *Psychological Review, 55*(4), 189-208. https://doi.org/10.1037/h0061626

Türkçapar H. (2008). *Bilişsel terapi: Temel ilkeler ve uygulama* (3. basım) [Cognitive therapy: Basic principles and applications (3rd ed.)]. HYB Yayıncılık.

Wilcox, G., & Nordstokke, D. (2019). Predictors of university student satisfaction with life, academic self-efficacy, and achievement in the first year. *Canadian Journal of Higher Education, 49*(1), 104–124. https://doi.org/10.7202/1060826ar

Wood, R.E., & Bandura, A. (1989). Impact of conceptions of ability on self-regulatory mechanisms and complex decision making. *Journal of Personality and Social Psychology, 56*, 407-415. https://doi.org/10.1037//0022-3514.56.3.407

Yıldız, Z. (2014). Social cognitive learning theory and religious education. *Review of the Faculty of Divinity, University of Süleyman Demirel, 2*(33), 147-161. https://dergipark.org.tr/en/download/article-file/794249

Yılmaz, M., Gürçay, D., & Ekici, G. (2007). Adaptation of the academic self-efficacy scale to Turkish. *Hacettepe University Journal of Education, 33*. http://www.efdergi.hacettepe.edu.tr/yonetim/icerik/makaleler/1047-published.pdf

Zajacova, A., Lynch, S.M., & Espenshade, T.J. (2005). Self-efficacy, stress, and academic success in college. *Research in Higher Education, 46*(6), 677-706. https://doi.org/10.1007/s11162-004-4139-z

Zimmerman, B.J., & Martinez-Pons, M. (1988). Construct validation of a strategy model of student self-regulated learning. *Journal of Educational Psychology, 80*(3), 284-290. http://doi.org/10.1037//0022-0663.80.3.284

APPENDIX

Perceived Academic Self-Efficacy Scale (Turkish Form)

Çocuklar için Algılanan Akademik Özyeterlik	Katılma Düzeyiniz			
	Kesinlikle Katılmıyorum	Katılmıyorum	Katılıyorum	Kesinlikle Katılıyorum
1. Okulda çok çalışırım.				
2. Yeterince çalışırsam sınıftaki en iyi notları ben alabilirim.				
3. Sınıf arkadaşlarımın çoğu matematiği sever çünkü matematik kolaydır.				
4. Öğretmenim beni daha çok sevse, daha iyi notlar alabilirim.				
5. Sınıf arkadaşlarımın çoğu ev ödevlerine benden daha çok çalışırlar.				
6. Fen Bilgisi dersinde iyi bir öğrenciyim.				
7. Liseden mezun olacağım.				
8. İyi bir okula okuyorum.				
9. Çok çalıştığımda her zaman iyi notlar alırım.				
10. Bazen, sınıf arkadaşlarımın zor olduğunu düşündüğü ödevler bana kolay gelir.				
11. Sosyal bilgiler dersinde iyi bir öğrenciyim.				
12. Şu an iyi bir işe sahip olan yetişkinler, büyük olasılıkla çocukken iyi bir öğrenciydi.				
13. Yeterince büyüdüğümde üniversiteye gideceğim.				
14. Sınıfımdaki en başarılı öğrencilerden biriyim.				
15. Okulda başarılı olmam kimsenin umurunda değil.				
16. Öğretmenim zeki olduğumu düşünür.				
17. Liseye gitmek önemlidir.				
18. Matematik dersinde iyi bir öğrenciyim.				
19. Sınıf arkadaşlarım genellikle benden daha iyi notlar alır.				
20. Okulda ne öğrendiğim önemli değil.				
21. Genellikle nasıl bir ödev verildiğini anlarım.				
22. Çoğunlukla matematikten iyi notlar alamam çünkü matematik çok zor.				
23. Okulda başarılı olmam önemli değil.				
24. Öğretmenim yüksek puan alan öğrencilere daha çok yardım ediyor.				
25. Türkçe dersinde iyi bir öğrenciyim.				
26. Okulda iyi notlar almak benim için zor değil.				
27. Zekiyim.				
28. Mümkün olan en kısa sürede okulu bırakacağım.				
29. Öğrenciler her zaman güzel notlar almasalar da öğretmenler onları sever.				
30. Öğretmen bir soru sorduğunda diğer öğrenciler yanıtı bilmese bile ben bilirim.				

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 2, 451–469
https://doi.org/10.21449/ijate.959440
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Teachers' perceived skills, challenges and attitudes towards distance education: A validity and reliability study

Derya Cobanoglu Aktan [1,*], **Begum Oztemur** [2]

[1]Hacettepe University, Faculty of Education, Department of Educational Sciences, Ankara, Turkiye
[2]Turkish Ministry of National Education, Düzce, Turkiye

ARTICLE HISTORY
Received: June 29, 2021
Revised: Mar. 15, 2022
Accepted: Apr. 20, 2022

Keywords:
Distance education,
Attitudes towards distance education,
Challenges in distance education,
Perceived distance education skills,
Covid-19 pandemic.

Abstract: This study aimed to develop a reliable and valid measurement tool to examine teachers' attitudes towards distance education and perceived distance education skills. The data of the study were collected from 2290 K-12 teachers. In the data analysis, reliability was calculated with stratified Cronbach's alpha coefficient, and for construct validity, EFA and CFA were performed. Bartlett and KMO tests were used for the suitability of the data for factor analysis. It was observed that the calculated correlations among the item and the total score for the 25-item trial form were above 0.20. As a result of the EFA, 7 items that loaded more than one factor, or with a factor loading less than 0.45 were excluded from the scale. Promax rotation revealed three factors with an eigenvalue greater than 1.00. The total explained variance of the final form of the scale (18 items) was found as 53.594%. The fit indices calculated in the confirmatory factor analysis (RMSEA = 0.053; CFI = 0.932; TLI = 0.918, SRMR = 0.055) confirmed the three-factor model. The results obtained showed that the model fits the data. The stratified alpha reliability for the whole scale was calculated as .848. The results of the study show that the scale can measure teachers' perceived skills, challenges they face, and their attitudes towards distance education reliably and validly.

1. INTRODUCTION

In December 2019, the pandemic that stemmed from Covid-19 virus affected the entire world in a short time and then turned into an intercontinental pandemic in March 2020 (World Health Organization [WHO], 2020a; 2020b). Within the measures taken to tackle the pandemic following the closing down of the educational institutions on March 23, 2020, the Turkish Ministry of National Education started the content preparation for distance education via both the Internet and TV. This sudden change affected both teachers and students. In this context, teachers' attitudes towards distance education and perceived distance education skills are examined in this study. Before focusing on the "teachers", distance education, what distance education is, and the studies about distance education in the field are reviewed in the following section.

*CONTACT: Derya COBANOGLU AKTAN coderya@gmail.com [1]Hacettepe University, Faculty of Education, Department of Educational Sciences, Ankara, Turkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

Moore (1972, p.76) defines distance education as "the family of instructional methods in which the teaching behaviors are executed apart from the learning behaviors . . . so that communication between the learner and the teacher must be facilitated by print, electronic, mechanical or other device". Distance education is the method of structuring courses and managing dialogue between teacher and learner to bridge with communications technology (Moore & Kearsley, 2011). In a general sense in distance education, learners stay at home or office and follow the courses, do the assignments, and interact with each other and the teacher via the Internet. In other words, the learners take the responsibility for their own learning, so that learner autonomy is of great importance (Ekmekçi, 2015). In distance education, in separate places, teachers and students use communication technologies to have one-way or two-way communication by using special software and special equipment (İşman, 2011).

It is contended that the first practice of distance education started in 1728 (Arat & Bakan, 2011), since then it has been carried out with various communication methods and materials. Early practices of distance education employed written materials, mail, and then it was followed by television (Uşun, 2006). In recent years, it has spread into computer environments with the advantages of internet technologies and mobile devices (Telli Yamamoto & Altun 2020).

Before the Covid-19 pandemic, the average age of the distance education participants was outside the compulsory education age group (Akinbadewa & Sofowora, 2020; Alharthi, 2020; Seage & Türegün, 2020). Most individuals chose distance education (Sheets, 1992; Wood, 1996) to get a degree at a higher education level and to meet the demands of knowledge-based economies (Levine, 2001). In this context, a student can enroll in a program given in an accredited institution and graduate from this institution without being physically present (Fornaciari et al., 1999; Kretovics, 1998).

After Covid pandemic, for a certain time, in Turkey distance education became a necessity rather than a choice. During that time, distance education provided a good opportunity for individuals who could not be physically present in the classrooms (Sousa & Florencıo Da Sılva, 2020). This necessity included all students. For a certain period of time, a voluntary basis of distance education became an obligation after the pandemic. Although there was a certain amount of distance education experience at the secondary, high school and university levels in Turkey, for the first time, it included all students at all levels (TEDMEM, 2021). Because of these changes, which are also the subject of this study, teachers had to teach in distance learning platforms (TEDMEM, 2022).

In this change, many teachers around the world were mostly unprepared to support continuity of learning and teaching with distance education. During those times, teachers took more responsibilities required by distance education. Many teachers made great efforts to improve their skills to use technology, digital content preparation, and distance learning while improving their knowledge of their field of interest (Orhan & Beyhan 2020). However, after face-to face formal education was interrupted, many teachers who had not received sufficient training in distance education and who had never had such an experience were caught off guard. In a study examining the distance education experiences of teachers during the Covid-19 pandemic, it was found that the vast majority of teachers (80% of 5.661 teachers) did not have distance education experience (Bahçeşehir University [BAU], 2020). The most common problems voiced by teachers were students' access to technology, knowledge of technology, internet connection, lack of teacher-student interaction, inadequate teaching time, assessment of learning, providing feedback to students, and learning motivation (Hebebci et al., 2020; Korkmaz & Toraman, 2020). To eliminate teachers' lack of knowledge in distance education and to increase their experience, some institutions and organizations such as the Ministry of Education, Teacher Network, and Istanbul Teacher Academies organized webinars on different platforms (e.g.

Zoom, YouTube, Twitter, Facebook) (ERG, 2020; Istanbul Teachers Academy, 2020; The Turkish Ministry of National Education [TMNE], 2020).

One difficulty of distance education is that teachers and students are in separate places and communicate using technology. To communicate effectively, to create a dialogue among learners and teacher, the content and teaching need to be organized with a certain structure. According to Moore (1972) the aim is to build a bridge across an understanding of a teacher and that of a learner. In distance education, teachers organize their courses to manage dialog via technology (Moore & Diehl, 2019). Thus, to teach in distance education, teachers are required to have different skill set, namely organizing course materials with certain structure, technology skills, and creating dialogue using course materials and technology. The fact that computer literacy has become a functional necessity in the learning environment and integrating technology into education has become even more prominent during those times. Technology skills are not enough by themselves to teach in distance education; however, they form, some of the essential parts to create a bridge for students' learning. It is reported that many educators lacked the most basic computerized communication technologies (CCT) skills, even if they had sufficient infrastructure and connectivity (UNESCO, 2020a; UNESCO, 2020b). Instructors generally use information communication technologies; for web searching, communicating, benefiting from electronic services, and making presentations, but they do not frequently use it for participating in forums, video and voice chat, creating multimedia, and presenting courses on the Internet (Düzakın & Yalçınkaya, 2008); therefore, it was stated that they need to improve their professional competencies with respect to quality distance education. (UNESCO, 2020b). It is also very important to assess teachers' readiness for online teaching, as it plays an important role in the effective delivery of online education (Miglani & Awadhiya, 2017).

There seems to be a global need to develop an understanding of educators' and schools' readiness for distance education and to modernize teacher education to meet the needs of knowledge-based global society. In times of crisis, it has also become important to increase teachers' applications in the use of CCT for pedagogy, digital literacy, and data assessment to enable more individualized learning (UNESCO, 2020b). It is important for teachers to keep up with the changes. In addition, determining to what extent they have sufficient knowledge to carry out the practices of distance education is also important to direct the training to be provided for teachers.

There are many studies on online teaching and distance education. The following part focuses on teachers' attitudes and perceptions of distance education. Higher education is the focus of most studies. The studies in the United States are reviewed by Shattuck (2019). In teaching online chapter of handbook of distance education, Shattuck (2019) summarizes characteristics of faculty members by answering "where, what, who, when, why, how" questions. Intrinsic and extrinsic motivators, age, gender, technology experience, and faculty rank on motivation, demotivators, faculty attitudes, values, and perceptions are listed among those faculty characteristics (Moore & Diehl, 2019). The relationship between technology acceptance and intentions to teach online was examined by Stewart, Bachman, and Johnson (2010). Dahlstrom and Brooks (2014) explored faculty members' perceptions of information and educational technology. AlShahrani (2014) investigated perceived self-efficacy in using technology and teaching online. A survey by Babson Research Group (Lammers et al., 2017) found that faculty are critical to the success of digital learning, and when they are supported. Ulmer et al. (2007) explored a link between attitudes and participation in online learning and acknowledged that faculty with experience in online distance education tended to have positive attitudes. Lin (2002) found that faculty was more likely to take part if they had a positive attitude toward distance education or had a positive distance education experience. Moreover, Zhen et al. (2008) explored faculties' teaching values and attitudes towards teaching online. According to

Zhen et al. (2008) if faculty members do not see intrinsic value and perceive their pedagogical values as being accommodated and encouraged, they might focus on demotivators and do not wish to teach online. In literature, the barriers that decrease faculty participation in distance online education have also been identified. Dillon and Walsh (1992), Berge et al. (2002), and Shea (2007) reported these barriers which negatively influence faculty participation in distance education; namely, lack of quality in online education, lack of time, lack of compensation, lack of incentives and/or rewards, lack of policies and institutional support, and lack of perceived student interaction.

When the studies conducted to examine the opinions of the teachers and faculty members on distance education in Turkey are reviewed, it is seen that Turkey did not benefit enough from the educational potential provided by the e-technology to meet the educational needs (Özkul, 2004). Further, even though various distance education applications are implemented, it is thought that the distance education applications are not efficient enough, and many of the web-based distance education programs do not go beyond downloading the lecture notes from websites (Gülnar, 2003). Orhan and Beyhan (2020), in their study, examined teachers' opinions on Zoom and stated that teachers see distance education as a supportive education as a continuation of formal education, while some teachers stated negative opinions. It is reported in the studies that some lecturers have negative attitudes towards distance education (Kaya et al., 2017; Yıldırım, 2020). Reasons for the demotivators (sources of negative attitudes) are reported as inadequate student participation, difficulties in preparation and presentation of course materials, and habits of face-to-face education interaction (Kaya, 2002). Faculty members needed training for web-supported education (Erişti et al., 2008; Soydal et al., 2012) and lacked necessary materials and equipment (Korkmaz & Tunç, 2010). It was also reported that teachers do not receive feedback from students during the lessons (Orhan & Beyhan, 2020) and in this specific context, lack of immediate feedback (course structure), complexity of the interface, lack of control in student-student interaction (dialogue), and lack of feedback (dialogue) in teacher-student interaction are considered as demotivators (Hamıutoğlu et al., 2018).

1.1. The Purpose of the Study

When the relevant literature for measuring attitudes and perceptions of distance education was reviewed, it was seen that there are scale development studies that focused on higher education institutions (Akaslan & Law, 2011; Dündar et al., 2017; Süer et al., 2005). The scale developed by Akaslan and Law (20119 composed of three factors: "readiness to e-learning", "acceptance of e-learning" and "e-learning education". Similarly, "Distance Education Attitude Scale" by Süer et al. (2005) included "trust in distance education" and "interest in distance education" factors. On the other hand, Dündar et al., (2017) developed a three-factor scale with a “cognitive”, “affective”, and “behavioral” factors. The only scale developed for K-12 level primary school teachers has two factors under the headings of the advantages of distance education and the limitations of distance education (Ağır, 2008). All these scales were created before the Covid-19 pandemic, when distance education was a choice rather than a necessity, and teachers were unprepared for distance education. Therefore, there was a need for a measurement tool to measure teachers' attitudes towards distance education, and their perceived distance education skills at all K-12 levels, during the closing down of the face-to-face formal education.

According to the studies, teachers’ attitudes, and their technical and pedagogical characteristics affect the success of online learning (Dillon & Guawardena, 1995; Leidner & Jarvenpaa, 1993; Volery & Lord, 2000). Therefore, differences in teachers’ attitudes, access to technical infrastructure, and tools will result in the difference in students’ learning. In this context, it is

important to have studies on measurement tools to measure teachers' attitudes towards distance education and their perceived distance education skills.

Considering the importance and the gap in the literature, the purpose of this study is to develop a reliable and a valid scale to evaluate attitudes towards distance education and perceived distance education skills of teachers working at primary and secondary education towards distance education.

2. METHOD

This section provides information about the study groups, the process of developing the scale, and the data analysis.

2.1. Study Group

The data used within the study were obtained from 2290 K-12 teachers (1145 of which were used in the Exploratory Factor Analysis and 1145 participants' data were used in the Confirmatory Factor Analysis). The data were collected from teachers working at all levels from 290 different primary school to high school in Turkey's Western Black Sea Region.

2.2. The Development of Item Pool

In order to write the items to be included in the scale, firstly the related literature was reviewed, scales developed for similar purposes were examined, and teachers' opinions about distance education were collected. When the studies on distance education were examined, it was seen that there were scales developed for primary school teachers (Ağır, 2008) and faculty members of higher education (Akaslan & Law, 2011; Dündar et al., 2017; Süer et al., 2005). Similarly, various qualitative studies on teachers' views on distance education were reviewed (Alshangeeti et al., 2009; Chao et al., 2006; Erişti et al., 2008; Göktaş & Kayri, 2005; Kaya, 2002; Lloyd et al., 2012; Miglani & Awadhiya, 2017). In addition to studies, reports were also reviewed (Bahçeşehir University [BAU], 2020; ERG, 2020; The Turkish Ministry of National Education [TMNE], 2020). Moreover, views of teachers were collected via open-ended questions, regarding opinions and difficulties about distance educations to generate items. When all these studies and views were examined, teachers' attitudes towards distance education and perceived skills of distance education dimensions were identified.

The pilot form of the scale was reviewed by two experts in the field of measurement and evaluation, two secondary school level teachers (mathematics, literacy), three elementary school teachers, and one psychological counseling and guidance teacher. Measurement and evaluation experts reviewed the items for content and item characteristics. Teachers assessed items for content representation. Reviewers assessed items as appropriate or inappropriate and also suggested revisions for the items if they thought it was necessary. Only minor wording revisions were suggested by the reviewers. The trial form was comprised of 25 items. Each item was scored as "*Strongly Disagree*", "*Disagree*", "*Undecided*", "*Agree*" and "*Strongly Agree*" according to the 5-point likert type grading scale.

2.3. Data Analysis

The data were collected using Google form. The Provincial Directorate of National Education shared the link of the form with teachers via SMS. No missing data was found in the data. Outliers were determined via Mahalanobis distances. The data from 168 participants were deleted according to their Mahalanobis distances. A Mahalanobis distance (χ^2 (25) = 38.104) was used to detect multivariate outliers. When the variance increase value [VIF] was analyzed for the remaining 977 participants' data, it was seen that it ranged between 1.202 and 2.997. Therefore, it can be interpreted that there is no multicollinearity problem for the data obtained from our sample because the VIF values were less than 10.

In this study, exploratory factor analysis [EFA] and confirmatory factor analysis [CFA] were performed for the construct validity of the teachers' attitudes towards the distance education scale. EFA aims to reach a few definable meaningful structures that these variables can explain together from many variables (items) and it is a method used to reveal whether there is a certain order among the responses of the respondents to the items in the measurement tool which has been developed (Büyüköztürk, 2004; Tavşancıl, 2006). In this study, EFA analysis was run in SPSS statistical software.

CFA was used to evaluate to what extent the factors formed from various variables theoretically matched the actual data. The extent to which a predetermined or constructed structure was verified by the collected data in CFA was examined. Some fit indices were used to determine the adequacy of the model tested in CFA (Büyüköztürk et al., 2004).

Confirmatory factor analysis [CFA] study was conducted (in MPlus) on data obtained from a different sample of 1145 teachers in order to provide evidence for the validity of the structure determined as a result of EFA and to reveal to what extent the observed structure was compatible with the data. Multiple fit indices were used for CFA and Chi-square fit test [Chi - Square Goodness], Comparative Fit Index [CFI], Root Mean Square Error of Approximation [RMSEA], Tucker-Lewis Index [TLI] and Standardized Root-Mean-Squared Residual [SRMR] fit indices were examined (Hu & Bentler, 1999; Çelik & Yılmaz, 2013; Kline, 2005).

Table 1. *Multivariate skewness and kurtosis test results.*

	Sample Value	x̄	ss	*p*
Skewness Value	36.134	7.992	0.3	0
Kurtosis Value	425.746	398.365	1.666	0

The normality of the data was examined in MPlus via multivariate skewness and kurtosis tests. The results are presented in Table 1. It can be interpreted that the data do not meet the assumption of normality, since the tests performed for skewness and kurtosis are statistically significant. Therefore, Maximum Likelihood Robust [MLR] was preferred as the estimation method in CFA. MLR method provides stronger estimation in non-normal data (Wang & Wang, 2019). In this study, stratified alpha value for reliability was calculated. When the literature is examined, it is recommended that the Stratified Cronbach's Alpha coefficient be used for the reliability of composite scores obtained from measurement tools containing sub-dimensions (Cronbach et al., 1965). Stratified Cronbach's alpha coefficient was calculated using the "sirt" package in the R program (Robitzsch, 2021).

3. FINDINGS

3.1. Findings on the Construct Validity Evidence of the Scale

3.1.1. *Exploratory factor analysis*

The item and the item-total scale correlations showed that there was no item below 0.20. The Bartlett test and Kaiser-Meyer-Olkin [KMO] values of the data obtained for the suitability of the data related to the scale trial form comprising of 25 items after item analysis were examined.

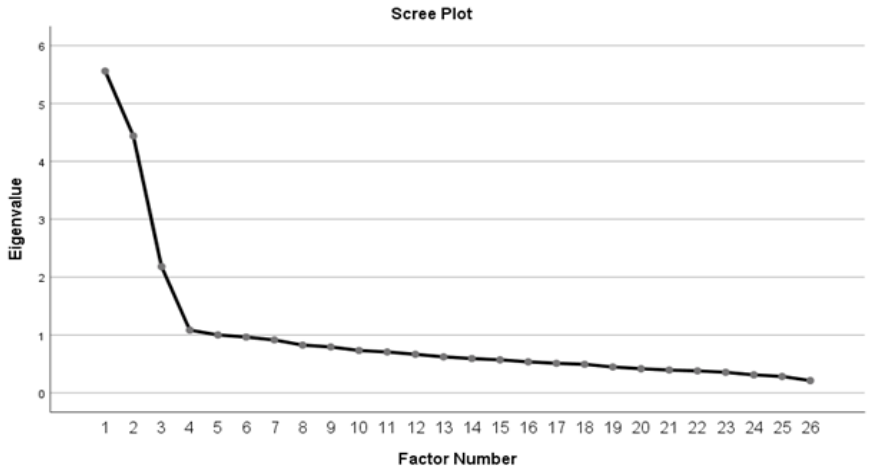
The calculated KMO value was found to be 0.878, and it was seen that for the Barlett test, the calculated chi-square statistics was also significant ($\chi^2 = 9170.480$, $df = 300$, $p < 0.01$). KMO values were determined to be quite high. It can be said that the sample size is suitable for factor analysis because the KMO value is high and the Bartlett test is significant. Exploratory factor analysis was conducted to determine the construct validity based on these data. In the exploratory factor analysis, the number of factors was determined according to the scree plot. According to the plot in Figure 1, the number of factors was found to be 3.

Table 2. *EFA eigenvalues and parallel analysis eigenvalues.*

Factor	Eigenvalues	PA eigenvalues
1	**5.552**	**1.3011**
2	**4.390**	**1.2563**
3	**2.025**	**1.2245**
4	1.073	1.1951

The parallel analysis supported this finding. According to the values given in Table 2, there are three factors, where the calculated eigenvalues are greater than the random eigenvalues generated in parallel analysis.

Figure 1. *Scree plot.*



After determining the number of factors, Promax rotation was used in factor analysis and Principal Axis Factoring was used as a method. The structure resulting from rotation helps to obtain items that can be classified meaningfully in only one category. Therefore, when using oblique rotation method, Promax rotation method is a good option to be fast and economical (Çokluk et al., 2016). Table 3 shows the factors and items after the rotation. After the rotation, the items that were not loaded under any factor were removed one by one, and then the items with a factor loading value below 0.45 were eliminated, starting with the item with the lowest value. Comrey and Lee (1992) suggest a scale of quality of factor loadings that is often referenced: .71 is excellent, .63 is very good, .55 is good, .45 is fair, and .32 is poor (Multiplicity et al., 2014). When the factor loading values are examined, item 1 and item 13 were discarded since the factor loading values were below 0.45. After this deletion, in EFA results, items that loaded more than one factor were also excluded from the scale. Following this rule, item 16, item 18 and item 24 were excluded from the scale since they loaded more than one factor. After removing these items from the scale and repeating the factor analysis, the factor loading of the item 19 was below 0.45 (please note that Table 3 presents item loadings with item 16, 18 and 24). Item 2 was also removed, because it was conceptually different from the items in that factor ("I think face to face education is a necessity for the best education" conceptually does not align with "challenges faced in distance education" factor). Thus, it was also excluded from the scale. Final factor loadings are given in Table 5.

Explained and total variances are presented in Table 4. The total variance percentage explained by the three factors is 53.594%. The variance explained by each factor is 24.361%, 20.016% and 9.218% and the eigenvalues calculated for each factor are 4.385; 3.603 and 1.659, respectively.

Table 3. *Factor loading values after rotation.*

	Factors		
	1	2	3
m1			.428
m4			.583
m5			.591
m8			.663
m9			.442
m11			.625
m12			.569
m13	.397		
m14	.637		
m15	.714		
m16	.335		.477
m18	.366	-.349	
m19			.489
m20	.685		
m21	.758		
m22	.807		
m24	-.516	-.494	.361
m2	-.301	.502	
m3		.591	
m6		.569	
m7		.653	
m10		.520	
m17		.462	
m23		.702	
m25		.724	

Table 4. *Eigenvalue and variance percentages for each factor.*

Factor	Values		
	Eigenvalues	Explained Variance (%)	Total Variance (%)
1	4.385	24.361	24.361
2	3.603	20.016	44.376
3	1.659	9.218	53.594

The final factor loading values obtained with promax rotation are presented in Table 5. According to the values specified in Table 5, the first factor on the scale comprises 7 items (m3, m6, m7, m10, m17, m23, m25); the second factor comprises 5 items (m14, m15, m20, m21, m22); and the third factor comprises 6 items (m4, m5, m8, m9, m11, m12). The names were given to each factor by considering the literature and their contents. The first factor was named as "challenges faced in distance education", the second factor as "perceived distance education skills", and the third factor as "positive attitudes toward distance education". The scale items and the factor loading values are presented in the Appendix Section. A confirmatory factor analysis was performed to provide evidence for the validity of the structure determined as a result of EFA (18 items and three factors). Data obtained from a different study group of 1145

people were used for CFA. In this data set, no missing data was observed, and the answers of 148 participants who showed outliers were deleted as a result of Mahalanobis distance. CFA was performed with the remaining data of 997 teachers.

Table 5. *Final factor loading values of the 18-item scale.*

Items	Factors		
	F1 (challenges faced)	F2 (perceived skills)	F3 (positive attitudes)
m4			.527
m5			.560
m8			.725
m9			.528
m11			.634
m12			.570
m14		.551	
m15		.700	
m20		.768	
m21		.856	
m22		.825	
m3	.558		
m6	.613		
m7	.700		
m10	.531		
m17	.460		
m23	.693		
m25	.702		

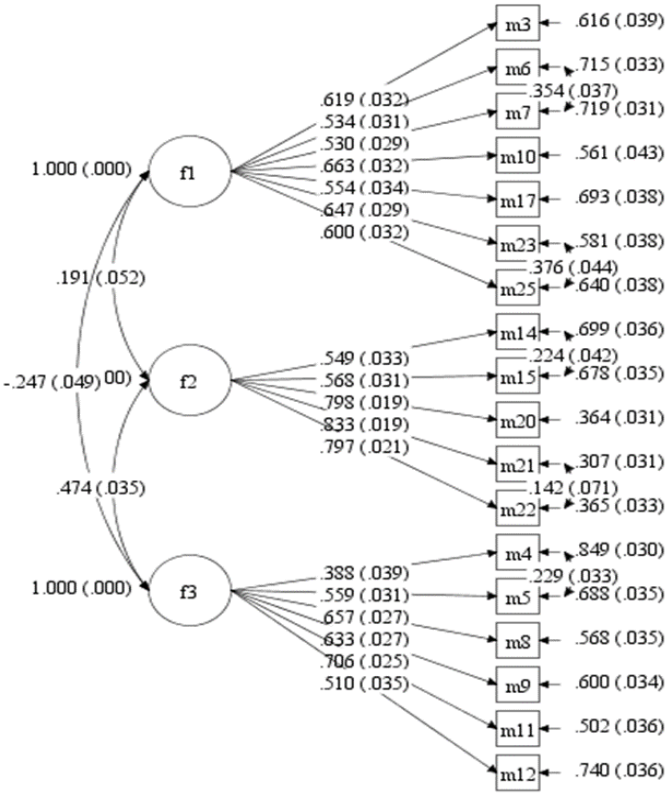
To test the data fit of the three-factor model, χ^2/df, RMSEA, CFI, TLI, SRMR values were calculated. The calculated fit indices were not in the range of acceptable values for good model fit, as seen in Table 6. Therefore, after an examination of the modification indices, modifications were carried out in order to decrease the chi-square value. According to their modification indices, the error terms of the items were correlated in the measurement model. The measurement model of the scale is presented in Figure 2.

The fit indices for the first and modified measurement model are presented in Table 6. The modified model fit indices are RMSEA = .053; CFI = .931; TLI = .918; SRMR = .055. These values show that model fit is ensured. The calculated value of χ^2= 486.856 (df = 127) was significant (p <.01) and χ^2/ df = 3.833. According to the literature, if the χ^2 is df < 3, it is the proof of perfect fit; if it is below 5, it is the proof of medium level of fit; if RMSEA and SRMR value is .80 or less it is acceptable fit; if the CFI and TLI value is higher than 0.90, it is accepted as an indicator of acceptable fit (Hu & Bentler, 1999; Kline, 2005; Şimşek, 2007; Yılmaz & Çelik, 2009).

Table 6. *The model fit indices and values calculated for the models.*

Model Fit Indices	(χ^2/df)	RMSEA	CFI	SRMR	TLI
First Model	6.018	.071	.874	.062	.854
Modified Model	3.833	.053	.931	.055	.918

Figure 2. *Model obtained by CFA.*



As seen in Figure 2, the item factor loading values are in a range of 0.307 and 0.849. As a result, it is seen that the factor loading values are acceptable. According to these results, it can be said that the model fits the data well.

Descriptive statistics for each factor were also explored. For the challenges faced in distance education factor, the maximum and minimum values were between 35 and 7, and the mean was found as 27.13, which indicates that participants thought they faced challenges during distance education. The distribution of the factor scores was negatively skewed, meaning that the number of the participants who faced high-level challenges was more than the number of participants who faced low-level challenges. Regarding the perceived skills factor, maximum and minimum values were between 25 and 5, and the mean was 17.87. This shows that participants perceived themselves as skilled in distance education. Finally, in terms of positive attitudes factor, maximum and minimum values were 30 and 6, and the mean was 13.46. The distribution of the scores was positively skewed, meaning that the number of participants with high positive attitudes levels was less than that of the low levels.

3.2. Reliability of the Scale

Cronbach Alpha internal consistency coefficients were calculated to determine the reliability of the final scale obtained for each factor and for total scale.

Table 7. *Cronbach Alpha values for the scale.*

Factor Number	Factor Name	Number of items	Cronbach Alpha
F1	Challenges faced in distance education	7	.804
F2	Perceived distance education skills	5	.865
F3	Positive attitudes toward distance education	6	.776
	Total (Stratified Alpha)	18	.848

Cronbach alpha values for the scale are presented in the Table 7. As seen in the Table 7 Stratified Cronbach's Alpha reliability of the whole scale was calculated as α = .848. In addition, the alpha reliability values obtained for each factor are respectively α = .804 for the 1st factor; α= .865 for the 2nd factor; and α = .776 for the 3rd factor. The obtained alpha coefficients are considered quite reliable for values between .60 and .79 in the literature and are highly reliable for values of .80 and above (Kalaycı, 2010). The overall reliability of the scale was calculated as .848 with stratified alpha.

4. DISCUSSION and CONCLUSION

In this study, a scale was developed to measure teachers' attitudes towards distance education, their perceived skills of distance education, and the challenges faced in distance education. The data were collected from 2290 primary and secondary education level teachers. The initial form of the scale comprised 25 items. To collect construct validity evidence, EFA and CFA were performed. In the EFA, 7 items were excluded from the scale. Three items, "I think I manage the classroom better in online lessons", "I use the existing materials in my online lessons", and "I think that in distance education, students regularly do their homework" were excluded from the scale because of their low factor loadings. In addition, "I am satisfied with distance education, because there are no distractions such as students talking among themselves or going out of the classroom", "I think it is necessary to examine the learning levels of students in distance education through exams", and "I think homework in distance education is sufficient in determining the learning levels of students" were eliminated from the scale because these items were loaded on more than one factor. Lastly "I think face to face education is necessity for the best education" was not included in the scale because of the content discrepancy with the factor.

The factors were named as "challenges faced in distance education", "perceived distance education skills" and "positive attitudes toward distance education". The first factor included items such as "In distance education, I feel like I'm talking by myself." The other items similarly are about challenges that teachers faced in distance education. The second factor, perceived distance education skills, included items about teachers' own skill perceptions such as "In online lessons, I can adequately employ multimedia such as graphics, sound and animation."

The third factor, positive attitudes toward distance education, is composed of items such as "I think distance education is suitable for student groups of all ages."

The measurement model obtained by EFA was verified by CFA with the data obtained from 1145 teachers. The fit indices of the model (three factors and 18 items) in the CFA were in the range of acceptable limits. In the model, positive correlation (.474) between "perceived distance education skills" and "positive attitudes towards distance education" factors was observed. In contrast, as expected, a negative correlation was found (-.247) between "challenges faced in distance education" and "positive attitudes towards distance education" factors. Lastly, there is a weak correlation (.191) between "challenges faced in distance education" and "perceived distance education skills" factors. The stratified reliability coefficients of the factors were calculated as .804, .865 and .776, respectively. In addition, for the whole scale stratified Cronbach's Alpha reliability coefficient value was found as .848. These evidences showed that the developed scale is valid and reliable in determining teachers' attitudes towards distance education, perceived distance education skills, and challenges faced in distance education.

In the scale developed by Ağır et al. (2008) some items in the disadvantages of distance education factor are similar to the challenges faced in distance education, however, the scale do not include any items regarding teachers' perceived distance education skills. In addition, the factor of positive attitudes towards distance education is similar to the factor of advantages of distance education. Süer et al.'s scale included two factors: trust in distance education and

interest in distance education. However, the researchers did not report any further information about the items and the concepts related to the items. Thus, it is not possible to make any comparison with the scale other than that it was administered to the different education level (higher education versus K-12). Dündar et al. (2017) reported that their scale included cognitive, affective, and behavioral factors related to the attitudes toward distance education. Similar to Süer et al. (2005), Dündar et al. also did not publish the items of their scale, but they just gave example items for the related factors.

According to the results of our study, the participants of the study reported that although they perceived themselves as skilled in distance education, they also faced challenges during the distance education and reported comparatively low level positive attitudes toward distance education. Research calls challenges faced in distance education as also demotivators of distance education. Considering this, to improve teachers' motivations and positive attitudes, teachers require support to overcome these challenges. This finding aligns with conclusions by Lin (2002) that faculty were more likely to participate if they had a positive attitude toward distance education or had a positive distance education experience. This is also supported by Shattuck's (2013) findings as Shattuck stated that, although faculty members are intrinsically motivated, they also value and need support services, including training opportunities in technology skills, design, and instructional support, and awareness of sound student support services. Although previous studies were conducted with faculty members, this study showed that K-12 teachers shared similar experiences in terms of challenges they faced during distance education.

Even if face-to-face education has started in Turkey, it is important to examine teachers' skills, attitudes and difficulties in order to be ready for the future or to benefit from distance education when necessary. In this context, the scale developed in this study provides a valid and reliable measurement tool to determine the attitudes and skill perceptions of teachers working in primary and secondary education. It is thought that developing a standardized measurement tool to measure teachers' perceptions of their attitudes and skills towards distance education will be beneficial in increasing teachers' positive attitudes and reducing the difficulties they face. It should be noted that the findings of this study are limited to the study participants and pandemic conditions. This study is limited to the group of teachers working in the Western Black Sea Region in Turkey. For further validation, the scale can be administered to groups of teachers working in different regions of Turkey, or at different education levels.

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, 35853172-600

Authorship Contribution Statement

All authors contributed to the manuscript equally.

Orcid

Derya Çobanoglu Aktan https://orcid.org/00000-002-8292-3815
Begum Oztemur https://orcid.org/0000-0001-5761-2567

REFERENCES

Adıyaman, Z. (2014). Uzaktan eğitim yoluyla yabancı dil öğretimi [Foreign language teaching through distance education]. *Sakarya University Journal of Education Faculty*, 4, 420-425. https://dergipark.org.tr/tr/pub/sakaefd/issue/11231/134140

Ağır, F. (2008). Uzaktan eğitime karşı tutum ölçeği geliştirmeye yönelik geçerlilik ve güvenirlik çalışması [Development of the attitude scale towards distance learning: Reliability and validity], *Education Sciences*, *3*(2), 128-13.

Akaslan, D., & Law, E.L.C. (2011). Measuring teachers' readiness for e-learning in higher education institutions associated with the subject of electricity in Turkey. *2011 IEEE Global Engineering Education Conference (EDUCON)*, 481-490. https://doi.org/10.1109/educon.2011.5773180

Akinbadewa, B.O., & Sofowora, O.A. (2020). The effectiveness of multimedia instructional learning packages in enhancing secondary school students' attitudes toward Biology. *International Journal on Studies in Education (IJonSE)*, *2*(2), 119-133. https://doi.org/10.46328/ijonse.19

Alharthi, M. (2020). Students' attitudes toward the use of technology in online courses. *International Journal of Technology in Education (IJTE),* *3*(1), 14-23. https://doi.org/10.46328/ijte.v3i1.18

Allen, B., Crosky, A., Yench, E., Lutze-Mann, L., Blennerhassett, P., Lebard, R., Thordarson, P., & Wilk, K. (2010). A model for transformation: A trans-disciplinary approach to disseminating good practice in blended learning in science faculty, *Curriculum, technology & transformation for unknown future*. *Proceedings of ascilite Sydney,* 36-48.

Alshangeeti, A., Alsaghier, H., & Nguyen, A. (2009). Faculty perceptions of attributes affecting the diffusion of online learning in Saudi Arabia: A quantitative study*, 4th International Conference on e-Learning*, Toronto, Canada.

Anderson, J. (2020). The coronavirus pandemic is reshaping education, *Quartz.* https://qz.com/1826369/how-coronavirus-is-changing-education/

Arat, T., & Bakan, Ö. (2011). Uzaktan eğitim ve uygulamaları [Distance education and its applications], *Selçuk Üniversitesi Sosyal Bilimler Meslek Yüksekokulu Dergisi, 14*(1), 363-374.

Arora, A.K., & Srinivasan, R. (2020). Impact of pandemic COVID-19 on the teaching–learning process: A Study of Higher Education Teachers. *Prabandhan: Indian Journal of Management*, *13*(4), 43-56. https://doi.org/10.17010/pijom/2020/v13i4/151825

BAU (2020, August 16). *BAU distance education report.* https://bau.edu.tr/haber/15707-%20bau-uzaktan-egitim-raporu

Bozkurt, A., Akgun-Ozbek, E., Yilmazel, S., Erdogdu, E., Ucar, H., Guler, E., Sezgin, S., Karadeniz, A., Sen-Ersoy, N., Goksel-Canbek, N., Dincer, G.D., Ari, S., & Aydin, C.H. (2015). Trends in distance education research: A content analysis of journals 2009-2013. *The International Review of Research in Open and Distributed Learning*, *16*(1). https://doi.org/10.19173/irrodl.v16i1.1953

Büyüköztürk, Ş. (2004). *Veri analizi el kitabı [Data analysis handbook]*. Pegem A Yayıncılık.

Cabı, E. (2018). Uzaktan eğitim ile bilgisayar okuryazarlığı öğretimi: Eğitmen deneyimleri [Computer literacy teaching with distance education: Instructor experiences]. *Başkent University Journal of Education*, *5*(1), 61-68. http://buje.baskent.edu.tr/index.php/buje/article/view/93

Can, E. (2019). Açık ve uzaktan yükseköğretim mezunları üzerine bir değerlendirme [An evaluation on open and distance higher education graduates]. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi*, *5*(3), 81-105. https://dergipark.org.tr/en/download/article-file/853649

Can, E. (2020). Coronavirüs (Covid-19) pandemisi ve pedagojik yansımaları: Türkiye'de açık ve uzaktan eğitim uygulamaları [Coronavirus (Covid-19) pandemic and its pedagogical reflections: Open and distance education practices in Turkey]. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi*, *6*(2), 11-53. https://dergipark.org.tr/en/download/article-file/1179832

Chao, T., Saj, T., & Tessier, F. (2006). Establishing a quality review for online courses. *Educause Quarterly*, *29*(3), 32-39.

Chen, G.D., Ou, K.L., Liu, C.C., & Liu, B.J. (2001). Intervention and strategy analysis for web group- learning. *Journal of Computer Assisted Learning*, *17*(1), 58-71. https://doi.org/10.1111/j.1365-2729.2001.00159.x

Comrey, A.L., & Lee, H.B. (1992). *A first course in factor analysis. (2nd ed.)*. Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L.J., Schönemann, P., & McKie, D. (1965). Alpha coefficients for Stratified-Parallel Tests. *Educational and Psychological Measurement, 25*(2), 291-312. https://doi.org/10.1177%2F001316446502500201

Czerniewicz, L. (2020, 15 March). What we learnt from "going online" during university shutdowns in South Africa. *Phil on EdTech*. https://philonedtech.com/what-we-learnt-from-going-online-during-university-shutdowns-in-south-africa/

Çelen, F.K., Aygül Ç., & Seferoğlu S.S. (2013). Analysis of teachers' approaches to distance education. *Procedia-Social and Behavioral Sciences*, *83*, 388-392.

Çelik, H., & Yılmaz, V. (2013). *LISREL 9.1 ile yapısal eşitlik modellemesi [Structural equation modeling with LISREL 9.1].* Anı Yayıncılık.

Çokluk, Ö., Şekercioğlu, G. & Büyüköztürk, Ş. (2014). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları [Multivariate statistics for social sciences: SPSS and LISREL applications]*. Pegem Akademi Yayınları.

Dillon, C.L., & Guawardena, C.N. (1995). A framework for the evaluation of telecommunications-based distance education. *In Proceedings of Selected Papers from the the 17th Congress of the International Council for Distance Education, Milton Keynes, UK: Open University.*

Dooley, K.E., & Murphrey, T.P. (2000). How the perspectives of administrators, faculty, and support units impact the rate of distance education adoption. *Online Journal of Distance Learning Administration*, *3*(4). https://www.learntechlib.org/p/92503/

Dougiamas, M. (2000). Improving the effectiveness of tools for the internet-based education. *9th. Annual Teaching Learning Forum*, Curtin University of Technology, USA.

Durak, G., Çankaya, S., Yunkul, E., Urfa, M., Topraklikloğlu, K., Arda, Y., & İnam, N. (2017). Trends in distance education: A content analysis of master's thesis. *TOJET: The Turkish Online Journal of Educational Technology*, *16*(1).

Dündar, S., Candemir, Ö., Demiray, E., Genç Kumtepe, E., Öztürk, S., Sağlık Terlemez, M., & Ulutak, İ. (2017). Anadolu Üniversitesi çalışanlarının açık ve uzaktan öğretime ilişkin tutumları [Attitudes of Anadolu University employees towards open and distance education]. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi, 3*(4), 187-227.

Düzakın E., & Yalçınkaya S. (2008). Web tabanlı uzaktan eğitim sistemi ve Çukurova Üniversitesi öğretim elemanlarının yatkınlıkları [Web-based distance education system and the predispositions of Çukurova University faculty members]. *Çukurova Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *17*(1), 225-244.

Eğitim Reformu Girişimi (ERG), (2020, July 7). *The effects of Coronavirus on education VII: The only thing that doesn't change is the importance of teachers*. https://www.egitimreformugirisimi.org/e-bulten/

Erişti, S.D., Şişman, E., & Yıldırım, Y. (2008). İlköğretim branş öğretmenlerinin web destekli öğretim ile ilgili görüşlerinin incelenmesi [Examination of primary school branch teachers' views on web-assisted teaching]. *İlköğretim Online, 7*(2), 384-400. https://app.trdizin.gov.tr/publication/paper/detail/TnpjM016RXg

Ferdig R.E., & Kennedy K. (2014). *Handbook of Research on K-12 Online and Blended Learning*. Carnegie Mellon University: ETC Press

Fornaciari, C.J., Forte, M., & Mathews, C.S. (1999). Distance education as strategy: How can your school compete?. *Journal of Management Education, 23*, 703-718.

Galusha, J.M., (1998). Barriers to Learning in Distance Education. *Institute of Education Sciences*. https://files.eric.ed.gov/fulltext/ED416377.pdf

Geng, S., Law, K.M.Y., & Niu, B. (2019). Investigating self-directed learning and technology readiness in blending learning environment. *International Journal Education Technology in Higher Education, 16*(17). https://doi.org/10.1186/s41239-019-0147-0

Gökmen, Ö., Uysal, M., Yaşar, H., Kırksekiz, A., Güvendi, G., & Horzum, M. (2017). Türkiye'de 2005-2014 Yılları Arasında Yayınlanan Uzaktan Eğitim Tezlerindeki Yöntemsel Eğilimler: Bir İçerik Analizi [Methodical Trends in Distance Education Theses Published Between 2005-2014 in Turkey: A Content Analysis]. *Eğitim ve Bilim, 42*(189). http://dx.doi.org/10.15390/EB.2017.6163

Gökdaş, İ., & Kayri, M. (2005). E-öğrenme ve Türkiye açısından sorunlar, çözüm önerileri [E-Learning-The problems and solution recommends terms of Turkey situation]. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi, 2*(2), 1-20.

Gülnar, B. (2003). *Bilgisayar ve İnternet Destekli Uzaktan Eğitim Programlarının Tasarım, Geliştirme ve Değerlendirme Aşamaları (SUZEP Örneği) [Design, Development and Evaluation Phases of Computer and Internet Supported Distance Education Programs],* (Master thesis, Selçuk University, Konya). https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Hamutoğlu, N.B., Sezen Gültekin, G., & Savaşçı, M. (2019). Öğretmen adaylarının uzaktan eğitime yönelik görüşleri: Açıköğretim uygulamaları [Pre-service teachers' views on distance education: Open education practices]. *Yükseköğretim Dergisi, 9*(1), 19–28. https://doi.org/10.2399/yod.18.023

Harris, D.A., & Krousgrill, C. (2008). Distance education: New technologies and new directions. *Proceedings of the IEEE*, *96*(6), 917-930. https://doi.org/10.1109/JPROC.2008.921612.

Hebebci, M.T., Bertiz, Y., & Alan, S. (2020). Investigation of views of students and teachers on distance education practices during the Coronavirus (COVID-19) Pandemic. *International Journal of Technology in Education and Science (IJTES), 4*(4), 267-282. https://files.eric.ed.gov/fulltext/EJ1271267.pdf

Holmberg, B. (1989). *Theory and practice of distance education*. New York: Routledge.

Horzum, M.B., Özkaya, M., Demirci, M., & Alpaslan, M. (2013). Türkçe uzaktan eğitim araştırmalarının incelenmesi [Examination of Turkish distance education research]. *İnönü Üniversitesi Eğitim Fakültesi Dergisi, 14*(2), 79-100.

Hu, L.T., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. https://doi.org/10.1080/10705519909540118.

Istanbul Teachers Academies (2020, July). *Special education academy*. https://istanbulakademi.meb.gov.tr/akademiler.php?cID=14

İşman, A. (2011). *Uzaktan eğitim [Distance Education]*. Pegem Akademi.

Johnson, A.E. (2008). A nursing faculty's transition to teaching online. *Nursing Education Perspectives, 29*(1), 17-22.

Kalaycı, Ş. (2010). *SPSS uygulamalı çok değişkenli istatistik uygulamaları [Multivariate statistics applications with SPSS]*. Asil Yayın Dağıtım Ltd. Şti.

Kaya, Z. (2002). *Uzaktan eğitim [Distance Education]*. Pegem A Yayıncılık.

Kaya, Z., & Önder, H. (2002). İnternet yoluyla öğretimde ergonomi [Ergonomics in ınternet teaching]. *The Turkish Online Journal of Educational Technology, 1*(1), 48-54. http://www.tojet.net/articles/v1i1/118.pdf

Kaya, M., Çitil Akyol, C., Özbek, R., & Pepeler, E. (2017). Lisansüstü eğitim programlarında 'uzaktan eğitim uygulamasına' yönelik 'Eğitim Bilimleri Bölümü' akademisyenlerinin görüşleri [Opinions of 'Educational Sciences Department' academicians on 'distance education application' in graduate education programs], *Elektronik Sosyal Bilimler Dergisi, 16*(64), 1616-1627. https://dergipark.org.tr/tr/pub/esosder/issue/31376/328632

Kline, R.B. (2005). *Methodology in the social sciences. Principles and practice of structural equation modeling.* Guilford Press.

Korkmaz, Ö., & Tunç, S. (2010). Mesleki-teknik eğitim öğretmenlerinin bilgisayar ve internet temelli öğretim materyallerinden yararlanmaya ilişkin görüşleri [Vocational-technical education teachers' opinions on using computer and internet-based teaching materials]. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi, 11*(3). 263-276. https://dergipark.org.tr/en/download/article-file/1492916

Köklü, N., Büyüköztürk, Ş. & Çokluk, Ö. (2007). *Sosyal bilimler için istatistik [Statistics for the social sciences]*. Pegem A Yayıncılık.

Kretovics, M. A. (1998). *Outcomes assessment: The impact of delivery methodologies and personality preference on student learning outcomes*. [Unpublished doctoral dissertation], Colorado State University.

Leidner, D.E., & Jarvenpaa, S.L. (1993). The information age confronts education: case studies on electronic classroom. *Information Systems Research*, *4*(1), 24–54.

Levine, A. (2001). The remaking of the American university, *Innovative Higher Education*, *25*, 253- 267.

Lloyd, S.A., Byrne, M.M., & McCoy, T.S. (2012). Faculty-perceived barriers of online education. *Journal of online learning and teaching*, *8*(1), 1-12.

Maloney, E.J., & Joshua Kim, J. (2020, May). Fall Scenario 13: A HyFlex Model. *Inside Higher Ed*. https://www.insidehighered.com/blogs/learning-innovation/fall-scenario-13-hyflex-model

Marvasi, M., Sebastian, G., & Lorenzo, S.L.J. (2019). Fostering researcher identity in STEM distance education: impact of a student led online case study. *FEMS Microbiology Letters*, *366*(6). https://doi.org/10.1093/femsle/fnz068

Miglani, A., & Awadhiya, A.K. (2017). Mobile learning: readiness and perceptions of teachers of Open Universities of Commonwealth Asia, *Journal of Learning for Development-JL4D, 4*(1), 58-71.

Moore, M. G. (1972). Learner autonomy: The second dimension of independent learning. *Convergence, 5(2),* 76–88.

Moore, M.G., & Kearsley, G. (2011). *Distance education: A systems view of online learning*, Belmot: CA: Wadsworth.

Moore, M.G., & Dielh W. (2019). *Handbook of Distance Education,* Routledge.

Kuru, E. (2019). Sosyal bilgiler öğretmen adaylarının dijital okuryazarlık kavramına ilişkin görüşleri [Opinions of social studies teacher candidates on the concept of digital literacy]. *Electronic Turkish Studies*, *14*(3). https://turkishstudies.net/turkishstudies?mod=makale_tr_ozet&makale_id=22563

Orhan, G., & Beyhan, Ö. (2020). Teachers' perceptions and teaching experiences on distance education through synchronous video conferencing during Covid-19 pandemic. *Social Sciences and Education Research Review, 7*(1), 18-44.

Özer, N., & Kır, Ş. (2018). Halk eğitim merkezinde görev yapan öğretmenlerin yetişkin eğitiminde uzaktan eğitimin uygulanabilirliğine ilişkin görüşleri [Opinions of teachers working in public education centers on the applicability of distance education in adult education]. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi*, *4*(4), 69-86. https://dergipark.org.tr/en/pub/auad/issue/42908/519039

Özkul, A.E. (2004, Mayıs). Açık ve Uzaktan Eğitimin Neresindeyiz? [Where are we in Open and Distance Education?], *Mersin Üniversitesi ÖYP-YUUP Uzaktan Eğitim Çalıştayı, Mersin.*

Robitzsch, A. (2021). Sirt: An R package for Supplementary Item Response Theory Models. Version 3.10-118. https://cran.r-project.org/web/packages/sirt/sirt.pdf

Sampson, N. (2003). Meeting the Needs of Distance Learners. *Language Learning & Technology, 7*(3), 103-118.

Seage, S.J., & Türegün, M. (2020). The effects of blended learning on STEM achievement of elementary school students, *International Journal of Research in Education and Science (IJRES), 6*(1), 133-14. https://files.eric.ed.gov/fulltext/EJ1231349.pdf

Seaman, J. (2009). *Online learning as a strategic asset. Volume II: The paradox of faculty voices: Viewsand experiences with online learning*. Washington, DC: Association of Public and Land-grant Universities and Babson Survey Research Group. https://files.eric.ed.gov/fulltext/ED517311.pdf

Sheets, M. (1992). Characteristics of adult education students and factors which determine course completion: a review, *New Horizons in Adult Education, 6*(1), 3-18.

Sousa, O.C., & Florencio Da Silva, R. S. (2020). Contributions of Technology to Distance Learning. How the University will need to reinvent itself to face the challenges of the 21st Century. *Revista Espacıo, 41*(2), 21.

Soydal, İ., Alır, G., & Ünal, Y. (2012). Türk üniversiteleri e-öğrenmeye hazır mı? Hacettepe Üniversitesi Edebiyat Fakültesi Örneği [Are Turkish universities ready for e-learning? Example of Hacettepe University Faculty of Letters]. Ankara: Hacettepe Üniversitesi Bilgi ve Belge Yönetimi Bölümü.

Süer, İ., Kaya, Z., Bülbül, H. İ., Karaçanta, H., Koç, Z., & Çetin, Ş. (2005). Gazi Üniversitesi'nin uzaktan eğitim potansiyeli [Distance education potential of Gazi University], *The Turkish Online Journal of Educational Technology, 4*(1), 107-113.

Şimşek, Ö.F. (2007). *Yapısal eşitlik modellemesine giriş, temel ilkeler ve LİSREL uygulamaları [Introduction to structural equation modeling, basic principles and LISREL applications]*. Ekinoks Yayınları.

Tabachnick, B.G., & Fidell, L.S. (2006). *Using multivariate statistics*. Pearson.

Tavşancıl, E. (2002). *Tutumların ölçülmesi ve SPSS ile veri analizi [Measuring attitudes and data analysis with SPSS]*. Nobel Yayınları.

TEDMEM, (2021). *2020 Education Evaluation Report*, TED. https://tedmem.org/yayin/2020-egitim-degerlendirme-raporu

TEDMEM, (2022). *2021 Education Evaluation Report*, TED. https://tedmem.org/yayin/2021-egitim-degerlendirme-raporu

Telli Yamamoto, S.G., & Altun, D. (2020). Coronavirüs ve çevrimiçi (online) eğitimin önlenemeyen yükselişi [Coronavirus and the unstoppable rise of online education]. *Üniversite Araştırmaları Dergisi, 3*(1), 25-34.

The Turkish Ministry of National Education [TMNE], (2020, July). *"Teachers' Lounge" starts in the TRT-EBA-TV*. http://www.meb.gov.tr/ogretmenler-odasi-kusagi-trt-eba-tvde-basliyor/haber/21224/tr

Tutsun, E. (2020). Açıköğretim hizmeti veren bir üniversitede uzaktan eğitim deneyimi olmayan personelin uzaktan eğitim hakkındaki görüşlerine ilişkin bir durum çalışması [A case study on the views of personnel who do not have distance education experience at a university providing distance education services]. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi*, *6*(3), 1-26. https://dergipark.org.tr/tr/pub/auad/isue/56247/672732

United Nations Educational, Scientific and Cultural Organization [UNESCO], (2020a). *International Task Force on Teachers for Education 2030*. https://teachertaskforce.org/sites/default/files/2020-04/COVID19_PDF_2PAGES_EN.pdf

United Nations Educational, Scientific and Cultural Organization [UNESCO], (2020b). *UNESCO Covid-19 Education Response Education Sector issue notes*. https://unesdoc.unesco.org/ark:/48223/pf0000373338/PDF/373338eng.pdf.multi

Uşun, S. (2006). *Uzaktan eğitim [Distance education]*. Nobel Yayıncılık.

Volery, T., & Lord, D. (2000). Critical success factors in online education, *International Journal of Educational Management, 14*(5), 216 – 223.

Wang J., & Wang X., (2012). *Structural equation modeling applications using Mplus*, Wiley Publication, Higher Education Press.

Wu C.H. (2008). An Examination of the Wording Effect in the Rosenberg Self-Esteem Scale Among Culturally Chinese People, *The Journal of Social Psychology, 148:5*, 535-552, https://doi.org/10.3200/SOCP.148.5.535-552

Yıldırım, Y. (2020). Fatih projesi kapsamında düzenlenen uzaktan hizmet içi eğitimlere yönelik öğretmen görüşlerinin incelenmesi [Examining the opinions of teachers about distance in-service trainings organized within the scope of Fatih project]. *Açıköğretim Uygulamaları ve Araştırmaları Dergisi, 6*(2), 76-90. https://dergipark.org.tr/en/pub/auad/issue/55662/761364

Yılmaz, V., & Çelik, H. (2009). *LISREL ile yapısal eşitlik modellemesi [Structural equation modeling with LISREL]*. Pegem Akademi.

Zhu, X., & Liu, J., (2020). Education in and after Covid-19: immediate responses and long-term visions, *Postdigital Science and Education, 2*, 695-699.

APPENDIX

Scale Items

Factor 1: Challenges faced in distance education
m3. I think, in distance education, my students are distracted.
m6. I can't understand whether students understand the lessons in distance education.
m7. In distance education, I feel like I'm talking by myself.
m10. The active participation rate of students in distance education is lower than that of face-to-face education.
m17. Since there are no exams in distance education, I think my students do not care as much as in face-to-face education.
m23. Not being able to make eye contact with students and not seeing their faces in distance education make it difficult for me to adjust my teaching speed.
m25. I find distance education platforms complicated.
Factor 2: Perceived distance education skills
m14. I think the EBA system is easy to use for the online lessons.
m15. I can use internet resources such as e-books and e-journals as a course material.
m20. I can visualize my online lessons with appropriate pictures in a clear and understandable way more easily.
m21. In online lessons, I can adequately employ multimedia such as graphics, sound and animation.
m22. I think online learning systems are easy to use.
Factor 3: Positive attitudes toward distance education
m4. I prefer distance education to face-to-face education.
m5. I think that distance education enables every student to learn at their own pace.
m8. I think distance education is suitable for student groups at all ages.
m9. I think the time allocated for each subject in distance education is sufficient.
m11. I can cover topics in distance education in depth.
m12. I think distance education is also suitable for disadvantaged students

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 2, 470–491
https://doi.org/10.21449/ijate.974214
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Reliability of Ratings of Multidimensional Fluency Scale with Many-Facet Rasch Model

Cigdem Akin Arikan[1,*], **Pinar Kanik Uysal**[2], **Huzeyfe Bilge**[1], **Kasim Yildirim**[2]

[1]Ordu University, Faculty of Education, Department of Measurement and Evaluation in Education, Türkiye
[2]Ordu University, Faculty of Education, Turkish Language Teaching Department, Türkiye
[3]Kafkas University, Faculty of Education, Turkish Language Teaching Department, Türkiye
[4]Muğla Sıtkı Koçman University, Faculty of Education, Primary School Teaching Department, Türkiye

ARTICLE HISTORY
Received: July 25, 2021
Revised: Feb. 04, 2022
Accepted: Apr. 23, 2022

Keywords:
Rater bias,
Rubric,
Reading fluency,
Prosody,
Many-facet Rasch model,
Rater reliability

Abstract: The aim of this study was to determine whether the reliability of raters was provided by assessing reading prosody using the Multidimensional Fluency Scale (MDFS). The study was completed with a cross-sectional design, and in line with this, the prosodic reading skills of 41 fifth-grade students were rated by elementary school classroom teachers and Turkish language arts teachers using the MDFS. Data obtained from the ratings were analyzed with the many-facet Rasch model (MFRM). When the findings are investigated, the reading prosody rubric used in the research served the purposes of the reading prosody criteria, the sub-dimensions of the rubric could be reliably differentiated, the determined criteria were reliable, and the criteria categories appear to be adequate. Additionally, the severity and leniency of raters were found to differ, and Turkish language arts teachers were found to perform more severe ratings than classroom teachers. It was found that raters were ranked reliably in terms of severity/leniency, and that their levels of severity/leniency differed from each other. Another result obtained is that the prosody criterion that students completed with the most difficulty was phrasing. Therefore, it was concluded that the MDFS is a reliable rubric and that researchers and teachers can reliably use it to assess prosodic reading skills.

1. INTRODUCTION

The interest shown in reading fluency has increased in recent years. One of the most important reasons for this is the understanding of the significant correlation between reading fluency and academic success (Baştuğ & Keskin, 2012; Buck & Torgesen, 2003; Hallman, 2009; National Reading Panel, 2000; Rasinski, 2004; Yıldız et al., 2019). While initially, it was common to deal with speed and accuracy in reading fluency explained with the automaticity theory, in recent years, the definition of fluency has expanded and begun to include different concepts (Godde et al., 2019). According to the accepted view, reading fluency comprises speed, accuracy, and prosody, with researchers (Godde et al., 2019; National Reading Panel, 2000;

*CONTACT: Cigdem Akin-Arikan ✉ akincgdm@gmail.com Ordu University, Faculty of Education, Department of Measurement and Evaluation in Education, Türkiye

e-ISSN: 2148-7456 /

Rasinski, 2004, 2010; Schwanenflugel et al., 2004; Ulusoy et al., 2011) stating the need to deal with these three elements when defining reading fluency (Benjamin & Schwanenflugel, 2010).

There are differences at the point of measurement and assessment when dealing with these three essential elements of reading fluency, with prosody being accepted as relatively more challenging to assess compared to speed and accuracy (Baştuğ, 2021; Valencia et al., 2010). Rubrics in which rater judgments come into play are used in the evaluation of prosody. This situation involves a variety of difficulties. Perhaps the most important of these is the degree to which raters are consistent in giving scores, the degree to which they are severe or lenient when giving scores, and the purpose served by the rubric used. The Multidimensional Fluency Scale (MDFS) (Zutell & Rasinski, 1991), commonly used in Turkey to assess prosodic reading, appears to have been examined for consistency between several raters to show reliability in general (Ceyhan, 2019; Kanık Uysal & Duman, 2020). However, it is accepted that this type of assessment is deficient in many aspects (Eckes, 2015), and it is recommended to note the severity of the raters (Bond & Fox, 2015). Though more comprehensive reliability studies were performed for the original version of the MDFS (Moser et al., 2014; Smith & Paige, 2019), these types of comprehensive analysis are not encountered for the version adapted to Turkish.

1.1. Prosody

Prosody is a comprehensive and well-established term used since very ancient times (Couper-Kuhlen, 1986; Crystal, 2008; Sinambela, 2017; Spafford et al., 1998; VandenBos, 2015; Xu & Liu, 2012). In literature, prosody is defined as the melody of a language (Sinambela, 2017), the flow of rhythm involving intonation, words and sentence length, and the stress patterns of a language (Spafford et al., 1998). It is a phonological feature of speaking related to a phoneme sequence, like stress, intensity or duration rather than a single section (VandenBos, 2015) and is a term used to express variations in pitch, loudness, tempo and rhythm in suprasegmental phonetics and phonology (Crystal, 2008). It is defined as the ability of readers to use phrasing and expression appropriately (Rasinski, 2004) and is a general language term describing rhythmic and tonal features of speech (Dowhower, 1991). While some of these definitions consider prosody as a speech term, some attribute it to a dimension of reading aloud. In addition to speech, prosody has been mentioned frequently in relation to reading skills in recent years and prosodic reading, also called expressive reading, means reading using the voice well with appropriate phrasing and reflecting the emotions in the text (Rasinski, 2004). Prosody includes elements like intonation, rhythm involving syllable-word-sentence length, stress, pitch, and tempo (Crystal, 2008; Dowhower, 1991; Palmer, 2010; Rasinski, 2004; Spafford et al., 1998; VandenBos, 2015; Xu & Liu, 2012).

The importance given by people to prosody, or their attempts to understand based on prosody, begin in infancy. A child develops a sensitivity to the mother tongue and prosodic features used by the mother, and these prosodic features play an essential role in early reading development (Godde et al., 2019). Research shows that even infants younger than 12 months use prosody as a primary clue to syntactic structures and that their babbling contains prosodic features (Kuhn & Stahl, 2013). Based on these results, it is possible to say that prosody is one of the most important elements affecting understanding, even from very young ages. People's contact with prosody begins when they are young and has an effect on their comprehension skills in future periods (Godde et al., 2019; Kuhn & Stahl, 2013). This relationship between prosody and comprehension, between understanding what is read and prosody, is frequently revealed and widely accepted (Çetinkaya et al., 2016; Godde et al., 2019; Schwanenflugel et al., 2004). Prosody has an important place in determining reading competence and identifying reading fluency (Keskin, 2012; Schreiber, 1991). However, the multiple dimensions of prosody and rubrics for measurement require great care in the measurement and assessment process.

Assessment of prosody is more difficult compared to measuring speed and accuracy (Grosjean & Collins, 1979; Moser et al., 2014; Valencia et al., 2010). Prosody involves reading by paying attention to many elements like intonation, stress, pauses, syntax and semantic groups, causing prosody to be the most difficult variable to measure among reading skills (Godde et al., 2019). After many years of measuring speed and accuracy, it was emphasized that measurement of fluency without prosody was not sufficient (Dowhower, 1991; Kuhn, 2007; Schreiber, 1991). After awareness of this deficiency, rubrics were developed to measure prosody. The most common among these are the MDFS (Zutell & Rasinski, 1991) and the Oral Reading Fluency Scale (U.S. Department of Education, 2002), with these two rubrics accepted as being the most commonly used rubrics to assess prosody (Morrison & Wilcox, 2020; Smith & Paige, 2019).

After Allington (1983), Zutell and Rasinski (1991) were the first to develop rubrics to assess prosody. Prepared as a task-specific rubric (Brookhart, 2013), in the MDFS the researchers dealt with three dimensions of phrasing, smoothness and pace, with each dimension organized on four levels. This rubric was updated by Rasinski (2004), and expression and volume was added to bring it to four dimensions. This update was completed to allow a separate assessment of the four basic features included in prosody. This multidimensional rubric comprises the dimensions of 'expression and volume', 'phrasing', 'smoothness' and 'pace'. Statements for each point are included in the rubric and raters perform the assessment in line with these statements. It was shown to be among the best scales to assess prosody (Benjamin et al., 2013) in many studies using this rubric (Aşıkcan, 2019; Morrison & Wilcox, 2020; Overstreet, 2014; Rasinski et al., 2017; Young & Rasinski, 2009).

In spite of the common use of rubrics to assess prosody, a variety of problems are encountered in using rubrics. Zutell and Rasinski (1991) stated that they developed the MDFS for use in class. In other words, the rubric is oriented toward in-class application. Additionally, these rubrics require judgments mediated by the assessor, and the professional experience of those evaluating the reading may affect the scores obtained from the rubric (Moser et al., 2014). To overcome these problems, it is recommended to provide training or directions to raters who will use the rubric (Zutell & Rasinski, 1991). Based on these views, the use of rubrics to assess prosody involves more than just listening to oral reading and giving scores, and requires many precautions to be taken in relation to reliability.

Studies show that interrater agreement is not sufficiently high in assessments made without training (Godde et al., 2017; Haskins & Aleccia, 2014). Though agreement rates were significant in these studies, the significance was not high enough to ensure use in a common fashion. Among studies using the MDFS, studies are encountered where the prosody assessment was excluded due to the lack of statistical agreement between two raters (Bilge, 2019); where there were high levels of agreement between scores given by two experts (Ceyhan, 2019; Kanık Uysal & Duman, 2020; Overstreet, 2014; Paige et al., 2021); and where assessments were made by a single rater without examining interrater reliability (Aşıkcan, 2019; Esmer, 2019; Kaya Tosun, 2019; Kızıltaş, 2019, Rasinski et al., 2017; Zimmerman et al., 2019). Based on these differences in the relevant literature, it is understood that there are different forms used when assessing prosody despite widespread use, and a need to investigate the interrater reliability related to the use of rubrics.

1.2. Rater Reliability

Most measurement processes in behavioral science involve errors; however, this problem is observed more frequently when measurement is made by raters (Shrout & Fleiss, 1979) and the fallibility of human raters has led to serious concerns related to the psychometric quality of scores given to those entering exams (Eckes, 2015). Many studies revealed that in situations where it is not possible to perform automatic rating for assessment of the performance of individuals, evaluation of student responses by several raters would ensure that more

definite/accurate results are obtained. However, in this situation, the scores of an individual are not just linked to their performance or the difficulty of the task, they are also related to rater behavior, or in other words, errors due to the raters. Personal bias error, one of the rater errors, occurs in three different ways. Raters may have a tendency to give lower scores than deserved by the student's performance in severity error (excessive negativity error), they may tend to give higher scores in the leniency error (excessive positivity error) or they may avoid giving low or high scores and give moderate scores, called the central tendency error (McMillan, 2017). If the scores given to the same individual are similar during assessment by several raters, it means reliability is provided or sufficient between raters. However, this situation is not always present in practice. As raters generally comprise an important source of variance, they threaten the validity of inferences made from the results (Eckes, 2015). For this reason, it is important to investigate the effect of raters on the assessment of the performance of individuals. The effects of raters can be identified by two general methods: generalizability theory (G-Theory) and the many-facet Rasch model (MFRM). However, there are some differences between the two approaches. G-theory provides information at the group level, while the MFRM provides individual-level information about all the variability sources (Barkaoui, 2008; Linacre, 1993). The MFRM, one of the item response theory (IRT) models, includes assessments of the effects of other possible sources of systematic error (raters, ratings, tasks, and items) (Sudweeks et al., 2004). The MFRM, from a micro perspective, has the advantages of simultaneously assessing the difficulty of test items, a student's ability, the severity/leniency of raters, and the consistency of scores on the same scale (Li et al., 2021). In addition, the MFRM exceeds G-Theory by presenting quality control fit statistics as well as calculating a measure and a standard error for each source (Linacre, 1993). Using the MFRM, each facet's contribution is analyzed independently of the other facets (Engelhard & Myford, 2003), which allows it to make more accurate estimates than scores obtained with G-Theory. It can be said that the MFRM should be preferred primarily in assessments where it is not possible to score objectively (İlhan, 2015).

As discussed earlier, prosody is a critical component of reading fluency. Teachers and researchers evaluate prosodic skills frequently. While the MDFS was prepared for in-class usage (Zutell & Rasinski, 1991), it is common to use this rubric in scientific research. Since the MDFS depends on human rating, and thus, errors are expected to occur, the question of whether there are significant differences between raters' ratings is important because decisions about prosodic skills are made depending on these assessments. The reason why the MFRM is used for this scale is that it is a stronger psychometric model than classical test theory (Haiyang, 2010) in terms of its ability to detect interactions between different error sources, and is recommended by researchers (Baird et al., 2013) to avoid the limitations of classical approaches. In the literature, there are studies that tested the interrater reliability of assessments performed using prosody rubrics. Moser et al., (2014) examined the interrater reliability of Zutell and Rasinski's (1991) MDFS, and found that the rubric was reliable. In the study by Smith and Paige (2019), it was determined that the MDFS was more reliable compared to the Oral Reading Fluency Scale (U.S. Department of Education, 2002). However, there is no study encountered in the literature revealing how reliability between multiple raters is provided for the version of the MDFS adapted to Turkish. As the use of the reading prosody rubric occurs in the Turkish lesson, the MDFS developed by Zutell and Rasinski (1991), updated by Rasinski (2004) and adapted to Turkish by Yıldız et al., (2009) appears to have been used by researchers in primary school teaching (Aşıkcan, 2019; Ceyhan, 2019; Kaya Tosun, 2019) and Turkish education in secondary schools (Armut & Türkyılmaz, 2017; Kanık Uysal & Duman, 2020). Based on this, in this study the aim was to identify the degree to which interrater reliability was provided for assessment of reading prosody by elementary school classroom teachers and Turkish language arts teachers using the MDFS. In line with this aim, answers to the following questions were sought:

1) Do teachers display differences in terms of severity/leniency when assessing students' prosodic reading?
2) Are there differences in terms of severity/leniency during assessment of students' prosodic reading according to teaching branch?
3) What are the results for task/criterion difficulty analysis related to students' prosodic reading?
4) What are the outcomes of central tendency behavior and bias analysis of raters?

2. METHOD

This research aimed to identify the reliability of results obtained from different raters using the MDFS (Zutell & Rasinski, 1991). In line with the aim of the research, the assessment results for prosodic reading of students by different raters were investigated with the MFRM. Descriptive research, which is a type of quantitative design, was used in the study (Fraenkel & Wallen, 2009).

2.1. Participants

2.1.1. *Students*

Prosodic reading data were obtained from fifth-grade students in a state middle school located in the central district of a metropolitan city in Turkey. The criterion sampling method was used for the selection of students included in the research. The reason for determining the criterion as fifth-grade level was that this class level is known by both elementary school classroom teachers and Turkish language arts teachers. Until the 2012-2013 academic year in Turkey, elementary school classroom teachers continued teaching until fifth grade, and the reading skills of students at this grade level were assessed by elementary school classroom teachers. However, since the 2012-2013 academic year, a 4+4+4 educational system has been implemented, and the fifth grade was moved to the middle school level. For this reason, the reading skills of students in fifth grade are currently assessed by Turkish language arts teachers. To find answers to one of the problems in this research, namely "Are there differences in terms of severity/leniency during assessment of students' prosodic reading according to teaching branch?" the study included both elementary school classroom and Turkish language arts teachers. Comparisons were made between branches in assessing the oral reading prosody of students in fifth grade, known by both departments. In order to meet this criterion, in other words, to include teachers who had taught at fifth-grade level, teachers with at least ten years of teaching experience were included in the study. All fifth-grade students attending the school in which the research was performed were invited, and 41 students from eight different classes who volunteered to participate and whose parents signed consent forms were included in the study. Of the students in the study group, 24 were girls (58%) and 17 were boys (42%).

2.1.2. *Raters*

Ten teachers participated in an assessment of oral reading records. For the determination of the teachers, the criterion sampling method of purposive sampling was used. In line with this, the criterion was that all teachers participating in the study had been employed for at least ten years. The situation leading to this criterion was that oral reading skills of students are assessed by elementary school classroom teachers and by Turkish language arts teachers in middle school. Before the data files were sent to the raters, they were given rater training. The raters were given information about the sub-dimensions of reading prosody, how it is assessed, and what requires attention during assessment, and the rubric to be used was described. In addition to the information given during training, a written form related to the rubric and the elements that require attention during the rating process was prepared and given to the raters. The raters assessed the voice recording for each student using the MDFS and included their scores in an

Excel table. Information related to the demographic characteristics of the raters is given in Table 1.

Table 1. *Information related to raters.*

Rater	Gender	professional experience (years)	Branch
R1	Male	20	Turkish Language Arts
R2	Female	10	Turkish Language Arts
R3	Female	14	Turkish Language Arts
R4	Female	18	Turkish Language Arts
R5	Female	15	Turkish Language Arts
R6	Male	15	Elementary School
R7	Female	17	Elementary school
R8	Female	10	Elementary school
R9	Female	30	Elementary school
R10	Female	17	Elementary school

Table 1 shows that the raters included five Turkish language arts and five elementary school classroom teachers and that their years of experience varied from 10 to 30 years.

2.2. Measurement Tools

In order to identify the prosodic reading skills of students, a narrative text was chosen. The MDFS was used to assess recordings of oral readings of this text.

2.2.1. *Narrative Text*

In order to measure reading prosody, a text was chosen in line with expert opinion from the Turkish textbook used in previous years and permitted by the Ministry of National Education and Board of Education and Discipline (MoNE, 2016). When choosing texts, opinions were sought from three Turkish language arts teachers, three elementary school classroom teachers and an academic in the field of Turkish education. The selected text was a story containing 275 words. When deciding on the type of text, again expert opinion was sought. It was concluded that stories were more suitable in reflecting prosodic reading elements (reflecting mutual dialogue and emotional variations). Students read the text aloud, a voice recording was made for each student, and a one-minute portion of the reading was assessed. In the literature, one-minute voice recording samples were stated to be sufficient for assessment of prosody (Rasinski et al., 2017; Zimmerman et al., 2019), with no significant difference found between assessments of one minute and three minutes (Valencia et al., 2010).

2.2.2. *Multidimensional Fluency Scale*

The MDFS was developed by Zutell and Rasinski (1991), updated by Rasinski (2004) and adapted to Turkish by Yıldız et al. (2009). It comprises four dimensions. These are "expression and volume", "phrasing", "smoothness" and "pace". These are rated from one (1) to four (4) points with a graded rating key. Scores obtained from the four dimensions comprise the total prosody scores, and so the lowest score that can be obtained is 4, while the highest score is 16. The rubric contains statements for each point and raters perform the assessment in line with these statements. For example, for the 'phrasing' dimension, '1' point is equivalent to the statement "reading is monotone, reader does not pay attention to units of meaning or word groups, mostly reads word-by-word", while '4' points are equivalent to the statement "generally reads by paying attention to word groups and units of meaning, reveals the emotional features of expressions in appropriate phrasing". Additionally, if total scores obtained from the rubric at the end of the assessment are less than 10, it means reading is inadequate in prosodic terms and requires development, while scores of 10 or more are accepted as adequate prosodic reading

(Rasinski et al., 2017). Many studies in the research stated that valid and reliable measures were obtained by rating using this rubric (Ceyhan, 2019; Kaya Tosun, 2019; Moser et al., 2014; Rasinski et al., 2017; Rasinski et al., 2009; Valencia et al., 2010; Zimmerman et al., 2019).

2.3. Procedure

The research data were collected in the fall semester of the 2019-2020 educational year. Permission for the research was granted by the Provincial Directorate of National Education (Number: 1 88023 89-44-E.22033 447) and an ethics committee report was obtained (Decision number: 2020-39). Additionally, detailed information was given to parents about the research regarding making voice recordings of the students, and the necessary permission was obtained by signing the 'Parental Consent Form'. Data for the research were collected from fifth-grade students in a state school in the center of the city. The school administration, guidance and counseling service and the Turkish language arts teachers in classes in which data would be collected were given detailed information about the content and aims of the research. When collecting data, care was taken to ensure students were in an environment in which they felt comfortable, with voice recordings made in the school meeting room to ensure a quiet environment. When taking voice recordings, each student was talked to for a few minutes before reading to minimize the student's agitation, and attempts were made to overcome problems with breath and agitation control.

2.4. Data Analysis

The MFRM was used for data analysis related to the MDFS. The MFRM is a member of the Rasch model family. The Rasch model was originally used for dichotomous data, while later, the Rasch model also began to be used for polytomous data. Parameters for the MFRM are expressed on a common scale called a 'logit scale'. The logit scale unit makes it possible to compare units on every facet with others (Linacre, 1994). The Rasch model is linked to the difficulty level of items for the competence levels of individuals. In addition to competence and item difficulty levels, the MFRM is a model paying attention to the potential effects on performance outcomes of performance criteria, measurement time, raters, and other sources of variance and their interactions (Eckes, 2015; Linacre, 2002). Additionally, the MFRM provides information about how well the values predicted by the model created by performance analysis for each individual, rater or task match the expected values (Sudweeks et al., 2004). The MFRM analysis was conducted using the FACETS program (Linacre, 1994). The facets in the MFRM are calibrated simultaneously on a single linear scale. Thus, it is possible to measure the severity or leniency of a rater on the same scale as the difficulty of tasks/items for the competence of individuals (Eckes, 2015).

2.4.1. *Separation index and separation index reliability*

The separation index and reliability are separately calculated for each facet in the model (Schumacker & Smith, 2007). The reliability of raters on a facet represents ratings being different in a reliable way, rather than similar in a reliable way (Haiyang, 2010). For this reason, it should not be considered a measure of the reliability of raters or consistency between raters (Sudweeks et al., 2004). If the aim is to separate individuals in terms of performance, the separation reliability value should be high (Myford & Wolfe, 2003). While the separation index has values from 1 to ∞, the reliability coefficient has values from 0 to 1 (Sudweeks et al., 2004). For individual facets, the reliability of the separation index may be interpreted similarly to the Cronbach alpha coefficient (Engelhard & Myford, 2003; Myford & Wolfe, 2003).

2.2.2. *Fit statistics and chi-square statistic*

For each facet in the research, two statistics are obtained: 'infit' and 'outfit' mean squares. Wright and Linacre (1994) and Linacre (2002) stated that the lower limit for these values was 0.5 and the upper limit was 1.5. Bond and Fox (2015) stated that if the infit and outfit mean

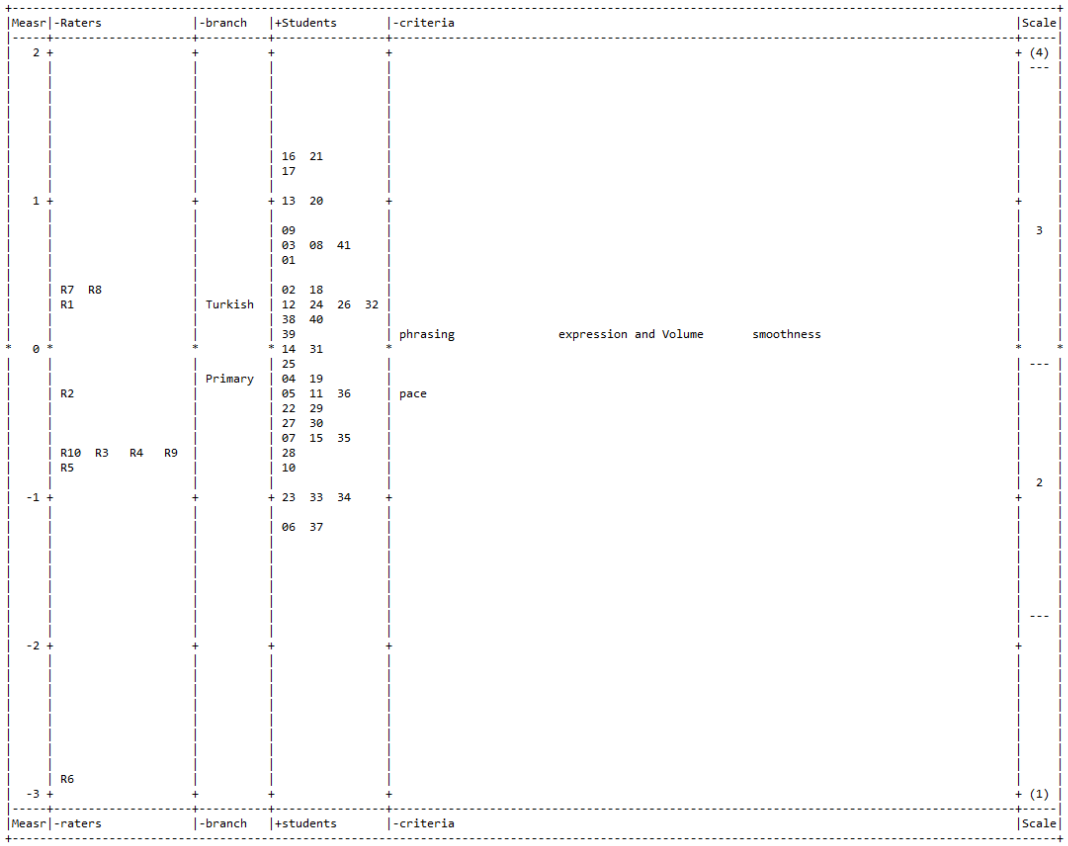
squares values were larger than 1.30, there was no fit and values below 0.70 represented overfitting. Fit values larger than 1 show more variability than expected between the scores of raters, while values below 1 are interpreted as showing less variability than expected (Eckes, 2015). Additionally, Linacre (2011) stated that the Rasch-Cohen kappa value may be used to assess whether raters acted independently or not, and that this value should be close to 0. The chi-square statistic (fixed) provides information about whether there is a significant difference present or heterogeneity between elements/levels of a facet. If the chi-square statistic is significant, it is interpreted that there is a difference between levels of the facet. A significant chi-square statistic for the rater facet shows that at least two raters do not share the same parameter (Eckes, 2005). In this research, there were four facets of rater, branch, student and criterion (dimension). The infit and outfit mean squares statistics, reliability values and separation rates were interpreted for each facet.

3. RESULTS

3.1. Findings of MFRM

This section includes findings obtained from analyzing reading prosody with the MDFS in practice. The logit-scale obtained as a result of the MDFS used by a total of 10 raters as Turkish and elementary school classroom teachers, assessing 41 students and four criteria on three facets is presented in Figure 1.

Figure 1. *Logit Scale for Four Facets*



The first column in Figure 1 shows the logit, the measurement unit of the logit scale. The ability levels of the students included in the study, difficulty level of the criteria, rater branches and rater severity/leniency levels are interpreted based on this measurement unit. The second column in the figure contains measurements belonging to raters and gives the opportunity to make interpretations about the severity/leniency of raters. This means the rater with the highest

logit score in this column performs the most severe rating and the rater with the lowest logit score performs the most lenient rating. When the figure is investigated, the most severe rating was given by the seventh rater (.42), while the most lenient rating was given by the sixth rater (-2.87). The third column of the figure lists the performance for the criteria in terms of the branch of the rater; in other words, it represents the raters' ability to give scores. The logit values for the Turkish teaching branch (.32) were found to be higher than the values for primary teachers (-.16). The fourth column of the figure lists the students from highest to lowest in terms of performance for the criteria found in the graded rating key. Thus, students number 16 and number 21 had the highest performance (1.3) and students number 4 and number 37 had the lowest performance (-1.25). The positive and negative values on the logit scale for student performance of criteria included on the reading prosody scale, in other words the spread over a wide range, shows that students could be well differentiated from each other. The final column of the logit scale lists the difficulty level of the dimensions. Accordingly, the 'phrasing' and 'expression and volume' dimensions (0.14) were the most difficult, while the 'pace' dimension (-.32), where students obtained the highest scores, was the easiest.

The logit scale provides important information about all facets and is included in measurement reporting in order to obtain more detailed information about all facets. Firstly, the findings for the measurement report related to raters are included in Table 2.

Table 2. *Measurement report related to raters.*

Raters	Mean	Measure	Model Standard error	Infit MnSq	Outfit MnSq
7	2.44	.42	.10	1.33	1.33
8	2.44	.41	.10	1.33	1.33
1	2.27	.33	.10	.65	.64
2	2.62	-.30	.10	1.08	1.10
4	2.83	-.68	.11	1.04	1.03
10	3.01	-.69	.11	.80	.79
3	2.84	-.69	.11	.58	.58
9	3.04	-.75	.11	.77	.80
5	2.88	-.77	.11	1.29	1.27
6	3.76	-2.87	.18	.92	1.14
Population		.88	.02	.26	.27
Sample		.93	.02	.28	.28
Model population RMSE = 0.11		S.S.=.88	Separation index = 7.64		Reliability = .98
Model sample RMSE = 0.11		S.S.=.92	Separation index = 8.06		Reliability = .98
Model fixed chi-square = 384.4	*df*=9		*p*=.00		
Model random chi-square = 8.8	*df*=8		*p*=.36		

When Table 2 is investigated, the reliability and separation index values related to the rater facet were .98 and 8.06, respectively. As the separation index approaches zero, the severity/leniency of raters is accepted as being more similar. The separation index reliability value is interpreted as showing how well separated raters are on the rater facet (Eckes, 2015). The high reliability and separation index values can be said to show that raters differed in their rating. In other words, it means there was unwanted variance between raters and this variance contributed to measurement error (Engelhard, 2002). Additionally, the chi-square statistic (χ^2=384.4, *df*=9, *p*=.00) shows that there was a statistically significant difference between raters in terms of severity and leniency. The negative values in the measurement column show that there were raters giving more generous scores compared to others, while positive values show more severe raters. Severe raters have a tendency to assign lower evaluations/scores consistently compared to other raters, while more lenient raters have a tendency to assign higher evaluations/scores (Myford & Wolfe, 2004). In this research, the Rasch-Cohen's kappa statistic

was -.13 and this shows that raters gave consistent scores, though only partially. The infit and outfit mean squares (.58/1.33) obtained in the study appear to be within the desired interval (Linacre, 2014).

Additionally, when the mean total scores given by the raters are examined, apart from one teacher (3.76), the other teachers had a tendency to give category means from 2.44 to 3.04. Accordingly, it is understood that students displayed prosodic skills at moderate and close to adequate levels. The general mean given to students was 2.81, which may be interpreted as showing that teachers viewed the prosodic skills of students as partly adequate. The measurement report related to the branches of the raters is given in Table 3.

Table 3. *Measurement report according to branch of raters.*

Branch	Measure	Standard error	Infit MnSq	Outfit MnSq
Turkish	.32	.05	.93	.92
Primary	-.16	.05	1.06	1.08
Population	.13	.00	.07	.08
Sample	.18	.00	.10	.11
Model population RMSE = .05		S.S.=.16	Separation index = 3.21	Reliability = .91
Model sample RMSE = .05		S.S.=.23	Separation index = 4.65	Reliability = .96
Model fixed chi-square = 22.6	*df*=1	*p*=.000		

When Table 3 is investigated, the reliability and separation index values related to the branch of the raters were found to be .96 and 4.65, respectively. The reliability and separation index values show that the scores given by raters differed in terms of branch. Additionally, the chi-square test results (χ^2=22.6, *df*=1, *p*=.00) reflect the significant difference between scores given by Turkish and primary teachers. According to Table 3, the infit (.93/1.06) and outfit (.92/1.08) mean squares appear to be within the desired interval. The measurement report related to the student facet is given in Table 4.

In Table 4, the reliability and separation index values for the student facet were .90 and 2.97, respectively. Additionally, the significant results of the chi-square test (χ^2=370.8, *df*=40, *p*=.00) and the high separation index and reliability indicate the students displayed differences in reading prosody skill levels. Additionally, when the infit and outfit mean squares are investigated, only individuals numbered 2, 3, 33 and 39 had fit values larger than 1.5, in other words, outside the accepted interval. Linacre (2003) stated that infit and outfit mean squares between 1.5 and 2.0 were not productive but not harmful, while values above 2.0 disrupted the model. For this reason, these individuals did not break the model. The measurement report related to the criterion facet is given in Table 5.

According to Table 5, the reliability and separation index values related to the criterion facet were .90 and 3.02, respectively. Additionally, when the 'there is no difference between criteria difficulty levels' hypothesis was tested with the chi-square test, it was significant (χ^2=29.3, *df*=3, *p*=.00). This means that the tasks on the MDFS differed in terms of difficulty levels.

Table 4. *Measurement report for students.*

Ind. No	Scores	Logit Score	S.E.	Infit MnSq	Infit MnSq	I. no	Scores	Logit Score	S.E.	Infit MnSq	Infit MnSq
16	137	1.30	.26	1.06	1.49	25	112	.21	.2	1.35	1.31
21	137	1.30	.26	.93	1.41	4	109	.21	.2	.94	.96
17	136	1.23	.26	1.24	1.10	19	108	.21	.21	.69	.66
13	132	.98	.25	.96	.86	5	106	.21	.21	1.14	1.10
20	132	.98	.25	.96	.89	11	106	-.33	.21	.86	.84
9	129	.80	.25	.98	.1.03	36	106	-.33	.21	1.25	1.23
3	128	.74	.24	1.52	1.55	22	105	-.38	.21	0.6	0.59
8	128	.74	.24	.53	.62	29	105	-.38	.21	1.18	1.19
41	127	.69	.24	.82	.77	30	103	-.46	.21	.79	.77
1	126	.63	.24	1.37	1.35	27	102	-.51	.21	.77	.78
2	122	.47	.23	1.58	1.64	7	101	-.55	.21	.97	.96
18	121	.37	.23	.49	.64	15	101	-.55	.21	.65	.66
24	120	.32	.23	.62	.57	35	101	-.55	.21	.74	.72
12	119	.27	.22	.86	.82	28	98	-.68	.21	.68	.68
26	119	.27	.22	.49	.47	10	95	-.81	.21	.75	.76
32	119	.27	.22	1.25	1.28	23	91	-.98	.21	.71	.68
38	118	.12	.22	1.22	1.17	33	91	-98	.21	1.91	1.83
40	117	.17	.22	.93	.94	34	90	-1.03	.21	.82	.81
39	116	.13	.22	1.82	1.73	6	85	-1.25	.21	.47	.48
14	114	.03	.22	.81	.76	37	85	-1.25	.21	.1.26	1.26
31	113	-.02	.22	.58	.56						

Model population RMSE = .22	S.S.=.65		Separation index = 2.93	Reliability = .90
Model sample RMSE = .22	S.S.=.66		Separation index = 2.97	Reliability = .90
Model fixed chi-square = 370.8	*df*=40	*p*=.00		
Model random chi-square = 36.02	*df*=39	*p*=.60		

Ind. No: Individual Number, S.E.: Standard Error

Table 5. *Measurement report for criterion facet.*

Criteria	mean	Logit	S.E.	Infit MnSq	Outfit MnSq
2 phrasing	2.74	.14	.06	1.09	1.14
3 smoothness	2.74	.13	.07	.91	.94
1 expression and volume	2.78	.06	.07	.91	.93
4 pace	2.98	-.32	.07	1.05	1
Mean	2.81	.00	.07	.99	1
Population		.19	.00	.08	.08
Sample		.22	.00	.09	.10

Model population: RMSE = .07	S.S.=.18	Separation = 2.57	Reliability = .87
Model sample: RMSE = .07	S.S.=.21	Separation = 3.02	Reliability = .90
Model fixed chi-square = 29.3, *df*=3, *p*=.00			
Model random chi-square = 2.7, *df*=2, *p*=.25			

3.2. Central Tendency Behavior

One of the rater errors frequently encountered with scores given with a graded key is central tendency behavior. Central tendency behavior indicates that when raters are assigning scores they tend to avoid giving high or low scores and give central scores. Values related to the rating categories (from 1-4) in the graded rating key used for this were investigated. The measurement report related to rating categories is presented in Table 6.

Table 6. *Statistics related to scale structure in rating categories.*

Branch	Category	*f*	%	Mean measure	Outfit MnSq	Rasch–Andrich threshold value	
						Measure	S.E.
Primary	1	74	9%	-.30	.8		
	2	162	20%	-.07	.6	-.84	.13
	3	326	40%	.61	1	-.37	.09
	4	258	31%	1.46	.9	1.21	.09
Turkish	1	74	9%	-0.36	.9		
	2	250	30%	-0.09	.7	-1.34	.13
	3	356	43%	0.39	.8	-0.14	.08
	4	140	17%	0.73	1	1.47	.10

When the table is investigated, the outfit mean squares varied from .6 to 1; in other words, they were within the desired interval (Linacre, 2014). Additionally, as the rating categories increase (from 1 to 4), a monotonous increase in threshold values is expected (Eckes, 2015). It appeared that the Rasch-Andrich threshold values monotonously increased and that all values were smaller than the logit value 5. When the frequency and percentage values related to rating categories are investigated on the table, it was identified that elementary school classroom teachers used rating categories 3 and 4, and that Turkish language arts teachers used rating categories 2 and 3 more often.

3.3. Bias Interaction

If the *t* values obtained from the interaction tables are outside the ±2 interval, they should be investigated for interaction effects. When the branch interaction of raters is investigated, the t values were within the desired interval. However, there was no statistically significant interaction effect according to branch. This study included 10 raters and 4 criteria. For this reason, there were a total of 40 interactions. The table related to the bias interaction according to criteria and raters is given in the appendix (Appendix). When the findings obtained for the rater-criteria interaction are investigated, the fifth, seventh and eighth raters were identified to have t values for the pace criterion outside the ±2 interval, while the third and sixth raters had *t* values for expression and volume outside the ±2 interval. Negative *t* values for the scores of these raters show that their scores were lower than expected, while positive values show that their scores were higher than expected. In other words, there was a difference between the scores expected and the scores observed for the rating by these raters, and bias was present. The eighth, seventh and sixth raters had positive bias and were more lenient raters, while the third and fifth raters had negative bias and were more severe raters. As the t values for the other raters were within the expected interval, rating bias can be ignored. Additionally, the interaction effect for rater and criteria facets was statistically significant ($\chi^2 = 58.1$, $df = 40$, $p = .03$).

4. DISCUSSION and CONCLUSION

The research aimed to investigate the reliability of results obtained with the MDFS assessed by elementary school classroom teachers and Turkish language arts teachers with the MFRM. When the results related to the first problem in the research, “Do teachers display differences in terms of severity/leniency when assessing students’ prosodic reading?” are investigated, the rating behavior of the raters included in the study was reliable, they were reliably ranked in terms of severity and leniency, and their severity/leniency levels were different from each other. When examined from the perspective of evaluators, the difference in severity/leniency levels is an unwanted situation and will restrict the ability of evaluators to take each other’s places (Eckes, 2015). Linacre (2012) emphasized that infit and outfit mean squares smaller than 0.5 show overfitting of the model and give misleading results, and indicate that evaluators did not use the full interval in the rubric. Additionally, values between 0.5 and 1.5 indicate that the values are efficient for measurement. In the results of the research, though there were

differences between teachers in terms of severity and leniency, the infit and outfit mean squares were within the values proposed by Linacre (2012). It was concluded that the sixth rater was the most generous and the seventh rater was that strictest among raters. Goodwin (2016) stated that rating judgments of raters may be affected by the examples rated previously. Additionally, halo effects, described as bias in rating caused by different aspects other than the judged dimension of a person, can lead to misjudgments (American Psychological Association, 2015). In this study, therefore, raters may have been affected by the reading performance of the previous student or by the halo effect when evaluating students. This may have caused differences in the severity/leniency levels of raters.

When the results related to the second problem in the research, "Are there differences in terms of severity/leniency during assessment of students' prosodic reading according to teaching branch?" are investigated, the two raters who were the most severe and most lenient were observed to be classroom teachers. In other words, raters who were Turkish language arts teachers performed more consistent and similar ratings when giving scores for the rubric, while classroom teachers gave different ratings compared to each other. This situation may have caused the rating reliability of raters to be low. When choosing evaluators, it is necessary to find those with appropriate educational backgrounds and experience in the field (Myford & Wolfe, 2003). The prejudices, attitudes, and personality traits of evaluators and the purpose of the assessment may cause a tendency to evaluate more severely (Eckes, 2015). Classroom teachers giving scores at the extremes compared to Turkish language arts teachers may be interpreted with a variety of variables. Turkish language arts teachers take many courses directly and indirectly related to language skills during undergraduate education, while classroom teachers take courses in different areas like mathematics, music, etc. In fact, Taşkaya and Muşta (2008) identified that classroom teachers felt they were inadequate and that the education they received was inadequate with regard to Turkish teaching. In this situation, it may be expected that the general knowledge and judgments related to language skills of classroom teachers will be different to those of Turkish language arts teachers. Some studies in the literature showed differences (Coşkun & Coşkun, 2014; Doğan, 2013) and similarities (Benzer & Eldem, 2013; Doğan, 2013; Saracoğlu et al., 2011) in terms of a variety of features between Turkish language arts teachers and classroom teachers. However, there are findings showing that teachers in both branches do not use rubrics often (Acar & Anıl, 2009; Benzer & Eldem, 2013). It may be considered that when teachers perform assessments with rubrics that they do not use much, their lack of familiarity with these tools may affect the judgments made and may lead to more personal behavior when giving scores. This situation may be interpreted as showing that teachers do not have adequate experience with the use of rubrics. Stevens and Levi (2005) stated that as experience is gained with rubrics, reliable ratings will increase. Mathson et al. (2006) considered that the lack of training on the assessment and teaching of fluency limits the use of rubrics in the classroom. From this perspective, the difference in scores between teachers may be explained by the lack of experience related to rubrics. It is possible that another source of difference between scores is the students they are in constant contact with in both branches. While the Turkish curriculum expects students to have upper-level reading skills, the Turkish curriculum used by classroom teachers requires more basic skills. In this situation, teachers have different expectations and it is probable that this difference was reflected in the assessment of prosody.

When the results related to the third problem in the research, "What are the results for task/criterion difficulty analysis related to students' prosodic reading?" are investigated, the prosody criteria can be ranked in order from more difficult to easier for students as phrasing, smoothness, expression and volume, and pace. This situation overlaps with the ranking accepted in the literature. According to researchers, students begin with letter-sound relationships and move toward higher units (for example, units of meaning), before completing

accurate, paced and prosodic reading in order (Baştuğ, 2021; Keskin, 2012; Mathson et al., 2006; Samuels, 2006). The findings in this study show that students received the highest scores for the pace dimension on the rubric. In the study by Godde et al. (2019), the dimension of prosody completed with the most difficulty was intonation and expressive reading (Godde et al., 2019). Considering that intonation and reading in meaningful units are closely related to each other (Godde et al., 2019), students who read quickly were determined not to understand the text because reading in meaningful units requires implicit clues in the text to be solved by readers (Rasinski, 2004; Schreiber, 1991). Based on this, the degree of difficulty emerging in student assessment according to the MDFS by teachers is supported by previous findings. However, the findings conflict with the finding of Godde et al. (2019) that expression was the most difficult. In this study, 'expression and volume' (mean 2.78) was the dimension with the second highest scores received by students and was easier than smoothness (mean 2.79). The lack of a large difference between these may be interpreted as showing that the difficulty of the two dimensions may occasionally change places; in all cases, the most difficult dimension was reading with intonation and meaningful groups, while pacing was the easiest dimension. In fact, Daane et al. (2005) stated that students who read quickly may develop reading skills for word groups.

When the results related to the fourth problem in the research, "What are the outcomes of central tendency behavior and bias analysis of raters?" are investigated, when the mean total scores given by the raters are examined, the general mean was 2.81. The classroom teachers were found to use categories 3 and 4 more often, while Turkish language arts teachers used categories 2 and 3 more often. In other words, it may be said that Turkish language arts teachers chose central categories as raters. One of the reasons for this may be that when raters gave scores to students, they saw students as having partially adequate prosodic skills; in other words, the raters displayed central tendency behavior (Myford & Wolfe, 2004). Results from assessments by teachers using rubrics are similar to the study by Daane et al. (2005). In the study by Daane et al. (2005), 83% of students were grouped in the 2nd and 3rd categories, dominantly in the 3rd category, and scores were closer to the 3rd category. For this reason, the scores given by teachers to students were similar to previous findings in terms of distribution. Another result of the research is that the eighth, seventh and sixth raters had positive bias, while the third and fifth raters had negative bias. The reliability of these raters for assessment of prosodic reading by students was lower compared to other raters. The findings obtained in the research are consistent with the results of other studies researching rater bias (Baştürk & Işıkoğlu, 2008; Köse et al., 2016; Özbaşı & Kumandaş-Öztürk, 2020; Şata & Karakaya, 2020; Yüzüak et al., 2015). Goodwin (2016) stated that additional training would be beneficial for rater bias. When the findings obtained in the research are investigated, it was concluded that the reading prosody rubric used in the research served the purpose of measuring reading prosody of students, the sub-dimensions on the rubric could reliably differentiate, the criteria determined were reliable, and the criteria categories were suitable and adequate for measurement. Another result is that the prosody criterion where students experienced the most difficulty was phrasing. When examined generally, it was concluded that the MDFS is a reliable rubric for use by researchers and teachers to evaluate prosodic reading skills. This result overlaps with the findings of two studies found in the literature using the generalizability theory to determine the reliability of the scale (Moser et al., 2014; Smith & Paige, 2019). Additionally, it may be said that it is necessary to train people for rating prosody using the MDFS (Bilge, 2019; Erguvan & Dünya, 2020; Kaya Uyanık et al., 2019; Smith & Paige, 2019; Zutell & Rasinski, 1991); however, training may not be adequate all the time (Barrett, 2001; Eckes, 2015; Yan, 2014). For this reason, to ensure the reliability of MDFS ratings, the use of at least two but preferably three texts (Moser et al., 2014), and the presence of at least two raters (Smith & Paige, 2019) are recommended. However, the reliability obtained for measurements with two raters may be

seriously misleading and it should not be forgotten that high rater reliability obtained with two raters does not always mean accurate rating (Eckes, 2015). For this reason, more reliable results will be obtained with the MFRM so as not to ignore the severity of raters in adapted or prepared rubrics; otherwise, the problems that could arise are actually ignored (Bond & Fox, 2015). In this study, the rubric developed by Zutell and Rasinski (1991), updated by Rasinski (2004) and adapted to Turkish by Yıldız et al. (2009) was used. Similar studies may investigate rater behavior using different rubrics/scales measuring prosody.

In this research, a fatigue effect may be present in the results obtained by raters evaluating forty students. For this reason, future studies may perform investigations on reading data obtained with more texts and fewer students. In this research, the rater reliability of the MDFS was researched. In addition to elements contributing to the literature in this way, there are a number of limitations of the study, including assessment of reading prosody with one text, and the inclusion of forty students and ten raters in the study.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Ordu University, 27.05.2020, 2020-39.

Authorship Contribution Statement

Cigdem Akin Arikan: Investigation, Review of Literature, Introduction, Methodology, Discussion and Conclusion. **Pinar Kanik Uysal**: Investigation, Introduction, Review of Literature, Methodology, Discussion and Conclusion. **Huzeyfe Bilge:** Introduction, Review of Literature, Discussion and Conclusion. **Kasim Yildirim:** Supervision.

Orcid

Cigdem Akin Arikan https://orcid.org/0000-0001-5255-8792
Pinar Kanik Uysal https://orcid.org/0000-0003-1208-9535
Huzeyfe Bilge https://orcid.org/0000-0001-7664-488X
Kasim Yildirim https://orcid.org/0000-0003-1406-709X

REFERENCES

Acar, M., & Anıl, D. (2009). Sınıf öğretmenlerinin performans değerlendirme sürecindeki değerlendirme yöntemlerini kullanabilme yeterlilikleri, karşılaştıkları sorunlar ve çözüm önerileri [Classroom teacher evaluation methods to use in the performance assessment process qualification of able, they comparison problems and solution proposals]. *Journal of TUBAV Science, 2*(3), 354-363.

American Psychological Associaton. (2015). Halo effect. In APA dictionary of psychology (2nd ed., p. 667).

Allington, R.L. (1983). Fluency: the neglected reading goal. *The Reading Teacher*, *36*(6), 556-561.

Armut, M., & Türkyılmaz, M. (2017). Ortaokul öğrencilerinin okuma becerileri üzerine bir inceleme [An investigation on reading skills of middle school students]. *Erzincan University Journal of Education Faculty, 20*(1), 217-236. https://doi.org/10.17556/erziefd.330587

Aşıkcan, M. (2019). *Üçüncü sınıf öğrencilerinin akıcı okuma becerilerinin geliştirilmesine yönelik bir eylem araştırması* [*An action research on ımproving fluent reading skills of third-grade primary school students*] [Unpublished doctoral dissertation]. Necmettin Erbakan University.

Baird, J.A., Hayes, M., Johnson, R., Johnson, S., & Lamprianou, I. (2013). Marker effects and examination reliability. A Comparative exploration from the perspectives of

generalisability theory, Rash model and multilevel modelling. Oxford: University of Oxford for Educational Assessment. Retrieved from http://dera.ioe.ac.uk/17683/1/2013-01-21-marker-effects-and-examinationreliability.pdf

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes* [Doctoral Dissertation, Available from ProOuest Dissertations and Theses database]. UMI No: 304360302.

Barrett, S., (2001). The impact of training on rater variability. *International Education Journal, 2*(1), 49-58.

Baştuğ, M., (2021). *Akıcı okumayı geliştirme: kavramlar, uygulamalar, değerlendirmeler [Developing reading fluency: concepts, practices, assessments]*. Pegem Akademi Yayıncılık

Baştuğ, M., & Keskin, H.K. (2012). Akıcı okuma becerileri ile anlama düzeyleri (basit ve çıkarımsal) arasındaki ilişki [The relationship between fluent reading skills and comprehension level (literal and ınferential)]. *Ahi Evran University Journal of Kırşehir Education Faculty, 13*(3), 227-244.

Baştürk, R., & Işıkoglu, N. (2008). Analyzing process quality of early childhood education with many facet rash measurement model. *Educational Sciences:Theory and Practice, 8*(1), 25-32.

Benjamin, R.G., & Schwanenflugel, P.J. (2010). Text complexity and oral reading prosody in young readers. *Reading Research Quarterly, 45*(4), 388-404. https://doi.org/10.1598/RRQ.45.4.2

Benjamin, R.G., Schwanenflugel, P.J., Meisinger, E.B., Groff, C., Kuhn, M.R., & Steiner, L. (2013). A spectrographically grounded scale for evaluating reading expressiveness. *Reading Research Quarterly, 48*(2), 105-133. https://doi.org/10.1002/rrq.43

Benzer, A., & Eldem, E. (2013). Türkçe ve edebiyat öğretmenlerinin ölçme ve değerlendirme araçları hakkında bilgi düzeyleri [Level of the information about Turkish and literature teachers' measurement and assessment materials]. *Kastamonu Education Journal, 21*(2), 649-664.

Bilge, H. (2019). *Okuma, yazma ve konuşma akıcılık ile okuduğunu anlama ve kelime hazinesi arasındaki ilişki [The relationships between reading, writing and speaking fluencies, reading comprehension and vocabulary*] [Unpublished doctoral dissertation]. Gazi University.

Bond, T.G., & Fox, C.M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (3th ed.). Routledge.

Brookhart, S. M. (2013). *How to Create and Use Rubrics for Formative Assessment and Grading*. ASCD.

Buck, J., & Torgesen, J. (2003). The relationship between performance on a measure of oral reading fluency and performance on Florida comprehensive assessment test. FCRR Technical Report# 1. *Florida Center for Reading Research.*

Ceyhan, S. (2019). *Etkileşimli sesli okumanın öğrencilerin okuduğunu anlama, okuma motivasyonu ve akıcı okumalarına etkisi [The effect of interactive reading aloud on the reading comprehension, reading motivation and reading fluency of students]* [Unpublished doctoral dissertation]. Gazi University.

Coşkun, E., & Coşkun, H. (2014). İlkokul ve ortaokullardaki bitişik eğik yazı uygulamalarına ilişkin öğretmen, öğrenci ve veli görüşleri [teachers', students' and parents' views on cursive italic handwriting]. *Mustafa Kemal University Journal of Social Sciences Institute, 11*(26), 209-223.

Couper-Kuhlen, E. (1986). *An Introduction to English Prosody*. Edward Arnold.

Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics* (6th ed.). Blackwell Publishing.

Çetinkaya, Ç., Ateş, S., & Yıldırım, K. (2016). Prozodik okumanın aracılık etkisi: Lise düzeyinde okuduğunu anlama ve akıcı okuma arasındaki ilişkilerin incelenmesi [The mediation effect of reading prosody: exploring the relations between reading fluency and reading comprehension at high school level]. *Turkish Studies, 11*(3). https://doi.org/10.7827/TurkishStudies.9339

Daane, M.C., Campbell, J.R., Grigg, W.S., Goodman, M.J., & Oranje, A. (2005). *The nation's report card: fourth-grade students reading aloud: NAEP 2002 special study of oral reading*. Washington, D.C.: U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics.

Doğan, B. (2013). Türkçe ve sınıf öğretmenlerinin okuma güçlüğüne ilişkin bilgileri ve okuma güçlüğü olan öğrencileri belirleyebilme düzeyleri [Determining Turkish language and elementary classroom teachers' knowledge on dyslexia and their awareness of diagnosing students with dyslexia]. *Research in Reading & Writing Instruction, 1*(1), 20-33.

Dowhower, S.L. (1991). Speaking of prosody: fluency's unattended bedfellow. *Theory Into Practice, 30*(3), 165-175. https://doi.org/10.1080/00405849109543497

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly: An International Journal, 2*(3), 197-221. https://doi.org/10.1207/s15434311laq0203_2

Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments* (Vol. 22). Peter Lang Edition. https://doi.org/10.1080/15366367.2018.1516094

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. A. Tindal & T. M. Haladyna (Eds.), *Large scale assesment for all students: Validity, technical adequacy, and implementation* (pp. 261-287). Erlbaum.

Engelhard, G., & Myford, C.M. (2003). Monitoring faculty consultant performance in the advanced placement English Literature and composition program with a many-faceted Rasch model. *ETS Research Report Series, 2003*(1), i-60.

Erguvan, İ.D., & Dünya, B.A. (2020). Analyzing rater severity in a freshman composition course using many facet Rasch measurement. *Language Testing in Asia, 10*(1), 1-20. https://doi.org/10.1186/s40468-020-0098-3

Esmer, B. (2019). *Okuduğunu anlama ile akıcı okuma, okur benlik algısı, okumaya adanmışlık ve okuyucu tepkisi ilişkileri [Direct and inferential relations among reading comprehension, silent and oral reading fluency, reading self-concept, reading engagement and response to picturebooks]* [Unpublished doctoral dissertation]. Gazi University.

Fraenkel, J.R., & Wallen, N.E. (2009). *How to Design and Evaluate Research in Education* (7th ed.). McGraw-Hill.

Godde, E., Bailly, G., Escudero, D., Bosse, M.L., & Gillet-Perret, E. (2017). Evaluation of reading performance of primary school children: objective measurements vs. subjective ratings.*WOCCI 2017-6th Workshop on Child Computer Interaction*, Nov 2017, Glasgow, United Kingtom.

Godde, E., Bosse, M.L., & Bailly, G. (2019). A review of reading prosody acquisition and development. *Reading and Writing, 33*(2), 399-426. https://doi.org/10.1007/s11145-019-09968-1

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

Grosjean, F., & Collins, M. (1979). Breathing, pausing and reading. *Phonetica, 36*(2), 98-114. https://doi.org/10.1159/000259950

Haiyang, S. (2010). An application of classical test theory and many facet Rasch measurement in analyzing the reliability of an English test for non-English major graduates. *Chinese Journal of Applied Linguistics, 33*(2), 87-102.

Hallman, J. (2009). Reading aloud: comprehending, not word calling. In R. Stone (Ed.), *Best practices for teaching reading: what award-winning classroom teachers do* (pp. 39-43). Corwin Press.

Haskins, T., & Aleccia, V. (2014). Toward a reliable measure of prosody: an investigation of rater consistency. *International Journal of Education and Social Science, 1*(5), 102-112.

İlhan, M. (2015). *Standart ve solo taksonomisine dayalı rubrikler ile puanlanan açık uçlu matematik sorularında puanlayıcı etkilerinin çok yüzeyli rasch modeli ile incelenmesi [The identification of rater effects on open-ended math questions rated through standard rubrics and rubrics based on the SOLO taxonomy in reference to the many facet Rasch model]* [Unpublished doctoral dissertation]. Gaziantep University.

Kanık Uysal, P., & Duman, A. (2020). The effects of fluency-oriented reading instruction on reading skills. *Pegem Journal of Education and Instruction, 10*(4), 1111–1146. https://doi.org/10.14527/pegegog.2020.034

Kaya Tosun, D. *(2019). Okuma çemberlerinin okuduğunu anlama, akıcı okuma, okuma motivasyonu ve sosyal beceriler üzerindeki etkisi ve okur tepkilerinin belirlenmesi [The effect of literature circles on reading comprehension, reading fluency, reading motivation and social skills and exploring of reader responses]* [Unpublished doctoral dissertation]. Gazi University.

Kaya Uyanık, G., Güler, N., Taşdelen Teker, G., & Demir, S. (2019). The analysis of elementary science education course activities through many-facet Rasch model. *Kastamonu Education Journal, 27*(1), 139-150. https://doi.org/10.24106/kefdergi.2417

Keskin, H.K. (2012). *Akıcı okuma yöntemlerinin okuma becerileri üzerindeki etkisi [Impact of reading fluency methods on reading skills]* [Unpublished doctoral dissertation]. Gazi University.

Kızıltaş, Y. (2019). *Ana dili farklı ilkokul öğrencilerinin akıcı okuma ve okuduğunu anlama becerilerinin incelenmesi [The study of reading fluency and reading comprehension skills of primary school whose mother tongue is different]* [Unpublished doctoral dissertation]. Gazi University.

Köse, İ.A., Usta, H.G., & Yandı, A. (2016). Sunum yapma becerilerinin çok yüzeyli Rasch analizi ile değerlendirilmesi [Evaluation of presentation skills by using many facets rasch model]. *Abant İzzet Baysal University Journal of Faculty of Education*, *16*(4), 1853-1864.

Kuhn, M.R. (2007). Effective oral reading assessment (or why round robin reading doesn't cut ıt). In J.R. Paratore & R.L. McCormack (Ed.), *Classroom literacy assessment: making sense of what students know and do* (pp. 101-112). The Guilford Press.

Kuhn, M.R., & Stahl, S.A. (2013). Fluency: developmental and remedial practices-revisited. In D.E. Alvermann, N.J. Unrau, & R.B. Ruddell (Ed.), *Theoretical models and processes of reading* (pp. 385-412). International Reading Association.

Li, G., Pan, Y., & Wang, W. (2021). Using generalizability theory and many-facet Rasch model to evaluate in-basket tests for managerial positions. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.660553

Linacre, J.M. (1993). Generalizability theory and many-Facet Rasch measurement, in *Paper presented at the Annual Meeting of the American Educational Research Association* (Atlanta, GA).

Linacre, J.M. (1994). *Many-Facet Rasch Measurement*. Chicago: Mesa Press.

Linacre, J.M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J.M. (2003). Size vs. significance: Infit and outfit mean-square and standardized chi-square fit statistic. *Rasch Measurement Transactions, 17*, 918.

Linacre, J.M. (2011). *Facets computer program for Many-Facet Rasch Measurement*. https://www.winsteps.com/facets.htm#:~:text=Facets%20is%20designed%20to%20handle,further%20measurement%20and%20structural%20facets

Linacre, J.M. (2012). *Many-facet Rasch measurement: Facets tutorials*. http://www.winsteps.com/tutorials.htm

Linacre, J.M. (2014). *A user's guide to FACETS Rasch-model computer programs*. http://www.winsteps.com/a/facets-manual.pdf

Mathson, D.V., Allington, R.L., & Solic, K.L. (2006). Hijacking fluency and instructionally informative assessments. In T. Rasinski, C. Blachowicz, & K. Lems (Eds.), *Fluency instruction: research-based best practices* (pp. 106-119). The Guilford Press.

McMillan, J.H. (2017). *Classroom Assessment. Principles and Practice that Enhance Student Learning and Motivation* (7th ed.), Pearson.

MoNE. (2016). *5. Sınıf Türkçe ortaokul ders kitabı [Turkish secondary school textbook for 5th graders*]. Devlet Kitapları.

Morrison, T.G., & Wilcox, B. (2020). Assessing expressive oral reading fluency. *Education Sciences, 10*(59). https://doi.org/10.3390/educsci10030059

Moser, G.P., Sudweeks, R.R., Morrison, T.G., & Wilcox, B. (2014). Reliability of ratings of children's expressive reading. *Reading Psychology, 35*(1), 58-79. https://doi.org/10.1080/02702711.2012.675417

Myford, C.M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386-422.

Myford, C.M., & Wolfe, E.W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189-227.

National Reading Panel. (2000). *Teaching children to read: an evidence-based assessment of the scientific research litearature on reading and its implications for reading instruction: reports of the subgroups*. Washington, DC: National Institute of Child Health Human Development.

Overstreet, T.B. (2014). *The effect of prosody instruction on reading fluency and comprehension among third-grade students* [Unpublished doctoral dissertation]. Andrews University.

Özbaşı, D. & Kumandaş-Öztürk, H. (2021). Öğretim materyallerinin çok yüzeyli Rasch analiziyle değerlendirilmesi [Evaluatıon of teaching materials with many-facet Rasch analysis]. *Trakya Journal of Education*, 11(1), 187-200.

Paige, D.D., Smith, G., Rupley, W., & Wells, W. (2021). Reducing high-attaining readers to middling: the consequences of inadequate foundational skills instruction in a high-ses district. *Literacy Research and Instruction*, *60*(1), 81-106. https://doi.org/10.1080/19388071.2020.1780653

Palmer, M.L. (2010). *The relationship between reading fluency, writing fluency, and reading comprehension ın suburban third-grade students* [Unpublished doctoral dissertation]. San Diego State University.

Rasinski, T. (2004). *Assessing Reading Fluency*. Honolulu, Hawai: Pasific Resources for Education and Learning.

Rasinski, T. (2010). *The Fluent Reader*. Scholastic.

Rasinski, T., Paige, D.D., Rains, C., Stewart, F., Julovich, B., Prenkert, D., Rupley, W.H., & Nicholas, W.D. (2017). Effects of intensive fluency instruction on the reading proficiency of third-grade struggling readers. *Reading and Writing Quarterly, 33*(6), 519-532. https://doi.org/10.1080/10573569.2016.1250144

Rasinski, T., Rikli, A., & Johnston, S. (2009). Reading fluency: More than automaticity? More than a concern for the primary grades?. *Literacy Research and Instruction, 48*(4), 350-361. https://doi.org/10.1080/19388070802468715

Samuels, S.J. (2006). Reading fluency:ıts past, present, and future. In T. Rasinski, C. Blachowicz, & K. Lems (Ed.), *Fluency instruction research-based best practices* (pp. 7-20). The Guildgord Press.

Saracoğlu, S., Dedebali, N.C., Dinçer, B., & Dursun, F. (2011). Sınıf, fen ve teknoloji ile Türkçe öğretmenlerinin öğretmen stillerinin incelenmesi [Investigatıon of teaching styles of primary school, science and techonology, Turkish teachers]. *Education Sciences, 6*(3), 2313-2327.

Schreiber, P.A. (1991). Understanding prosody's role in reading acquisition. *Theory into Practice, 30*(3), 158-164. https://doi.org/10.1080/00405849109543496

Schumacker, R.E., & Smith, E.V. (2007). A Rasch perspective. *Educational and Psychological Measurement, 67*(3), 394-409. https://doi.org/10.1177/0013164406294776

Schwanenflugel, P.J., Hamilton, A., Kuhn, M.R., Wisenbaker, J.M., & Stahl, S.A. (2004). Becoming a fluent reader: reading skill and prosodic features in the oral reading of young readers. *Journal of Educational Psychology, 96*(1), 119-129. https://doi.org/10.1037/0022-0663.96.1.119

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428. https://doi.org/10.1037/0033-2909.86.2.420

Sinambela, S.E. (2017). Prosody as a tool for assessing reading fluency of adult esl students. *Advances in Language and Literary Studies, 8*(6), 83-87. https://doi.org/10.7575/aiac.alls.v.8n.6p.83

Smith, G.S., & Paige, D.D. (2019). A study of reliability across multiple raters when using the NAEP and MDFS rubrics to measure oral reading fluency. *Reading Psychology, 40*(1), 34-69. https://doi.org/10.1080/02702711.2018.1555361

Spafford, C.S., Pesce, A.J.I., & Grosser, G.S. (Ed.). (1998). *The Cyclopedic Education Dictionary*. Delmar Publishers.

Stevens, D.D., & Levi, A.J. (2005). *Introduction to Rubrics: an Assessment Tool to Save Grading Time, Convey Effective Feedback, and Promote Student Learning*. Stylus.

Sudweeks, R.R., Reeve, S., & Bradshaw, W. S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*(3), 239-261. https://doi.org/10.1016/j.asw.2004.11.001

Şata, M., & Karakaya, İ. (2020). Investigation of the use of electronic portfolios in the determination of student achievement in higher education using the many-facet Rasch measurement model. *Educational Policy Analysis and Strategic Research*, *15*(7-21).

Taşkaya, S.M., & Muşta, M.C. (2008). Sınıf öğretmenlerinin Türkçe öğretim yöntemlerine ilişkin görüşleri [Teachers' opinions on Turkish teaching methods]. *Elektronik Sosyal Bilimler Dergisi, 7*(25), 240-251.

U.S. Department of Education. (2002). *National assessment of educational progress (NAEP) 2002 oral reading fluency study*. Washington, DC: Institute of Education Sciences, National Center for Education Statistics.

Ulusoy, M., Ertem, İ.S., & Dedeoğlu, H. (2011). Evaluating pre-service teachers' oral reading records prepared for the grades 1-5 considering the prosodic competences. *Gazi University Journal of Gazi Educational Faculty, 31*(3), 759-774.

Valencia, S.W., Smith, A.T., Reece, A.M., Li, M., Wixson, K.K., & Newman, H. (2010). Oral reading fluency assessment: Issues of construct, criterion, and consequential validity. *Reading Research Quarterly, 45*, 270-291. https://doi.org/10.1598/RRQ.45.3.1

VandenBos, G.R. (2015). *APA Dictionary of Psychology*. American Psychological Association.

Xu, Y., & Liu, F. (2012). Intrinsic coherence of prosodic and segmental aspects of speech. In O. Niebuhr (Ed.), *Understanding prosody: the role of context, function and communcation* (Vol. 13, pp. 1-26). De Gruyter.

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing, 31*(4), 501-527. https://doi.org/10.1177/0265532214536171

Yıldız, M., Kanık Uysal, P., Bilge, H., Wolters, A.P., Saka, Y., Yıldırım, K., & Rasinski, T. (2019). Relationship between Turkish eighth-grade students' oral reading efficacy, reading comprehension and achievement scores on a high-stakes achievement test. *Reading Psychology, 40*(4), 1-21. https://doi.org/10.1080/02702711.2018.1555363

Yıldız, M., Yıldırım, K., Ateş, S., & Çetinkaya, Ç. (2009). An evaluation of the oral reading fluency of 4th graders with respect to prosodic characteristic. *International Journal of Human Sciences, 6*(1), 353-360.

Young, C., & Rasinski, T. (2009). Implementing readers theatre as an approach to classroom fluency instruction. *The Reading Teacher, 63*(1), 4-13. https://doi.org/10.1598/RT.63.1.1

Yüzüak, A., Yüzüak, B., & Kaptan, F. (2015). Performans görevinin akran gruplar ve öğretmen yaklaşımları doğrultusunda çok-yüzeyli Rasch ölçme modeli ile analizi [A many-facet Rasch measurement approach to analyze peer and teacher assessment for authentic assessment task]. *Journal of Measurement and Evaluation in Education and Psychology, 6*(1), 1-11.

Zimmerman, B.S., Rasinski, T., Was, C.A., Rawson, K.A., Dunlosky, J., Kruse, S.D., & Nikbakht, E. (2019). Enhancing outcomes for struggling readers: Empirical analysis of the fluency development lesson. *Reading Psychology, 40*(1), 70-94. https://doi.org/10.1080/02702711.2018.1555365

Zutell, J., & Rasinski, T. (1991). Training teachers to attend to their students' oral reading fluency. *Theory Into Practice, 30*(3), 211-217. https://doi.org/10.1080/00405849109543502

APPENDIX

Bias Interaction

Observd Score	Expctd Score	Observd Count	Obs-Exp Average	Bias- Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	Sq	Nu	pua	measr-	N	Ölçütler	measr-
161	156.59	41	.11	-1.01	.60	-1.69	40	.0984	1.1	1.4	36	6	P6	-2.87	4	hız	-.32
127	113.23	41	.34	-.64	.23	-2.86	40	.0067	.6	.7	3	3	P3	-.69	1	ifade ve ses düzeyi	.14
136	124.91	41	.27	-.60	.24	-2.45	40	.0190	1.5	1.4	35	5	P5	-.77	4	hız	-.32
106	96.86	41	.22	-.38	.20	-1.85	40	.0714	1.3	1.3	17	7	P7	.33	2	anlama üniteleri ve tonlama	.13
106	96.86	41	.22	-.38	.20	-1.85	40	.0714	1.3	1.3	18	8	P8	.33	2	anlama üniteleri ve tonlama	.13
130	123.05	41	.17	-.35	.23	-1.52	40	.1363	1.1	1.0	34	4	P4	-.68	4	hız	-.32
156	153.85	41	.05	-.29	.39	-.75	40	.4572	1.0	1.3	26	6	P6	-2.87	3	pürüzsüzlik	.06
121	114.80	41	.15	-.28	.22	-1.30	40	.2003	1.1	1.1	32	2	P2	-.30	4	hız	-.32
128	123.39	41	.11	-.23	.23	-1.01	40	.3177	.7	.8	29	9	P9	-.75	3	pürüzsüzlik	.06
125	120.44	41	.11	-.22	.22	-.98	40	.3315	1.0	.9	10	10	P10	-.69	1	ifade ve ses düzeyi	.14
96	91.64	41	.11	-.18	.20	-.89	40	.3793	.6	.6	21	1	P1	.29	3	pürüzsüzlik	.06
105	100.84	41	.10	-.17	.20	-.85	40	.4027	.7	.7	31	1	P1	.29	4	hız	-.32
123	121.74	41	.03	-.06	.22	-.27	40	.7850	.6	.6	9	9	P9	-.75	1	ifade ve ses düzeyi	.14
100	98.66	41	.03	-.06	.20	-.27	40	.7863	1.3	1.3	27	7	P7	.33	3	pürüzsüzlik	.06
100	98.66	41	.03	-.06	.20	-.27	40	.7863	1.3	1.3	28	8	P8	.33	3	pürüzsüzlik	.06
116	114.77	41	.03	-.05	.21	-.26	40	.7978	.7	.7	24	4	P4	-.68	3	pürüzsüzlik	.06
123	122.12	41	.02	-.04	.22	-.19	40	.8491	.7	.7	30	10	P10	-.69	3	pürüzsüzlik	.06
97	96.74	41	.01	-.01	.20	-.05	40	.9588	1.2	1.2	7	7	P7	.33	1	ifade ve ses düzeyi	.14
97	96.74	41	.01	-.01	.20	-.05	40	.9588	1.2	1.2	8	8	P8	.33	1	ifade ve ses düzeyi	.14
153	153.23	41	-.01	.03	.34	.08	40	.9372	.9	1.1	16	6	P6	-2.87	2	anlama üniteleri ve tonlama	.13
103	104.05	41	-.03	.04	.20	.22	40	.8307	1.3	1.3	2	2	P2	-.30	1	ifade ve ses düzeyi	.14
114	115.03	41	-.03	.05	.21	.22	40	.8297	1.3	1.3	5	5	P5	-.77	1	ifade ve ses düzeyi	.14
112	113.34	41	-.03	.06	.21	.28	40	.7811	.4	.4	13	3	P3	-.69	2	anlama üniteleri ve tonlama	.13
119	120.54	41	-.04	.07	.21	.33	40	.7405	.7	.7	20	10	P10	-.69	2	anlama üniteleri ve tonlama	.13
102	104.16	41	-.05	.09	.20	.44	40	.6608	1.0	1.0	12	2	P2	-.30	2	anlama üniteleri ve tonlama	.13
111	113.08	41	-.05	.09	.21	.43	40	.6659	.8	.8	14	4	P4	-.68	2	anlama üniteleri ve tonlama	.13
129	130.87	41	-.05	.10	.23	.43	40	.6660	.9	.9	39	9	P9	-.75	4	hız	-.32
87	89.86	41	-.07	.12	.21	.58	40	.5620	.4	.4	11	1	P1	.29	2	anlama üniteleri ve tonlama	.13
103	105.94	41	-.07	.12	.20	.60	40	.5502	.9	.9	22	2	P2	-.30	3	pürüzsüzlik	.06
118	121.84	41	-.09	.18	.21	.83	40	.4091	.8	.8	19	9	P9	-.75	2	anlama üniteleri ve tonlama	.13
111	115.13	41	-.10	.18	.21	.87	40	.3895	1.1	1.1	15	5	P5	-.77	2	anlama üniteleri ve tonlama	.13
126	129.74	41	-.09	.19	.22	.86	40	.3959	.8	.8	40	10	P10	-.69	4	hız	-.32
84	89.75	41	-.14	.24	.21	1.17	40	.2471	.9	.8	1	1	P1	.29	1	ifade ve ses düzeyi	.14
118	123.28	41	-.13	.25	.21	1.16	40	.2534	.5	.5	33	3	P3	-.69	4	hız	-.32
111	116.80	41	-.14	.25	.21	1.23	40	.2266	1.3	1.3	25	5	P5	-.77	3	pürüzsüzlik	.06
107	112.98	41	-.15	.26	.21	1.25	40	.2198	1.4	1.4	4	4	P4	-.68	1	ifade ve ses düzeyi	.14
108	115.03	41	-.17	.30	.21	1.48	40	.1476	.4	.4	23	3	P3	-.69	3	pürüzsüzlik	.06
97	107.79	41	-.26	.45	.20	2.21	40	.0329	1.3	1.3	37	7	P7	.33	4	hız	-.32
97	107.79	41	-.26	.45	.20	2.21	40	.0329	1.3	1.3	38	8	P8	.33	4	hız	-.32
147	153.19	41	-.15	.60	.29	2.09	40	.0427	.8	1.0	6	6	P6	-2.87	1	ifade ve ses düzeyi	.14
115.3	115.23	41.0	.00	-.02	.23	-.01			1.0	1.0	Mean (Count: 40)						
17.7	16.92	.0	.13	.31	.07	1.21			.3	.3	S.D. (Population)						
18.0	17.13	.0	.14	.31	.07	1.22			.3	.3	S.D. (Sample)						

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 2, 492–514
https://doi.org/10.21449/ijate.877035
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Investigating the Impact of Rater Training on Rater Errors in the Process of Assessing Writing Skill

Mehmet Sata[1,*], **Ismail Karakaya**[2]

[1]Agri Ibrahim Cecen University, Faculty of Education, Department of Educational Sciences, Agri, Türkiye
[2]Gazi University, Faculty of Gazi Education, Department of Educational Sciences, Ankara, Türkiye

ARTICLE HISTORY
Received: Feb. 08, 2021
Revised: Apr. 15, 2022
Accepted: May 14, 2022

Keywords:
Rater training,
Rater errors,
Many facet Rasch model,
Validity,
Reliability.

Abstract: In the process of measuring and assessing high-level cognitive skills, interference of rater errors in measurements brings about a constant concern and low objectivity. The main purpose of this study was to investigate the impact of rater training on rater errors in the process of assessing individual performance. The study was conducted with a pretest-posttest control group quasi-experimental design. In this research, 45 raters were employed, 23 from the control group and 22 from the experimental group. As data collection tools, a writing task that was developed by IELTS and an analytical rubric that was developed to assess academic writing skills were used. As part of the experimental procedure, rater training was provided and this training was implemented by combining rater error training and frame of reference training. When the findings of the study were examined, it was found that the control and experimental groups were similar to each other before the experiment, however, after the experimental process, the study group made more valid and reliable measurements. As a result, it was investigated that the rater training given had an impact on rater errors such as rater severity, rater leniency, central tendency, and Halo effect. Based on the obtained findings, some suggestions were offered for researchers and future studies.

1. INTRODUCTION

Cognitive skills are divided into two categories: lower-order and higher-order. Lower-order cognitive skills in Bloom's Taxonomy include behaviors that belong to the remembering and understanding levels and these behaviors do not change from learner to learner, are measured by traditional tools, and are result-oriented. Higher-order cognitive skills, on the other hand, are process-oriented, are measured by complementary measurement and assessment tools (such as essay, portfolio, performance task, etc.), and their acquisition takes more time compared to lower-order cognitive skills (Kutlu et al., 2014). Kutlu et al. (2014) indicated that higher-order cognitive skills are a combination of cognitive, affective, and psychomotor characteristics of an individual when he/she displays his/her talents. Since higher-order cognitive skills are a significant indicator of the development of success, measuring them reliably and validly is of paramount importance (Haladyna, 1997).

*CONTACT: Mehmet Sata msata@agri.edu.tr Agri Ibrahim Cecen University, Faculty of Education, Department of Educational Sciences, Agri, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

It was indicated that measuring and assessing higher-order cognitive skills with traditional measurement tools is not appropriate, and complimentary measurement and assessment tools should be employed more for this purpose (Ebel, 1965; Kutlu et al., 2014). It seems more appropriate to use performance assessment to measure and assess higher-order cognitive skills consistently and accurately (Johnson et al., 2008). Performance assessment was defined as the activities that are done to determine an individual's strengths and weaknesses by observing him/her and take actions to make these better (Bennet, 1998). Performance assessment is different from traditional assessment methods due to the following characteristics: a) performance assessment is based on real-life events, b) performance assessment is process-oriented, and c) performance assessment prompts the individual to think more (Brown & Hudson, 1998; Khattri et al., 1995; Moore, 2009).

While performance assessment provides significant advantages in measuring higher-order cognitive skills, the objectivity of measurements is an important implication problem in the process of assessing an individual's performance. It is quite difficult in practice for performance assessment to be as objective as traditional assessment methods (Romagnano, 2001). When the literature is examined, it was seen that many methods were suggested and employed for the objectivity of measurements in performance assessment-based studies. These methods are automated scoring (Attali et al., 2010; Burstein et al., 1998; Landauer et al., 2003), employing more than one rater (Gronlund, 1977; Kubiszyn & Borich, 2013), using rubrics (Dunbar et al., 2006; Ebel & Frisbie, 1991; Kutlu et al., 2014; Oosterhof, 2003), and rater training (Bernardin & Buckley, 1981; Haladyna, 1997; Ilhan & Cetin, 2014; Lumley & McNamara, 1995). It was emphasized that regardless of the method used in the performance assessment process, it is often quite difficult to ensure consistency between raters and assessors (Haladyna, 1997). In other words, regardless of the method employed, there is a possibility that some external factors other than individual performance often interfere with the measurements in the process of performance assessment. These inconsistencies that occur in the performance assessment process are defined as "rater effects/behaviors / bias" (Farrokhi et al., 2011; Haladyna, 1997; Ilhan, 2015).

If one or more of the rater behaviors are involved in the performance of the individual in the performance assessment process, the amount of error in the estimations made while determining the individual's ability level will be high, so the validity of the inferences made according to these values may be low. Rater behaviors directly threaten validity because they are attributed to a variance that is unrelated to the measured structure (Abu Kassim, 2011; Brennan, Gao & Colton, 1995; Congdon & McQueen, 2000; Farrokhi et al., 2011). In this context, it is important to determine rater behaviors in the process of scoring individual performance and to bring these behaviors to a minimum or controllable level or eliminate them (Kim, 2009; Linacre, 1994).

In the performance evaluation process, one of the methods used to reduce or control rater behaviors that interfere with measurements is rater training. Rater training is widely used to reduce the variance of raters (Brijmohan, 2016). The main purpose of rater training is to explain the assessment tools to raters through sample applications and to establish a common understanding and conceptualization among raters (Fahim & Bijani, 2011). Rater training can reduce, but not eliminate, variability in rater behaviors. One of the purposes of rater training is to increase the consistency between raters and within raters by observing factors such as experience, scoring style or scoring preference, giving feedback to raters (Kim, 2009).

Many rater training patterns/models have been proposed to reduce the raters' biases, increase the accuracy of the assessment, improve observation skills, and increase behavioral accuracy and rater reliability (Woehr & Huffuct, 1994; Zedeck & Cascio, 1982). The most preferred of these patterns are; i) Self-Leadership Training (SLT), ii) Behavioral Observation Training (BOT), iii) Rater Variability Training (RVT), iv) Performance Dimension Training (PDT), v)

Rater Error Training (RET), vi) Frame-of-Reference Training (FORT). It is seen that each rater training in the literature has different approaches. Regardless of which rater training patterns are used, the main target training is expected to increase rater reliability and accuracy and decrease rater behaviors. Since this research did not examine which rater design is better, this discussion was not entered into. In this study, RET and FORT designs were combined based on the literature in order to get maximum efficiency from rater training.

Rater training methods were used in this study to determine rater behaviors and reduce them to a controllable level in the performance assessment process. The main purpose of rater training is to enable raters to develop a common-sense towards student performance and criteria of assessment preferences (Eckes, 2008; Shale, 1996). In other words, it is ensured that the assessment is done validly and reliably (Moser et al., 2016). Since the scores students get from an open-ended exam consist of both the performance of the student and the rater's interpretation of the student's performance, it creates a constant validity concern in the test results (Ellis, Johnson & Papajohn, 2002; McNamara, 1996). If the decisions that are made based on test results are vital, rater behaviors should be determined and these behaviors should be reduced to an acceptable level (Ellis et al., 2002). When the literature was examined, it was seen that many rater training designs were suggested and used (Bernardin & Buckley, 1981; Feldman et al., 2012; Haladyna, 1997; Hauenstein, & McCusker, 2017; Stamoulis & Hauenstein, 1993; Weigle, 1998; Zedeck & Cascio, 1982).

When the literature is examined, it is seen that there are many rater trainings, but the existence of such a study in the national literature has been the main motivation for conducting this study. In addition, it is thought that the relevant study is important in terms of testing the effectiveness of rater training in the evaluation of compositions. Another originality of the study is that the second language academic writing skills of Turkish students were measured for the first time with a combined rater design. In this regard, rater training was provided in this study by combining rater error training and frame of reference training designs, and its impact on rater behaviors was investigated.

For the purpose of this study, the following hypotheses were tested:

1. Before training, the raters in the experimental and control groups showed rater behaviors in the process of assessing the writing performance of students,

2. After training, the raters in the experimental group showed fewer rater behaviors than those in the control group in the process of assessing the writing performance of students.

2. METHOD

2.1. Research Design

In this study, a quasi-experimental design with control & experimental groups and pretest & posttest was employed. This pattern is a relational design because the same people are measured twice on the dependent variable. However, it is also defined as an unrelated design due to the comparison of the measurements of the experimental and control groups consisting of different participants (Howitt & Cramer, 2008). Because of these two features, pre-test post-test control group design is defined as a mixed design in the quantitative studies (Buyukozturk, 2011).

2.2. Study Group

Since there is no assumption that results obtained through Rasch models can be generalized to the universe, universe and sample were not identified in this study, instead, a study group was chosen. There were two groups involved in the study: raters and students. There were 64 raters, 12 of whom were male, and 52 of whom were female; while individuals consisted of 39 students. Both individuals/students and raters were student teachers of English at Gazi University, English Language Teaching (ELT) department. Raters were the 3rd-grade students

who took the Measurement and Evaluation course, while individuals were the 1st-grade students, who took the 'Advanced Reading and Writing' course. The average age of the raters was 21.84, and they had not participated in any rater training and thus, had no experience in scoring before. The raters in the study were randomly divided into two groups (33 for the control group, and 31 for the experimental group). All the participants took place voluntarily in the research. However, 7 raters who participated in the pre-test but did not participate in the post-test were excluded from the study. Later, pre-test scores were analyzed and misfit was detected with 12 raters. These raters were also excluded from the study because the misfit negatively affected the model-data fit of the study. As a result, the study was conducted with a total number of 45 raters, 22 in the experimental group and 23 in the control group.

2.3. Data Collection Tools

A writing task (argumentative essay), personal information form, and analytical rubric were used as data collection tools in the study.

2.3.1. *The Writing task*

An argumentative essay task, which was prepared by the International English Language Testing System (IELTS) and was published as sample, was used to measure the academic writing skills (related performance) of individuals (see Appendix 1). One of the reasons for choosing this writing task is that it is authentic and reflects a real-life situation, and this provides a more valid framework for measuring student performance. Before the participants were given this task, they were informed that the researchers would not grade this task, it would be used only for academic purposes, the participation was voluntary, and they should not write their personal information on the sheets. The participants were told that they had 40 minutes to complete the task, and they are required to write an essay of within at least 250 words. The writing task was completed by 39 participants, and they were above B1 level. Later, these essays were numbered and duplicated for the rating purpose. The essays were written in the spring semester of the 2017-2018 academic year.

2.3.2. *Personal information form*

A personal information form was prepared by the researcher to collect the interests, attitudes, anxieties, and demographic information of the raters towards academic writing. A rating scale was also included in the personal information form, in which raters would write the score they gave to the essays on each criterion.

2.3.3. *Analytical rubric for academic writing skill*

To assess the essays, the researchers and a Ph.D. student from the ELT department who is knowledgeable about academic writing developed an analytical rubric for academic writing skills. While developing the rubric, a systematic process with certain steps was followed because the validity and reliability of the measurements obtained from the measurement tools developed without following a systematic process may be negatively affected. Therefore, reliability and validity should be taken into account in the process of developing rubrics (Moskal, 2000). During the development of the analytical rubric, Goodrich (1997), Haladyna (1997), Kutlu et al. (2014) and Moskal's (2000) suggestions were taken into consideration.

First of all, as the aim is to assess student teachers' academic writing skills, the purpose of the rubric was determined accordingly. In the second stage, the criteria for assessing performance (academic writing skill) were determined and sample rubrics in studies such as Weigle (2002), Hughes (2003), Brown (2004), Brown (2007), and Brookhart (2013) were examined in detail. Upon reviewing the literature, seven main criteria and 20 sub-criteria were selected and a draft form was created. Then, the draft form of the rubric was given to 11 field expertsto assess the criteria in the draft by using a measurement tool with a triple grading as (1) sufficient, (2)

sufficient but should be corrected, and (3) insufficient. After the opinions of field experts were taken into account as academic writing competencies, they were presented as evidence for the content validity. For the content validity of each criterion, Lawshe's (1975) approach was taken into consideration. Since there are 11 field experts in this study, it was taken into consideration that the content validity rate (CVR) should be equal to or greater than a minimum 0.591 value in order for any criterion to have sufficient coverage in academic writing skills (Wilson, Pan & Schumsky, 2012). The CVR value for each criterion was calculated and six criteria that were less than 0.591 were removed from the draft form. Moreover, based on the feedback received from field experts, two criteria were divided into two sub-criteria. As a result, a measurement tool consisting of six main criteria and 16 sub-criteria was obtained as the final form. The final form the rubric was presented in Table 1.

Table 1. *Criteria included in the measurement of writing skill.*

Criteria		Scoring					
Main Criteria	Sub-criteria	0 score	1 score	2 score	3 score	4 score	Total score
Organization	Title of Essay	( )	( )	( )	( )	( )	( )
	Introduction-Body-Conclusion	( )	( )	( )	( )	( )	( )
	Thesis Statement	( )	( )	( )	( )	( )	( )
	Topic Sentence	( )	( )	( )	( )	( )	( )
	Supporting Sentence	( )	( )	( )	( )	( )	( )
	Appropriate Length	( )	( )	( )	( )	( )	( )
Content	Topic Relevance	( )	( )	( )	( )	( )	( )
	Idea Development	( )	( )	( )	( )	( )	( )
Coherence	Coherence	( )	( )	( )	( )	( )	( )
Cohesion	Linking	( )	( )	( )	( )	( )	( )
Grammar	Accuracy of Grammatical Forms	( )	( )	( )	( )	( )	( )
	Syntatic Complexity	( )	( )	( )	( )	( )	( )
Vocabulary	Word Choice	( )	( )	( )	( )	( )	( )
	Lexical Range	( )	( )	( )	( )	( )	( )
Mechanics	Spelling	( )	( )	( )	( )	( )	( )
	Punctuation	( )	( )	( )	( )	( )	( )

When Table 1 is examined, it can be seen that the measurement tool consists of six main and 16 sub-criteria with a five-point rating. Upon reviewing the student teachers' essays, it was seen that most of them did not give a title to their essays even though they were told to do it, so the sub-criterion of 'Title of Essay' was excluded from the study because it distorted the data structure. After the CVR value was calculated for each criterion, the content validity index (CVI) value was calculated for the measuring tool, and this value was found to be 0.750. As a result, since the calculated CVI value is greater than 0.591, it was accepted that the prepared rubric had a sufficient scope for measuring academic writing skills. The content validity index is the prerequisite of the construct validity process (Lawshe, 1985). It was decided that the last version of the form had a five-point rating and the use of the analytical rubric was appropriate since the performance (academic writing), which was determined, was divided into sub-dimensions (Kutlu et al., 2014).

After providing evidence for the content validity of the developed analytical rubric, the evidence for the construct validity was collected. Exploratory factor analysis (EFA) was conducted to provide evidence for construct validity. Before EFA, Kaiser-Meyer-Olkin (KMO) and Barlett sphericity tests were conducted to determine whether the relevant data set had a factorizable structure. It is stated that for a data set to be factorizable, the KMO value should be 0.70 and the Barlett sphericity test should be significant (Cokluk et al., 2012). The KMO

value for the relevant data set was found to be 0.875 and the Barlett spherical test was found to be statistically significant (χ^2 (sd) = 956.427 (105); p = 0.000).

It was determined that there were no losses and misfit in the data set and the relationships between variables were linear. Test of normality was performed for each criterion and it was investigated that except two, all criteria showed normal distribution. During the process of EFA, the average score for each criterion rated by 45 raters in the experimental and control groups for the 39 student essays was analyzed. As a result of EFA analysis, it was found that the criteria were collected under a single factor and the variance was 70.05% (the factor loadings of the criteria for the relevant data set were as follows; 0.842; 0.855; 0.936; 0.968; 0.644; 0.860; 0.960; 0.987; 0.945; 0.605; 0.911; 0.891; 0.899; 0.861 and 0.622).

After collecting the evidence for the validity of the measurements obtained from the developed analytical rubric, McDonald ω coefficient was used for proof of the reliability of the measurements (McDonald, 1999). The reason for using the McDonald ω coefficient is to obtain more consistent and unbiased estimates of reliability (Osburn, 2000) in such measurements, since the factor loads of variables are different from each other (congeneric measurements). As a result of the analysis, McDonald ω coefficient was found to be 0.971 (95% Confidence Interval: 0.956-0.980). When the evidence obtained for reliability and validity is considered, it can be said that the measurements obtained from the analytical rubric to assess academic writing skills (related performance) are reliable and the inferences made based on these measurements are valid.

2.4. Experimental Procedure

In this section, information about the experimental procedure (rater training) that was applied to the experimental group was presented. First of all, all the raters in both experimental and control groups were informed about the developed analytical rubric and the performance to be assessed (academic writing skill). Moreover, they were informed about rubrics, their types, and how they were prepared the reason for this was to ensure that the experimental and control groups would have similar characteristics and experiences. In this way, it was aimed to reduce the possibility of mixing the different variance sources (variance unrelated to the structure) to the performance (related structure) to be determined. Next, the experimental and control groups were given detailed information about the analytical rubric's criteria and rating. All of these procedures took a total of three weeks, one hour each week, before the experimental procedure. In addition, the raters were not given any information as to whether they were in the experimental or control groups. After these procedures, all raters were given a 'rater file' that contained pre-prepared and numbered student essays, analytical rubric, and personal information form (pre-test), and they were given a week to assess student essays according to the developed analytical rubric. At the end of one week, rater files were collected and the assessments were transferred to the computer environment and the data set was analyzed. As a result of the analysis, it was identified that the experimental and control groups displayed similar rater behaviors in the process of assessing students' essays. Later on, the rater training was launched. Detailed information about the rater training was presented in the next section.

2.4.1. *Rater training*

To create a common structural framework (academic writing skill) among the raters in the study, rater error training (RET) and frame of reference training (FORT) were combined and applied. These two pieces of training were combined because although the RET is useful in terms of defining rater behaviors, it is not effective on rater accuracy, and the FORT is useful and effective on rater accuracy (Murphy & Balzer, 1989; Sulsky & Day, 1992). In other words, both patterns are chosen because they are complementary to each other.

The basic assumption of the RET design is that being familiar with common rater behaviors and encouraging raters to avoid these mistakes will directly lead to a decrease in rater behavior and thus more effective performance evaluation (Woehr & Huffuct, 1994). Studies have not found any evidence that RET design has a positive effect on scoring features such as inter-rater reliability (Bernardin & Pence, 1980; Borman, 1975). Although rater behaviors such as rater strictness and generosity decreased in the RET design, it was reported that scoring accuracy also decreased (Bernardin & Pence, 1980). Many researchers citing these results stated that the RET design was an inappropriate approach.

Although there are many rater training designs, it has been stated that the most preferred method is frame-of-reference traning (FORT) (Roch et al., 2012). The main reason for this is the use of a common conceptualization of performance for raters when performance is observed and evaluated (Aguinis et al., 2009; Athey & McIntyre, 1987). One reason for the effective use of the frame of reference training is that it encompasses performance theory, which is an explanation of various performance dimensions. Performance theory explains how rater behavior matches the appropriate dimension, how the effectiveness of rater behavior is evaluated, and how different judgments combine with the scoring dimension of performance (Sulsky & Day, 1992).

A rater module has been developed by the researchers for rater training. The rater training was given to student teachers of English who have taken the measurement and evaluation course for a total of four weeks and one hour each week. The rater training was implemented based on the sequence of the application in the rater module attached.

2.5. Data Analysis

Many Facet Rasch Model (MFRM) and independent samples *t*-test were used in the analysis of the data set. There are three dimensions in the study: raters, students, and criteria, and a fully crossed pattern was used because the raters assessed all students based on all the criteria.

2.5.1. *Many facet rasch model*

In the basic Rasch model, the individual and test items or performance tasks are assessed and the skill differences of the individuals and the difficulty levels of the items are placed on an equally spaced scale. It is claimed that the obtained results are independent of the sample (Sudweeks et al., 2005). In the Many Facet Rasch Model, many variability sources (such as rater, item, task, individual, time) can be placed on a single equally spaced scale (Linacre, 1993). MFRM is also known as facet models (Eckes, 2015). Although the MFRM model takes into account all variability sources, it also focuses on the interaction of these variability sources with each other (Abu Kassim, 2007). The Many Facet Rasch Model is a linear model that calibrates all parameters and converts the observations in the ranking scale to an equidistant logit scale (Bond & Fox, 2015). Logistic transformation of sequential category probabilities (log odds) enables independent variables such as peer assessment, assessment criteria, and open-ended items to be seen as dependent variables (Esfandiari, 2015). The Many Facet Rasch Model provides researchers with information that the models based on classical test theory and generalizability theory cannot provide (Lunz et al., 1990).

In this study, because academic essays written by a group of students were assessed by a group of raters, the model of the research was defined as follows:

$$\log\left(\frac{P_{bkpx}}{P_{bkpx-1}}\right) = \theta_b - \beta_k - \alpha_p - \tau_x \quad (1)$$

Pbkpx = the probability of giving an x score to a student's certain criterion by the rater

Pbkpx−1 = the probability of giving an x-1 score to a student's certain criterion by the rater

θb = b. the student's proficiency level,

k = k. the difficulty of the criterion,

αp = p. the severity of the rater,

τx = difficulty of getting an x score instead of x-1.

Assumptions to be met for the Many Facet Rasch Model are one-dimensionality, local independence and model data fit. First of all, when the one-dimensionality assumption is examined, it can be said that the related assumption is met, since the developed analytical rubric, as shown in the data collection tools section, has one factor. After the one-dimensionality assumption was met, G^2 statistics, which was proposed by Chen and Thissen (1997) was used to test local independence assumption. According to this statistic, the standardized LD χ^2 value estimated between each variable pair is below 10 and the marginal fit χ^2 value estimated for each variable is close to zero, which indicates local independence. In this context, estimates were made according to the generalized partial credit model and it was found that the standardized LD χ^2 values ranged from -0.4 to 4.5, and the marginal fit χ^2 values were close to zero, and as a result, local independence was achieved. Finally, standardized residual values were examined for model-data fit. For the model-data fit, it has been stated that the number of standardized residual values outside the ± 2 range should not be more than 5% of the total number of observations, and the standardized residual values outside the ± 3 range should not be more than 1% of the total data number (Linacre, 2017). Since the total number of observations for the pretest application is 39x45x15 = 26.325, the number of standardized residual values outside the ± 2 range is 1.067 (4.05%) and the number of standardized residual values outside the ± 3 range is 164 (0.62%). It was observed that model-data fit was achieved for pre-test application. The total number of observations for the post-test application was 26.322 (3 missing data), while the number of standardized residual values outside the ± 2 range was 995 (3.78%) and the number of standardized residual values outside the ± 3 range was 186 (0.71%) and it was accepted that model data fit was achieved for posttest. As a result, all the assumptions were met and the process of analysis was started.

3. FINDINGS

Findings were provided under headings as two hypotheses were tested. Besides, the measurement reports used in determining the rater errors were given in the appendices.

3.1. Research Findings of Pre-Rater Training

Literature warns that many rater errors get involved in the measurements when assessing the performance of an individual (Royal & Hecker, 2016). The present study examined the most frequently occurring rater errors such as rater severity, rater leniency, central tendency, and halo effect. Before the rater training, the rater facet measurement report given in Appendix 3 was examined for the rater severity and rater leniency errors involved in the measurements in the assessment of student compositions. Measurement reports were calculated for each facet and surface interactions (common interactions) in the MFRM. These measurement reports consisted of two parts as a group and individual levels. Measurement reports were first assessed at the group level and then at the individual level (MyFord & Wolfe, 2004). The current research considered this path.

Regarding the pre-test measurement report related to the rater facet in Appendix 2, group-level statistics (separation rate, separation index, and separation index reliability values) were found high. This indicates that the raters exhibited different errors in the process of assessing individual performance. The fixed effective chi-square value was examined to identify whether the raters showed different errors in the performance determination process, and this value was

found significant ($\chi^2(44) = 2\ 835.70$; $p < .05$). After determining that different rater errors were involved in the measurements at the group level, we attempted to identify which rater or raters showed different errors in assessing the individual performance by examining the statistics at the individual level. The logit value is one of the most important statistics at the individual level. By using the logit value, the *t*-value for each rater was obtained, and this value was compared with the critical *t* value in the *t* distribution table, and the rater error was determined. As is seen in Appendix 2, the *t*-value was calculated for each rater. Since there were 22 raters in the experimental group and 23 raters in the control group, the small sample size was taken as the basis. Besides, the degree of freedom was taken as 21 and the statistical significance level was taken as $\alpha = .05$. Considering *t* distribution table, the critical *t* value is 2.831. Accordingly, if the *t*-value calculated for each rater is greater than +2.831, it is assumed that the rater exhibits the leniency behavior, if it is less than -2.831, the rater exhibits the severity behavior. Figure 1 presents the graphical representation of the rater leniency and severity in the experimental and control groups that appeared in the pre-test scoring.

Figure 1. *The t-values obtained from the pre-tests of the experimental and control groups. (each point in the figure represents a rater).*

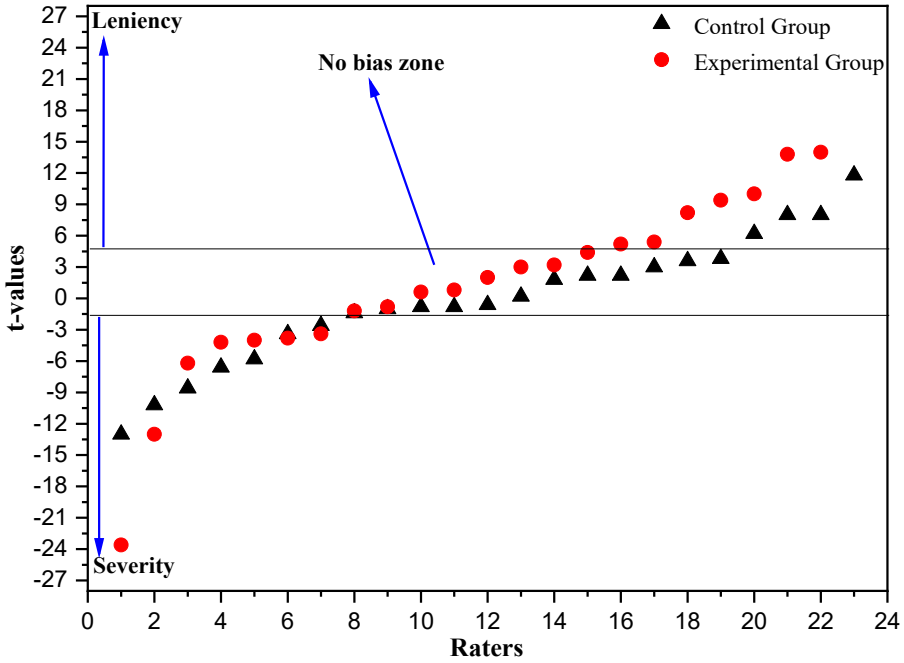


Figure 1 indicates that the raters in the control and experimental groups exhibited similar rater errors. Regarding the raters in the control group, 11 raters (47.82%) did not show leniency and severity (ideal scorer), but six raters (26.09%) showed severity, and six raters (26.09%) showed leniency. For the raters in the experimental group, five raters (22.73%) did not exhibit leniency and severity (ideal scorer), but 10 raters (45.46%) showed leniency, and seven raters (31.81%) demonstrated severity. To determine whether the logit values of the experimental and control groups differed from each other in terms of severity and leniency, independent samples t-test was conducted (see Table 2).

Table 2. *Independent samples t-test results related to the difference between pre-test scores of experimental and control groups.*

Test	Group	N	$\overline{X}$	S	*t*	*df*	*p*
Pre-test	Control	23	-0.04	0.38	0.735	43	0.466
	Experimental	22	0.05	0.43			

Note. *$p < .05$ Criteria: "Control=1"; "Experimental=2"

The logit values obtained from the pre-test of the raters in the experimental and control groups were not statistically significant ($t_{(43)}$ =0.735; $p > 0.05$). In other words, both groups showed similar rater errors and were involved in the measurements at a similar rate before the rater training.

Another rater error was central tendency. To determine the central tendency involved in the measurement of the individual performance, firstly, category statistics were calculated. Table 3 presents category statistics for pre-test results.

Table 3. *Category statistics calculated for pre-test of experimental and control groups.*

Scoring categories	Frequency	%	Cumulative %	Average logit measure	Expected logit measure	Outfit
0	595	2	2	-0.14	-0.33	1.20
1	2.194	8	11	0.11	0.11	1.00
2	6.435	24	35	0.56	0.59	1.00
3	10.132	38	74	1.10	1.11	1.00
4	6.969	26	100	1.67	1.64	1.00

As is seen in Table 3, extreme categories were preferred less, while middle categories were preferred more. In such a case, either the raters showed central tendency or the students (whose assessment preference was determined) were at the intermediate level. Therefore, referring only to category statistics at group level does not provide enough information; other statistics should also be examined. One of these statistics is the measurement report calculated for the individual / student facet. The measurement report emphasized that separation rate, separation index and separation index reliability were high. In other words, students were successfully distinguished according to their performance levels. Besides, the significant chi-square value was interpreted as statistical evidence that students were significantly differentiated according to their performance level (χ^2 (38) = 7 695.00; p = <.05). Based on these findings, it can be said that there was no central tendency at the group level, and the current situation in category statistics was due to the performance level of the students. After determining that central tendency did not interfere with the measurements at the group level, statistics at individual level were analyzed. One of these statistics is the in-compliance and out-of-compliance values estimated for each rater. The in-compliance and out-of-compliance values given in Appendix 3 were between acceptable ranges (0.50 to 1.50). The category statistics for each rater should be examined for the final decision whether central tendency inferred with the measurements at the individual level. The category statistics were calculated for each rater, and rater 11 and 23 from the control group and rater 2, 4, 6, 9, and 14 from the experimental group were found to show central tendency during the process of determining their assessment preference at the group level.

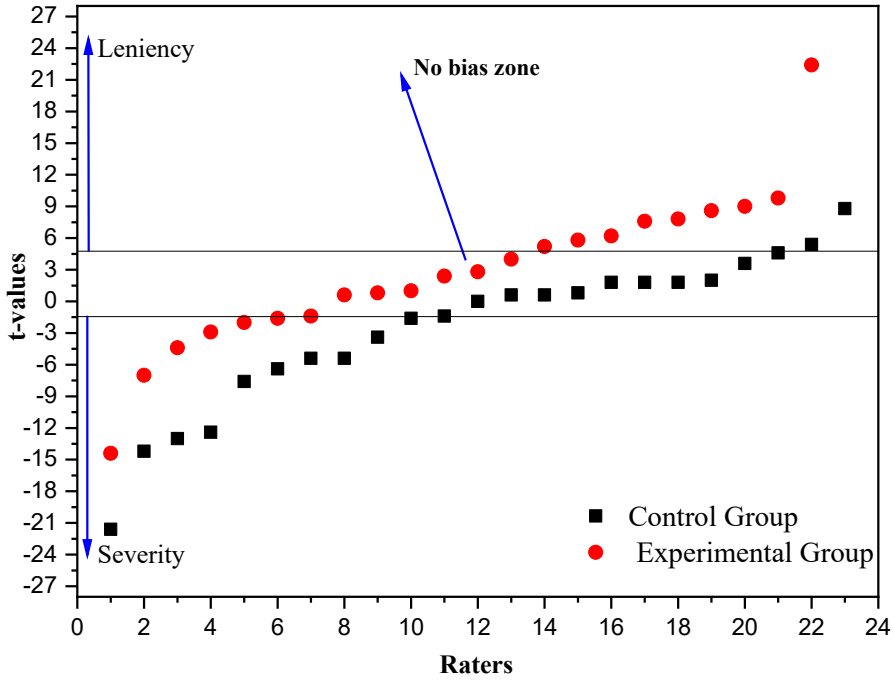
Finally, halo effect was investigated. First, group level statistics were examined. Thus, the criterion facet measurement report was studied. The separation rate, separation index and separation index reliability were found high. This shows that the difficulty levels of the criteria were different from each other and that halo behavior did not interfere with the measurements at the group level. Accordingly, halo effect was not involved in group-level measurements. It is recommended to examine the suitability values of the raters by equalizing the criteria difficulties to determine whether halo behavior is interfered with the measurements at the individual level when assessing individual performance (Linacre, 2017). If there is a rater that fits perfectly with one or both of the fit values, it is considered to show halo behavior (İlhan, 2015; Linacre, 2017). In this context, the criterion difficulties were equalized, the analysis was repeated, and the fit values of the raters were examined. According to results, rater 2 from the control group as well as rater 17 and 22 from the experimental group showed halo effect. Linacre (2017) suggests re-examining the suitability statistics of raters by equating criterion

difficulties to determine whether there is a halo effect in assessing individual performance. Therefore, MFRM analysis was repeated by setting the criterion difficulties equal to zero and the suitability statistics of the raters were examined. After the criterion difficulties were equalized, raters with a fit statistics (infit and outfit) perfectly fit with the data (with UI = 1.00 or UD = 1.00) were considered to show halo effect (Ilhan, 2015; Linacre, 2017). After the criterion difficulties were equalized, the measurement report regarding the rater aspect of the control group was obtained. Besides, in the analyses performed without equalizing the criterion difficulty, when the difference between the difficulty levels of the criteria is large, the suitability statistics are significantly greater than 1 and when the difference between the criterion difficulties is small, the relevant rater is considered to show halo behavior in performance assessment (MyFord & Wolfe, 2004). Therefore, considering the analyses performed without equalizing the criterion difficulty, it was found that three raters (1, 17 and 23) from the control group and one rater (15) from the experimental group had halo effect. As a result, it was determined that 4 raters from the control group and 3 raters from the experimental group displayed halo effect.

3.2. Research Findings of Post-Rater Training

Before rater training, upon examining rater behaviors that were involved in measurements in the process of assessing individual performance, the rater behaviors were re-examined after the rater training given. First, the effect of rater training on rater severity and rater leniency was examined. After the rater training, MFRM analysis was made for the final test, and the measurement report regarding the rater facet was presented in Appendix 3. This analysis underlined that the separation rate, separation index and separation index reliability values were high. The high values indicated that rater errors interfered with the measurements. The significance of the statistical significance test also supported this result ($\chi^2(44) = 2\ 334.60$; $p < 0.05$). After examining the statistics at the group level, the statistics at the individual level were studied. In this context, the *t*-values obtained by using the logit value were used and the *t*-value for each rater was calculated as shown in Appendix 3. Figure 2 presents the *t*-values calculated for the post-test of the experimental and control groups.

Figure 2. *t*-values obtained from the post-tests of the experimental and control groups (each point in the figure represents a rater).



As is seen in Figure 2, in the control group, 11 raters (47.83%) did not show severity or leniency, eight raters (34.78%) showed severity, and four raters (17.39%) showed leniency. In the experimental group, nine raters (40.91%) did not show severity or leniency, but 10 raters (45.46%) demonstrated leniency, and three raters (13.63%) showed severity. Besides, the raters in the experimental group got closer to the point where there was no severity and leniency. Independent samples *t*-test was conducted to determine whether the logit values obtained from the post-test of the experimental and control groups differed from each other in terms of severity and leniency. Thereby, for the effectiveness of the experimental process, the difference between raters' logit measures in the post-test and logit measures in the pre-test was taken. The differences in logit measures between post-test scores of experimental and control group was presented in Table 4.

Table 4. *Independent samples t-test result regarding the differences in logit measures between post-test scores of experimental and control groups.*

	Group	N	$\overline{X}$	*S*	*t*	*df*	*p*
Logit difference measures	Control	23	-0.09	0.16	2.708	43	0.010*
	Experimental	22	0.09	0.28			

Note. *$p < .05$ Criteria: "Control=1"; "Experimental=2"

The *t* value indicated that it was statistically significant (t_{43} =2.708; $p < 0.05$; $\eta^2 = 0.15$). Based on this finding, the rater training was effective, and this effect was great. Although rater training increases the harmony between raters, it can reveal severity and leniency due to rater drift (Moore, 2009). According to the findings of the present study, after the rater training, there were drifts in the scoring of some raters; therefore, severity and leniency emerged.

Regarding the effect of rater training on central tendency, only the statistics at the individual level were examined because it was not significant at the group level before rater training. Therefore, category statistics for each rater were examined and three raters (2, 5 and 16) from the control group and one rater (number 2) from the experimental group were observed to display central tendency during the process of determining individual performance.

The effect of rater training on halo effect was examined with the statistics at the individual level. First, after the criteria difficulties were equalized, the fit statistics of each rater were examined, and one rater (17) from the control group was found to demonstrate halo effect. In the analyzes performed without equalizing the criterion difficulties, it was found that five raters (2, 5, 11, 12 and 16) from the control group and two raters (2 and 18) from the experimental group exhibited halo effect. As a result, it was found that two raters from six experimental groups from the control group displayed halo behavior in the process of assessing individual performance. In other words, six raters from the control group and two raters from the experimental group showed halo effect in the process of assessing individual performance.

4. DISCUSSION, CONCLUSION and SUGGESTIONS

Rater training was used to determine the rater errors involved in the measurements in the process of assessing individual performance and to reduce these behaviors or bring them to a controllable level. The findings were discussed under two headings in terms of before and after the experimental procedure.

4.1. Conclusions of Pre-Rater Training and Discussion

Before the rater training, it was found that raters in both the experimental and control groups displayed similar behaviors in the process of assessing individual performance. The literature emphasizes that the severity and leniency of individual performance always interfere with the measured structure during the performance assessment process (Abu Kassım, 2007; Knoch et

al., 2018; Saritas-Akyol & Karakaya, 2021). Accordingly, rater's severity and leniency are important in intra-rater and inter-rater mismatches (Kane et al., 1995).

Before rater training, the raters in both groups were observed to have central tendency at the individual level (only some raters, not the whole group) while assessing individual performance. Esfandiari (2015) found that some raters showed central tendency when assessing academic writing skills, but they did not demonstrate it at the group level. A similar study was conducted by Engelhard (1994) who found that the scores of the students involved 80% of central tendency while assessing the academic writing skills. In another study, raters who did not have previous scoring experience displayed more central tendency than experienced raters (Leckie & Baird, 2011). Accordingly, the fact that the raters in both groups did not have previous scoring experience can be considered as one of the reasons for the presence of central tendency in the process of assessing the individual performance. Besides, the central tendency appeared less in performance assessment compared to severity and leniency. This indicates that the most common errors in performance assessment are severity and leniency (Cronbach, 1990).

Considering halo effect, it did not interfere with group-level measurements, but it did at the individual level. Literature advocates that halo effect is often involved in measurements and is the most studied error (Esfandiari, 2015). Engelhard (1994) also found the presence of halo effect in performance assessment. Similarly, Farrokhi and Esfandiari (2011) examined the interference of halo behavior with performance in the peer assessment, self-assessment and teacher assessment process. They observed that halo effect appeared in all three assessment types. In their study, Wu and Tan (2016) informed that some of the raters showed halo effect. In the present study, in order to prevent halo effect, the students' identity and socio-demographic information were not shared with the raters, however, halo effect was found to interfere with the measurements. This result is also supported by literature.

4.2. Conclusions of Post-Rater Training and Discussion

Before rater training, severity, leniency, central tendency and halo effect of the raters in both experimental and control groups were determined. Then, the experimental group went thorugh rater training on the aforementioned rater errors. The findings were reported based on the literature.

Considering the effect of rater training on rater severity and leniency, despite the rater training, it was found that severity and leniency were involved in the measurements during the process of assessing individual performance in both experimental and control groups. One of the reasons for this situation is thought to be the occurrence of rater drift when performance assessment spreads over time (Harik et al., 2009; Moore, 2009). The literature emphasizes that rater errors can change over time (Myford & Wolfe, 2009). When the amount of severity and leniency was examined after the rater training, it was found that the level of severity and leniency in the experimental group decreased from 77% to 59%, while severity and leniency in the control group was 52% both in the pre- and post-tests. No statistical difference was observed between the logit values showing the severity and leniency levels of the raters in the experimental and control groups before the rater training, but there was a statistical difference between the experimental and control groups after the rater training. For the practical significance of this difference, the effect size was calculated (Pallant, 2007). According to the calculated effect size value, rater training had a great effect on the rater severity and leniency, and 15% of the variability in rater severity and leniency could be explained by rater training. The literature supports this finding. Bijani (2018) found that rater training decreased the level of rater severity. Fahim and Bijani (2011) observed that rater severity and leniency involved in scoring during the assessment of students' second language writing skills decreased when rater training was given. Another study displayed that rater training had little effect on rater severity

and leniency, but it had a significant effect on rater consistency (Davis, 2016). On the other hand, Weitz et al. (2014) advocated that raters scored stricter after rater training. In the study conducted by Kondo (2010), rater severity and leniency were found to be similar before and after rater training. It seems normal to have different results considering the different designs and combinations of rater training given in the literature.

When the relationship between rater training and central tendency was examined, central tendency fell from 23% to 5% in the experimental group. Therefore, it can be argued that rater training had an effect on central tendency, which is often involved in measurements when assessing individual performance. Baird et al. (2013) argued that central tendency generally occurred because of inexperienced raters who used measurement tools with multiple ratings. According to Feldman et al. (2012), if central tendency interfered with the measurements in the performance assessment process, it could jeopardize the validity of the measurements by reducing the discrimination of the individual's performance level. Accordingly, it can be interpreted that the rater training provided contributes to the validity of the measurements. May (2008) stated that rater error training design was effective in reducing central tendency. Considering the combination of rater error training and frame of reference training in the present study, the findings confirmed literature. However, Bernardin (1978) and Knoch et al. (2007) found that central tendency increased after rater training contrary to expectations.

Finally, the effect of rater training on halo behavior was examined. In the control group, while there were three raters (13.04%) with halo effect in the pre-test, this number increased to six (26.09%) in the post-test. In the experimental group, four raters (18.18%) demonstrated halo effect in the pre-test results, but it decreased to two (9.09%) in the post-test. Thus, it can be argued that rater training was effective in reducing halo effect. Feldman et al. (2012) stated that halo effect increased systematic error in performance assessment, but decreased rater accuracy, and therefore, had a significant effect on the validity of the measurements. In this context, it contributed to the validity of the measurements obtained after the rater training. Bijani (2018) found that rater training reduced halo effect. Weitz et al. (2014) stated that rater training increased raters' awareness of halo effect. In the study conducted by Pulakos (1984), rater error training design was found to be effective in reducing halo effect. Similarly, Borman (1975) concluded that rater training reduced halo effect. Accordingly, the literature supports the findings of the present study.

Based on the findings, some suggestions are as follows:

- Findings showed that one or more rater behaviors were involved in the measurements during the performance assessment processes. In this context, it is expected that the analysis and determination of rater errors in the performance assessment process will contribute to the reliability of the measurements and the validity of the inferences made from the measurements.
- Rater training was found to reduce rater errors. Accordingly, it will be beneficial to provide rater training to raters or assessors for more fair and valid measurements in the performance assessment process.
- In the present study, rater error training and frame of reference training were combined and applied. Considering that there are many rater training designs and combinations, studies can be conducted to determine more effective rater designs.
- The present study was conducted with a large group (n = 22). Literature underlines the effectiveness of smaller groups (n =5 or 6). Future studies can apply the same design in smaller groups and examine its effectiveness.

Acknowledgments

This paper was produced from the first author’s doctoral dissertation.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Gazi University, 03/04/2018, 80287700-302.08.01-54466.

Authorship Contribution Statement

Mehmet Sata: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing, original draft. **Ismail Karakaya**: Methodology, Supervision, and Validation. Authors may edit this part based on their case.

Orcid

Mehmet Sata https://orcid.org/0000-0003-2683-4997
Ismail Karakaya https://orcid.org/0000-0003-4308-6919

REFERENCES

Abu Kassim, N.L. (2011). Judging behaviour and rater errors: an application of the many-facet Rasch model. *GEMA Online Journal of Language Studies*, *11*(3), 179-197.

Abu Kassim, N.L. (2007). Exploring rater judging behaviour using the many-facet Rasch model. *Paper Presented in the Second Biennial International Conference on Teaching and Learning of English in Asia: Exploring New Frontiers (TELiA2),* Universiti Utara, Malaysia.

Aguinis, H., Mazurkiewicz, M.D., & Heggestad, E.D. (2009). Using web-based frame-of-reference training to decrease biases in personality-based job analysis: An experimental field study. *Personnel Psychology*, *62*(2), 405-438. https://doi.org/10.1111/j.1744-6570.2009.01144.x

Athey, T.R., & McIntyre, R.M. (1987). Effect of rater training on rater accuracy: Levels–of–processing theory and social facilitation theory perspectives. *Journal of Applied Psychology*, *72*, 567–572. https://doi.org/10.1037/0021-9010.72.4.567

Attali, Y., Bridgeman, B., & Trapani, C. (2010). Performance of a generic approach in automated essay scoring. *Journal of Technology, Learning, and Assessment*, *10*(3), 1-16.

Baird, J.A., Hayes, M., Johnson, R., Johnson, S., & Lamprianou, I. (2013). *Marker effects and examination reliability. A Comparative exploration from the perspectives of generalisability theory, Rash model and multilevel modelling*. Oxford: University of Oxford for Educational Assessment.

Bennet, J. (1998). *Human resources management*. Singapore: Prentice Hall.

Bernardin, H.J. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. *Journal of Applied Psychology*, *63*(3), 301-308. http://dx.doi.org/10.1037/0021-9010.63.3.301

Bernardin, H.J., & Buckley, M.R. (1981). Strategies in rater training. *Academy of Management Review,* 6(2), 205-212.

Bernardin, H.J. & Pence, E.C. (1980). Effects of rater training: New response sets and decreasing accuracy. *Journal of Applied Psychology*, *65*, 60-66. https://doi.org/10.1037/0021-9010.65.1.60

Bijani, H. (2018). Investigating the validity of oral assessment rater training program: A mixed-methods study of raters' perceptions and attitudes before and after training. *Cogent Education*, *5*(1), 1-20. https://doi.org/10.1080/2331186X.2018.1460901

Bond, T., & Fox, C.M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences.* Routledge. https://doi.org/10.4324/9781315814698

Borman, W.C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. *Journal of Applied Psychology*, *60*(5), 556-560. https://doi.org/10.1037/0021-9010.60.5.556

Brennan, R.L., Gao, X., & Colton, D.A. (1995). Generalizability analyses of work key listening and writing tests. *Educational and Psychological Measurement*, *55*(2), 157-176. https://doi.org/10.1177/0013164495055002001

Brijmohan, A. (2016). *A many-facet RASCH measurement analysis to explore rater effects and rater training in medical school admissions* [Doctoral dissertation]. https://hdl.handle.net/1807/74534

Brookhart, S.M. (2013). *How to create and use rubrics for formative assessment and grading*. ASCD.

Brown, H.D. (2004). *Language assessment: Principles and classroom practices*. Pearson Education.

Brown, H.D. (2007). *Teaching by principles: An interactive approach to language pedagogy*. Pearson Education.

Brown, J.D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL quarterly*, *32*(4), 653-675. https://doi.org/10.2307/3587999

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M.D. (1998). Automated scoring using a hybrid feature identification technique. *In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, Montreal, Quebec, Canada. https://doi.org/10.3115/980845.980879

Büyüköztürk, Ş. (2011). *Deneysel desenler- öntest-sontest kontrol grubu desen ve veri analizi [Experimental designs-pretest-posttest control group design and data analysis]*. Pegem Akademi.

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

Congdon, P., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, *37*(2), 163-178. https://doi.org/10.1111/j.1745-3984.2000.tb01081.x

Cronbach, L.I. (1990). *Essentials of psychological testing*. Harper and Row.

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

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117-135. https://doi.org/10.1177/0265532215582282

Dunbar, N.E., Brooks, C.F., & Miller, T.K. (2006). Oral communication skills in higher education: Using a performance-based evaluation rubric to assess communication skills. *Innovative Higher Education*, *31*(2), 115-128. https://doi.org/10.1007/s10755-006-9012-x

Ebel, R.L. (1965). *Measuring educational achievement*. Prentice- Hall Press.

Ebel, R.L., & Frisbie, D.A. (1991). *Essentials of educational measurement*. Prentice Hall Press.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155-185. https://doi.org/10.1177/0265532207086780

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments.* Peter Lang.

Ellis, R.O.D., Johnson, K.E., & Papajohn, D. (2002). Concept mapping for rater training. *TESOL Quarterly*, *36*(2), 219-233. https://doi.org/10.2307/3588333

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93-112. https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

Esfandiari, R. (2015). Rater errors among peer-assessors: applying the many-facet Rasch measurement model. *Iranian Journal of Applied Linguistics*, *18*(2), 77-107. https://doi.org/10.18869/acadpub.ijal.18.2.77

Fahim, M., & Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. *Iranian Journal of Language Testing*, *1*(1), 1-16.

Farrokhi, F., & Esfandiari, R. (2011). A many-facet Rasch model to detect halo effect in three types of raters. *Theory & Practice in Language Studies*, *1*(11), 1531-1540. https://doi.org/10.4304/tpls.1.11.1531-1540

Farrokhi, F., Esfandiari, R., & Schaefer, E. (2012). A many-facet Rasch measurement of differential rater severity/leniency in three types of assessment. *JALT Journal*, *34*(1), 79-101.

Farrokhi, F., Esfandiari, R., & Vaez Dalili, M. (2011). Applying the many-facet Rasch model to detect centrality in self-assessment, peer-assessment and teacher assessment. *World Applied Sciences Journal*, *15*(11), 76-83.

Feldman, M., Lazzara, E.H., Vanderbilt, A.A., & DiazGranados, D. (2012). Rater training to support high-stakes simulation-based assessments. *Journal of Continuing Education in the Health Professions*, *32*(4), 279-286. https://doi.org/10.1002/chp.21156

Goodrich, H. (1997). Understanding Rubrics: The dictionary may define" rubric," but these models provide more clarity. *Educational Leadership*, *54*(4), 14-17.

Gronlund, N.E. (1977). *Constructing achievement test*. Prentice-Hall Press.

Haladyna, T.M. (1997). *Writing test items in order to evaluate higher order thinking*. Allyn & Bacon.

Harik, P., Clauser, B.E., Grabovsky, I., Nungester, R.J., Swanson, D., & Nandakumar, R. (2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement*, *46*(1), 43-58. https://doi.org/10.1111/j.1745-3984.2009.01068.x

Hauenstein, N.M., & McCusker, M.E. (2017). Rater training: Understanding effects of training content, practice ratings, and feedback. *International Journal of Selection and Assessment*, *25*(3), 253-266. https://doi.org/10.1111/ijsa.12177

Howitt, D., & Cramer, D. (2008). *Introduction to statistics in psychology*. Pearson Education.

Hughes, A. (2003). *Testing for language teachers.* Cambridge University Press.

İlhan, M. (2015). *Standart ve SOLO taksonomisine dayalı rubrikler ile puanlanan açık uçlu matematik sorularında puanlayıcı etkilerinin çok yüzeyli Rasch modeli ile incelenmesi [The identification of rater effects on open-ended math questions rated through standard rubrics and rubrics based on the SOLO taxonomy in reference to the many facet rasch model]* [Doctoral dissertation, Gaziantep University]. https://tez.yok.gov.tr/UlusalTezMerkezi/

İlhan, M., & Çetin, B. (2014). Rater training as a means of decreasing interfering rater effects related to performance assessment. *Journal of European Education*, *4*(2), 29-38. https://doi.org/10.18656/jee.77087

Johnson, R.L., Penny, J.A., & Gordon, B. (2008). *Assessing performance: Designing, scoring, and validating performance tasks*. Guilford Press.

Kane, J., Bernardin, H., Villanueva, J., & Peyrefitte, J. (1995). Stability of rater leniency: Three studies. *Academy of Management Journal*, *38*, 1036-1051.

Khaatri, N., Kane, M.B., & Reeve, A.L. (1995). How performance assessments affect teaching and learning. *Educational Leadership*, *53*(3), 80-83.

Kim, Y.K. (2009). *Combining constructed response items and multiple choice items using a hierarchical rater model* [Doctoral dissertation, Columbia University]. https://www.proquest.com/

Knoch, U., Fairbairn, J., Myford, C., & Huisman, A. (2018). Evaluating the relative effectiveness of online and face-to-face training for new writing raters. *Papers in Language Testing and Assessment*, *7*(1), 61-86.

Knoch, U., Read, J., & von Randow, T. (2007). Re-training writing raters online: How does compare with face-to-face training?, *Assessing Writing*, *12*(2), 26-43. https://doi.org/10.1016/j.asw.2007.04.001

Kondo, Y. (2010). Examination of rater training effect and rater eligibility in L2 performance assessment. *Journal of Pan-Pacific Association of Applied Linguistics*, *14*(2), 1-23.

Kubiszyn, T., & Borich, G. (2013). *Educational testing and measurement*. John Wiley & Sons Incorporated.

Kutlu, Ö., Doğan, C.D., & Karaya, İ. (2014). *Öğrenci başarısının belirlenmesi: Performansa ve portfolyoya dayalı durum belirleme [Determining student success: Determining the situation based on performance and portfolio]*. Pegem Akademi

Landauer, T.K., Laham, D., & Foltz, P.W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Lawrence Erlbaum Associates, Inc.

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

Lawshe, C.H. (1985). Inferences from personnel tests and their validity. *Journal of Applied Psychology*, *70*(1), 237-238. https://doi.org/10.1037/0021-9010.70.1.237

Leckie, G., & Baird, J.A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement, 48*(4), 399-418. https://doi.org/10.1111/j.1745-3984.2011.00152.x

Linacre, J.M. (1993). Rasch-based generalizability theory. *Rasch Measurement Transaction*, *7*(1), 283-284.

Linacre, J.M. (1994). *Many-facet Rasch measurement*. Mesa Press.

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

Lumley, T., & McNamara, T.F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54-71. https://doi.org/10.1177/026553229501200104

Lunz, M.E., Wright, B.D. & Linacre, J.M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3*(4), 331-345. https://doi.org/10.1207/s15324818ame0304_3

May, G.L. (2008). The effect of rater training on reducing social style bias in peer evaluation. *Business Communication Quarterly*, *71*(3), 297-313. https://doi.org/10.1177/1080569908321431

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

McNamara, T.F. (1996). *Measuring second language performance*. Longman.

Moore, B.B. (2009). *Consideration of rater effects and rater design via signal detection theory* [Doctoral dissertation, Columbia University]. https://www.proquest.com/

Moser, K., Kemter, V., Wachsmann, K., Köver, N.Z., & Soucek, R. (2016). Evaluating rater training with double-pretest one-posttest designs: an analysis of testing effects and the moderating role of rater self-efficacy. *The International Journal of Human Resource Management*, 1-23. https://doi.org/10.1080/09585192.2016.1254102

Moskal, B.M. (2000). *Scoring rubrics: What, when and how?*.

Murphy, K.R. & Balzer, W.K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology*, *74*, 619-624. https://doi.org/10.1037/0021-9010.74.4.619

Myford, C.M., & Wolfe, E.M. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale use. *Journal of Educational Measurement*, *46*(4), 371-389. https://doi.org/10.1111/j.1745-3984.2009.00088.x

Myford, C.M., & Wolfe, E.W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189-227.

Oosterhof, A. (2003). *Developing and using classroom assessments*. Merrill-Prentice Hall Press.

Osburn, H.G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological methods*, *5*(3), 343. http://dx.doi.org/10.1037/1082-989X.5.3.343

Pallant, J. (2007). *SPSS survival manual, a step by step guide to data analysis using spss for windows*. McGraw-Hill.

Pulakos, E.D. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology*, *69*(4), 581-588. http://psycnet.apa.org/doi/10.1037/0021-9010.69.4.581

Roch, S.G., Woehr, D.J., Mishra, V., & Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology*, *85*(2), 370-395. https://doi.org/10.1111/j.2044-8325.2011.02045.x

Romagnano, L. (2001). The myth of objectivity in mathematics assessment. *Mathematics Teacher*, *94*(1), 31-37.

Royal, K. D., & Hecker, K. G. (2016). Rater errors in clinical performance assessments. *Journal of veterinary medical education*, *43*(1), 5-8. https://doi.org/10.3138/jvme.0715-112R

Sarıtaş-Akyol, S., & Karakaya, İ. (2021). Investigating the consistency between students' and teachers' ratings for the assessment of problem-solving skills with many-facet Rasch measurement model. *Eurasian Journal of Educational Research*, *91*, 281-300. https://doi.org/10.14689/ejer.2021.91.13

Shale, D. (1996). Essay reliability: Form and meaning. In: White, E. Lutz, W. & Kamusikiri S. (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 76–96). MLAA.

Stamoulis, D.T. & Hauenstein, N.M.A. (1993). Rater training and rating accuracy: Training for dimensional accuracy versus training for ratee differentiation. *Journal of Applied Psychology*, *78*(6), 994-1003. https://doi.org/10.1037/0021-9010.78.6.994

Sudweeks, R.R., Reeve, S. & Bradshaw, W.S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, *9*, 239-261. https://doi.org/10.1016/j.asw.2004.11.001

Sulsky, L.M., & Day, D.V. (1992). Frame-of-reference training and cognitive categorization: An empirical investigation of rater memory issues. *Journal of Applied Psychology*, *77*(4), 501-510. https://doi.org/10.1037/0021-9010.77.4.501

Weigle, S.C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263-287. https://doi.org/10.1177/026553229801500205

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

Weitz, G., Vinzentius, C., Twesten, C., Lehnert, H., Bonnemeier, H., & König, I.R. (2014). Effects of a rater training on rating accuracy in a physical examination skills assessment. *GMS Zeitschrift für Medizinische Ausbildung*, *31*(4), 1-17.

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

Woehr, D.J., & Huffuct, A.I. (1994). Rater training for performance appraisal. A qantitative review. *Journal of Occupational and Organizational Psychology*, *67*(3), 189-205. https://doi.org/10.1111/j.2044-8325.1994.tb00562.x

Wu, S.M., & Tan, S. (2016). Managing rater effects through the use of FACETS analysis: the case of a university placement test. *Higher Education Research & Development*, *35*(2), 380-394. https://doi.org/10.1080/07294360.2015.1087381

Zedeck, S., & Cascio, W.F. (1982). Performance appraisal decisions as a function of rater training and purpose of the appraisal. *Journal of Applied Psychology*, *67*(6), 752-758. https://doi.org/10.1037/0021-9010.67.6.752

APPENDIX

Appendix 1. *Academic writing sample task.*

ACADEMIC WRITING SAMPLE TASK 2A

You should spend about 40 minutes on this task.

Write about the following topic:

The first car appeared on British roads in 1888. By the year 2000 there may be as many as 29 million vehicles on British roads.

Alternative forms of transport should be encouraged and international laws introduced to control car ownership and use.

To what extent do you agree or disagree?

Give reasons for your answer and include any relevant examples from your knowledge or experience.

Write at least 250 words.

Appendix 2. *Measurement report of the rater surface of the pre-test measurements of the experimental and control groups.*

Rater Code	Logit	Standart Error	Infit	Outfit	Obs %	Exp %	Rasch Kappa	t-values
OKP01	0.09	0.05	1.26	1.26	41.60	27.60	0.193	1.80
OKP02	-0.51	0.05	1.07	1.09	40.70	27.60	0.181	-10.20
OKP03	-0.04	0.05	0.90	0.93	44.10	27.40	0.230	-0.80
OKP04	0.19	0.05	0.74	0.77	38.40	27.30	0.153	3.80
OKP05	-0.13	0.05	0.74	0.76	42.90	27.80	0.209	-2.60
OKP06	0.15	0.05	0.86	0.87	37.30	26.70	0.145	3.00
OKP07	-0.04	0.05	1.00	1.01	41.40	30.10	0.162	-0.80
OKP08	-0.03	0.05	0.79	0.82	40.80	27.80	0.180	-0.60
OKP09	-0.07	0.05	0.76	0.80	42.00	28.30	0.191	-1.40
OKP10	-0.65	0.05	1.07	1.06	33.20	28.20	0.070	-13.00
OKP11	-0.17	0.05	1.29	1.22	42.00	28.20	0.192	-3.40
OKP12	-0.05	0.05	1.29	1.29	40.50	27.70	0.177	-1.00
OKP13	-0.29	0.05	1.08	1.07	24.50	24.30	0.003	-5.80
OKP14	-0.43	0.05	0.79	0.79	42.00	28.10	0.193	-8.60
OKP15	0.11	0.05	0.84	0.85	41.40	27.80	0.188	2.20
OKP16	0.59	0.06	0.84	0.84	40.10	27.90	0.169	11.80
OKP17	0.18	0.05	1.23	1.20	44.70	28.50	0.227	3.60
OKP18	0.40	0.05	1.18	1.10	42.80	27.90	0.207	8.00
OKP19	0.01	0.05	0.76	0.86	41.50	28.20	0.185	0.20
OKP20	0.40	0.05	1.14	1.11	43.00	28.30	0.205	8.00
OKP21	0.31	0.05	0.78	0.81	43.20	28.70	0.203	6.20
OKP22	0.11	0.05	0.89	0.92	42.50	28.50	0.196	2.20
OKP23	-1.13	0.05	0.93	0.92	42.10	28.00	0.196	-6.60
ODP01	0.26	0.05	0.72	0.75	40.10	31.10	0.131	5.20
ODP02	0.41	0.05	1.36	1.43	39.10	29.30	0.139	8.20
ODP03	0.16	0.05	0.68	0.69	41.20	30.70	0.152	3.20
ODP04	0.69	0.06	1.44	1.44	44.80	30.40	0.207	13.80
ODP05	0.22	0.05	0.74	0.78	41.80	30.00	0.169	4.40
ODP06	-0.65	0.05	1.45	1.49	44.00	30.30	0.197	-13.00
ODP07	0.27	0.05	1.09	1.05	40.10	29.90	0.146	5.40
ODP08	-0.06	0.05	0.60	0.60	43.50	29.90	0.194	-1.20
ODP09	0.03	0.05	1.26	1.31	37.60	28.20	0.131	0.60
ODP10	-0.21	0.05	1.10	1.08	39.70	29.50	0.145	-4.20
ODP11	-0.31	0.05	0.73	0.87	41.30	29.30	0.170	-6.20
ODP12	-1.18	0.05	0.69	0.69	35.80	28.30	0.105	-23.60
ODP13	-0.04	0.05	0.93	0.99	40.10	27.80	0.170	-0.80
ODP14	-0.17	0.05	1.07	1.15	41.50	28.20	0.185	-3.40
ODP15	-0.20	0.05	1.49	1.47	39.70	27.50	0.168	-4.00
ODP16	0.15	0.05	0.72	0.74	41.70	27.60	0.195	3.00
ODP17	-0.19	0.05	0.91	0.92	42.80	26.70	0.220	-3.80
ODP18	0.10	0.05	1.19	1.16	39.60	27.10	0.171	2.00
ODP19	0.50	0.06	1.04	1.08	38.80	27.30	0.158	10.00
ODP20	0.04	0.05	0.86	0.93	44.50	27.60	0.233	0.80
ODP21	0.47	0.06	0.83	0.89	42.30	27.80	0.201	9.40
ODP22	0.70	0.06	1.11	1.10	23.90	24.00	-0.001	14.00
Mean	0.00	0.05	1.00	1.02				
S.D.(Population)	0.40	0.00	0.25	0.25				
S.D. (Sample)	0.40	0.00	0.25	0.25				
Model. Population : RMSE = 0.05 Adj. (True) S.D. = 0.39 Separation = 7.63 Strata = 10.51 Reliability (not inter-rater) = 0.98								
Model. Sample: RMSE = 0.05 Adj. (True) S.D. = 0.40 Separation = 7.72 Strata = 10.63 Reliability (not inter-rater) = 0.98								
Model. Chi-square (Fixed) : 2.835.70 d.f. = 44 *significance (probability)* = .00								
Model. Chi-square (Normal) : 43.30 d.f. = 43 *significance (probability)* = .46								

Note. OKP: rater who took the pre-test from the control group ODP: rater who took the pre-test from the experimental group

Appendix 3. *Measurement report of the rater surface of the pre-test measurements of the experimental and control groups.*

Rater Code	Logit	Standart Error	Infit	Outfit	Obs %	Exp %	Rasch Kappa	t-values
SKP01	0.23	0.06	1.18	1.16	46.90	30.20	0.239	4.60
SKP02	-0.71	0.05	1.33	1.34	41.80	30.90	0.158	-14.20
SKP03	0.09	0.05	0.85	0.86	45.90	31.30	0.213	1.80
SKP04	0.18	0.05	0.88	0.90	40.90	29.80	0.158	3.60
SKP05	-0.27	0.05	1.45	1.47	44.10	31.30	0.186	-5.40
SKP06	0.09	0.05	0.82	0.85	44.50	31.00	0.196	1.80
SKP07	0.03	0.05	1.27	1.19	43.30	31.80	0.169	0.60
SKP08	-0.07	0.05	0.91	0.94	45.70	31.20	0.211	-1.40
SKP09	0.03	0.05	1.08	1.11	43.70	31.00	0.184	0.60
SKP10	-0.62	0.05	1.26	1.20	44.40	31.10	0.193	-12.40
SKP11	-0.08	0.05	1.39	1.29	45.20	31.50	0.200	-1.60
SKP12	-0.27	0.05	1.45	1.41	40.90	30.90	0.145	-5.40
SKP13	-0.65	0.05	1.14	1.23	35.60	29.70	0.084	-13.00
SKP14	-0.38	0.05	0.87	0.86	46.80	31.40	0.224	-7.60
SKP15	-0.32	0.05	0.92	0.92	45.60	31.50	0.206	-6.40
SKP16	0.44	0.06	1.45	1.49	42.30	31.40	0.159	8.80
SKP17	0.04	0.05	1.05	1.07	47.20	31.70	0.227	0.80
SKP18	0.27	0.06	0.96	0.96	45.20	31.70	0.198	5.40
SKP19	-0.17	0.05	0.81	0.85	42.10	31.70	0.152	-3.40
SKP20	0.09	0.05	0.96	0.95	45.10	31.40	0.200	1.80
SKP21	0.10	0.05	1.00	0.98	43.70	31.80	0.174	2.00
SKP22	0.00	0.05	0.99	0.96	46.00	31.90	0.207	0.00
SKP23	-1.08	0.05	0.78	0.82	45.10	31.60	0.197	-21.60
SDP01	0.19	0.05	0.61	0.63	42.20	33.20	0.135	2.80
SDP02	0.26	0.06	1.40	1.49	39.70	30.20	0.136	5.20
SDP03	0.29	0.06	0.90	0.89	46.70	32.90	0.206	5.80
SDP04	1.12	0.07	1.23	1.28	44.60	32.30	0.182	22.40
SDP05	0.04	0.05	0.69	0.75	39.70	31.50	0.120	0.80
SDP06	-0.08	0.05	0.94	1.04	44.60	32.30	0.182	-1.60
SDP07	0.12	0.05	0.83	0.89	41.60	31.70	0.145	2.40
SDP08	-0.22	0.05	0.62	0.64	41.00	32.20	0.130	-4.40
SDP09	-0.16	0.05	0.61	0.62	38.90	30.40	0.122	-2.90
SDP10	0.05	0.05	1.08	1.02	42.20	31.80	0.152	1.00
SDP11	-0.35	0.05	0.87	0.96	41.50	31.00	0.152	-7.00
SDP12	-0.72	0.05	0.64	0.67	35.00	29.50	0.078	-14.40
SDP13	0.43	0.06	0.74	0.78	42.60	29.80	0.182	8.60
SDP14	-0.07	0.05	0.70	0.71	41.50	29.10	0.175	-1.40
SDP15	-0.10	0.05	1.28	1.30	39.20	29.40	0.139	-2.00
SDP16	0.20	0.05	0.70	0.73	42.80	29.30	0.191	4.00
SDP17	0.45	0.06	0.93	0.91	46.70	28.90	0.250	9.00
SDP18	0.03	0.05	1.39	1.38	38.70	28.30	0.145	0.60
SDP19	0.38	0.06	1.06	1.07	37.70	31.50	0.120	7.60
SDP20	0.39	0.06	0.84	0.84	43.60	29.40	0.201	7.80
SDP21	0.31	0.06	0.76	0.80	45.30	30.20	0.216	6.20
SDP22	0.49	0.06	1.02	1.00	24.00	23.90	0.001	9.80
Mean	0.00	0.05	1.00	1.02				
S.D.(Population)	0.38	0.00	0.24	0.24				
S.D. (Sample)	0.39	0.00	0.24	0.24				
Model. Population : RMSE = 0.05 Adj. (True) S.D. = 0.38 Separation = 7.06 Strata = 9.75 Reliability (not inter-rater) = 0.98								
Model. Sample: RMSE = 0.05 Adj. (True) S.D. = 0.38 Separation = 7.14 Strata = 9.86 Reliability (not inter-rater) = 0.98								
Model. Chi-Square (Fixed) : 2.334.60 d.f. = 44 *significance (probability)* = .00								
Model. Chi-Square (Normal) : 43.10 d.f. = 43 *significance (probability)* = .46								

Note. SKP: rater who took the post-test from the control group SDP: rater who took the post-test from the experimental group

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 2, 515–533
https://doi.org/10.21449/ijate.993805
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Comparison of Inter-Rater Reliability Techniques in Performance-Based Assessment

Sinem Arslan Mancar[1,*], **H. Deniz Gulleroglu**[2]

[1]Independent Researcher
[2]Ankara University, Faculty of Educational Sciences, Department of Educational Sciences, Educational Measurement and Evaluation, Ankara, Türkiye

ARTICLE HISTORY
Received: Sep. 10, 2021
Revised: Jan. 29, 2022
Accepted: May 17, 2022

Keywords:
Inter-rater reliability, Performance-based assessment, Generalizability theory, International baccalaureate diploma programme, Scientific literacy.

Abstract: The aim of this study is to analyse the importance of the number of raters and compare the results obtained by techniques based on Classical Test Theory (CTT) and Generalizability (G) Theory. The Kappa and Krippendorff alpha techniques based on CTT were used to determine the inter-rater reliability. In this descriptive research data consists of twenty individual investigation performance reports prepared by the learners of the International Baccalaureate Diploma Programme (IBDP) and also five raters who rated these reports. Raters used an analytical rubric developed by the International Baccalaureate Organization (IBO) as a scoring tool. The results of the CTT study show that Kappa and Krippendorff alpha statistical techniques failed to provide information about the sources of the errors causing incompatibility in the criteria. The studies based on G Theory provided comprehensive data about the sources of the errors and increasing the number of raters would also increase the reliability of the values. However, the raters raised the idea that it is important to develop descriptors in the criteria in the rubric.

1. INTRODUCTION

The characteristics that individuals should possess in the 21st century have become highly differentiated and diversified, compared to previous centuries. A new generation of learners should be capable of collaborating and managing the complexities of the global world. Getting ahead in 21st century society requires acquiring a set of critical skills and adopting specific characteristics. Apart from the general knowledge and skills that learners should have, they are expected to be "global citizens" who have the ability to use basic sciences to solve the problems encountered in daily life by applying their advanced critical thinking, problem-solving, productivity, creativity, communication, awareness of ethical rules, information literacy, technology literacy, global awareness, innovation, and collaboration skills effectively (Ananiadou & Claro, 2009; MEB, 2016; National Research Council, 2012; OECD, 2017; Partnership for 21st Century Learning, 2007). This means that learners need to communicate effectively, think critically, analyse local and global issues, challenges and opportunities, become information literate, reason logically, interpret scientific data in terms of cognitive

*CONTACT: Sinem Arslan Mancar snmars89@gmail.com Independent Researcher

e-ISSN: 2148-7456 /

competencies, play a key role as a team member, cooperate with others, be aware of the importance of social impact in terms of interpersonal competencies, be aware of the significant impact of ethics, and have intellectual openness and self-regulation in terms of intrapersonal competencies (Collins, 2014; IBO, 2014a; IBO, 2014b; IBO, 2014c; Marzano & Heflebower, 2012; National Research Council, 2012; Schleicher, 2015; Trilling & Fadel, 2009; Uçak & Erdem, 2020).

Today's learners have started to live in the information age because of growing up in a fast-paced digital world. Moreover, technological innovations have accelerated the transmission and processing of information. These aspects related to information have also revealed the "information literacy" and the concept of the term has been determined as one of the important learner skills. Information literacy was defined by Paul G. Zurkowski, the president of the Information Industry Association in 1974 as "the person who uses scientific information resources effectively to reach a knowledge-based solution related to problems, and who has the skills to use various information sources" (p.6). Information literate individuals who have developed scientific thinking skills and who could use science for personal and social purposes are candidates for being a scientific literate. The definition and components of information and scientific literacy concepts have evolved together with the times as one of the most fundamental and continuous parts of the scientific process is information literacy (Klucevsek, 2017).

Scientific literacy is defined as the use of scientific knowledge by a global citizen in order to identify science-related issues, draw conclusions with a scientific method, and utilize that knowledge for the benefit of society and the individual (Bybee, 1997; Holbrook & Rannikmae, 2009; Hurd, 1998; Maienschein, 1998; Nbina & Obomanu, 2010; OECD, 2017; Turgut 2007). Being a qualified scientific literate requires being able to explain the facts and concepts scientifically, develop and evaluate scientific inquiry methods and interpret the findings logically (MEB, 2016; OECD, 2017; Rychen & Salganik, 2003). It is known that scientific literacy skills are tested in standardised tests globally such as the International Mathematics and Science Trends Research (TIMSS) and the International Student Assessment Program (PISA) and in national or international educational programmes such as International General Secondary Education Certificate (IGCSE), Advanced Level (A-Level) and International Baccalaureate (IB) (IBO, 2014a; IBO, 2014b; IBO, 2014c; Mullis & Martin, 2017; OECD, 2017; Syllabus Cambridge IGCSE Global Perspectives, 2015). In terms of scientific literacy, TIMSS tests are based on a comprehensive analysis of mathematics and science curricula and mainly focus on facts and processes while PISA tests measure mathematics and scientific literacy skills, as well as the application of these skills to real-life situations (OECD, 2017). In science literacy skills of A–level and the International Baccalaureate (IB) Diploma Programme (DP) (also known as IBDP) learners are assessed by both performance assessment and final exams (IBO, 2014a; IBO, 2014b; IBO, 2014c; Cambridge International Examinations, 2015).

Scientific literacy is assessed by national or international tests and educational programmes, as abovementioned. While the characteristics today's learners should have are so diverse, it is inevitable that the assessment and measurement tools be used to assess the relevant characteristics and also change, transform or diversify. Dietel, Herman, and Knuth (1991) define assessment as "any process and test used to learn more about the current level of knowledge possessed by the learner" (online document). Testing is defined as a "single-occasion, unidimensional, time- based" usually in the form of a multiple choice or short answer (Law & Eckes, 1995). Learners were assessed only by true-false tests, multiple choice tests and short-answer tests for a long time. Currently, due to the nature of the 21st century learner, it is realized that there is not only one way of gathering information about learner learning as alternative assessment tools are supportive approaches to the assessment of learner's higher-order skills with the traditional assessment tools (Coombe *et al., 2012)*. Furthermore, testing is

viewed as just one aspect of assessment, and the term "assessment" is widely used (Kulieke *et al.*, 1990).

In alternative assessment, there are three approaches: Authentic, performance based, and constructivist (Simonson *et al.,* 2000). Similarly, Reeves (2000) suggests that three key approaches be used in assessment; namely, cognitive, performance, and portfolio. As researchers and educators use the terms "performance based assessment," "alternative assessment", and "authentic assessment" interchangeably, performance based assessment will be used to refer to alternative assessment and discussed throughout this study. Tasks and context in performance based assessment are more closely aligned with learners' context in the classroom and in real life situations. In other words, the nature of the task and context in which assessment takes place represents real life problems or issues (Coombe *et al.,* 2012*)*. Therefore, performance based assessment is a valuable tool to observe learners' skills as to how to use science knowledge to solve problems encountered in daily life as it is compatible with the nature of scientific literacy (Kutlu *et al.,* 2008). Performance tasks and contexts enable learners to apply their skills to various simulations related to real life simulations.

Performance based assessment tools are also based on the process of learning which focuses on the growth and the performance of the learner. According to Law and Eckes (1995), if learners fail to perform a given task or context at a specific time, they can still demonstrate their abilities at a later stage and in a different situation as it is not a one-time test. Furthermore, performance based assessment focuses more on the process than on the product (pass or fail), which makes assessment formative. As a result, teachers may monitor and assess their learner's strengths and weaknesses in a variety of scenarios and can improve their syllabi based on the needs of the learners (Law & Eckes, 1995; Reeves, 2000). For this reason, performance based assessment also tends to prioritize more individualized and constructive feedback.

The key feature of performance based assessment is that the learners need to create their own work such as projects, portfolios, reports, experiments, or performance, which is scored against specific criteria (Kutlu *et al.,* 2008; Simonson *et al.,* 2000). In this context, various assessment tools such as checklists, grading scales, and rubrics are used by educators and researchers (Aktaş & Alıcı, 2017). A rubric that includes the specification of the skill being examined and the constituents of various levels of performance success is defined as a set of achievement criteria with the highest and lowest degrees (Callison, 2000). Constructing an appropriate rubric is the core element to meaningful performance based assessment and there are two types of rubrics commonly used to score learners' performance; namely, holistic and analytical (Mertler, 2001; Moskal, 2000).

Holistic rubrics that assess a learner's overall performance and achievement on a qualitative level provide an overall description of various levels of performance and result in a single score or grade (Goodrich Andrade, 2001; Moskal, 2000). Holistic rubrics can also be developed and applied more rapidly. By contrast, analytic rubrics view performance as being made up of many components and provide separate scores, indicators, and descriptions for each component. The educators can monitor a reflector's performance against each of the well-defined assessment criteria (Mertler, 2001). Then, by collecting the scores calculated separately, the total score related to the performance is obtained (Moskal, 2000). Therefore, it provides more detailed information that may be useful when providing feedback.

Performance-based assessment raises some concerns about subjectivity, reliability, and validity. One of the crucial points of performance-based assessment is to conduct highly reliable measurement and evaluation practices to make accurate decisions about the learners. Analytic rubrics are often preferred with the advantage of dividing the performance process or product into specific sections, ensuring that these sections are scored to meet predetermined criteria. In this case, it is thought that errors caused by the person who measured during the scoring process,

in other words, by the rater, will have less impact. However, determining whether the aspect to be measured exists in the individual based on the opinion of a single rater may also decrease the reliability of the assessment. Accordingly, it is believed that assessments with more than one rater will increase reliability (Abedi *et al.*, 1995). On the other hand, the raters are considered as a significant source of error in the assessments made based on the opinion of the rater (Airasian, 1994; Anadol & Doğan, 2018). At this point, while the increase in the number of raters is crucial for the accuracy of the decisions taken, the higher number of raters is seen as a potential source of error that is thought to be involved in the measurement. Accordingly, various error sources may be encountered such as the individual characteristics of the raters, the number of the raters, the differences of the raters' opinions, and the surrounding variables affecting the rater (Turgut & Baykul, 2010). The measurements are objective to the extent that the raters are given the same score on the same answer, and only in this case the rater reliability is ensured (Shavelson & Webb, 1991; Turgut & Baykul, 2010). In performance-based assessments, before making decisions about individuals, it is necessary to examine the consistency between raters to determine the reliability of the measurements made.

There are many methods and techniques to analyse inter-rater reliability based on Classical Test Theory (CTT), Item Response Theory (IRT) and Generalizability Theory (G Theory) (Baykul, 2015). The variety of theories and techniques causes differentiation of the reliability coefficients obtained, but also provides different information from applications. In this study, the consistency between different numbers of raters was analysed by using Kappa and Krippendorf alpha statistical techniques based on CTT. Within the scope of the G Theory, between the raters of the fully crossed pattern (s x i x r), the G and Phi coefficients that emerged because of the D study were determined and inter-rater reliability analyses were conducted.

In CTT (Lord, 1959; Novick, 1966), observed scores (X) from psychometric instruments are thought to be composed of a true score (T) that represents the subject's score that would be obtained if there was no measurement error, or an error (especially random errors) component (E) that is due to measurement error, such that "Observed Score = True Score + Measurement Error", or in abbreviated symbols, "X = T + E" (Baykul, 2015). Such errors may arise from the individual's performing measurements, the properties measured, the measuring environment, and the measuring technique (Atılgan *et al.*, 2007; Shavelson & Webb, 1991). Since the reliability coefficient for only one type of error is calculated at one time with the CTT, it is necessary to analyse each possible source of error separately. In addition, the inability to calculate the interaction of error sources together seems to be a limitation for the CTT. However, the limitations of techniques developed based on CTT to be used in determining rater reliability revealed the need to examine these techniques.

In this research, the Kappa statistical technique was chosen, because it was not affected by the subcategories included in the analytic rubric and it showed consistency only because of the change in the number of raters. However, analysis was carried out using the Fleiss Kappa statistical technique, since there were more than two independent categories and the need to determine the consistency of two independent and more than two scoring points independently. Thus, the consistency between the assessments of different numbers of raters independent from each other was determined. Krippendorff's statistical technique was preferred due to its advantages such as being used in this study in different number of sample cases, being easily applied to each scale type, and being used in cases where the number of raters is more than two.

The G Theory which was founded in 1940 is a continuation of CTT and Analysis of Variance (ANOVA). The fact that various and many error sources can be determined separately with a single analysis in G Theory increases the importance and usefulness of the theory. In addition, obtaining the Coefficient of Reliability (G) that reveals errors arising from the interaction of various error sources both individually and with each other is the reason why G Theory is

preferred (Brennan, 2001; Shavelson & Webb, 1991). In this research, the variances arising from the items in the analytic rubrics and the raters were determined and possible sources of errors were interpreted. In D studies, consistency analyses are performed in cases where different numbers of raters are included, and suggestions are developed for this situation.

A review of the related literature both nationally (Atılgan, 2005; Bıkmaz Bilgen, 2017; Büyükkıdık, 2012; Güler, 2009; Güler; 2011; Özmen Hızarcıoğlu, 2013) and internationally (Abedi *et al.,* 1995; Goodrich Andrade, 2001, Gwet, 2002; Lane & Sabers, 1989; Marzano, 2002; Oakleaf, 2009) shows that there are many studies on inter-rater reliability. However, there is no related study conducted in the field of IBDP, one of the programmes that are internationally accepted and has standardized assessment and evaluation practices. "Individual Investigation" is a core part of the internal assessment for science subjects. Learners select a real-life issue from Physics, Chemistry, or Biology and investigate it in order to produce a scientific report about it. The main aim of this component is to convert a situation that learners wonder about a scientific issue and solve it by using a scientific method. Within the scope of this aim, learners are expected to become individuals who are aware of the nature of science, have analytical and critical thinking skills, have an ability to apply scientific research methods, and use their scientific knowledge effectively in solving real life issues. However, they produce a scientific report in which they demonstrate these skills under the guidance of the teacher (IBO, 2014a; IBO, 2014b; IBO, 2014c; IBO, 2015). It is clearly seen that individual investigation work and its report are exceptionally good examples of performance based assessment and context of investigation coincides with scientific literacy skills. Therefore, scientific literacy skills of the learner are assessed through an individual investigation process and learner's report. In this specific research, the analytic rubric used in the assessment of learner's performance was prepared by IBO experts (IBO, 2015).

This research is thought to reveal whether the analytic rubrics used internationally is used effectively, to ensure their deficiencies, if any, and to contribute to the IBDP internal assessment process since it will serve as an example for performance studies on the assessment of scientific literacy at national and international levels to be carried out in the future. In addition, due to the COVID-19 pandemic we experienced, IBDP final exams were cancelled. While calculating the graduation scores of learners about science courses, individual investigation reports in this research would be predominantly taken as a basis. At this point, inter-rater reliability has become even more important. For this reason, it is thought that the research would serve as an example for the importance of the number of raters in determining the performance based assessment and the reliability of the decisions taken.

In line with the main aim of the research, this research strives to address the following research questions:

- Is there a statistically significant difference between the scores obtained from two, three, and five raters according to Kappa statistical technique?
- Is there a statistically significant difference between the scores obtained from two, three, and five raters according to Krippendorff alpha statistical technique?
- In the pattern where all sources of variability are fully crossed (s x i x r), does the consistency between different numbers of raters differ significantly in the G and Phi coefficients?
- Are the reliability coefficients obtained from the analysis findings based on the CTT and G theory consistent with each other?

2. METHOD

2.1. Research Model

This research is built on the basis of applying different techniques based on CTT and G Theory in order to analyse the level of inter-rater reliability, examining their restrictions and finding out which of these techniques provides more comprehensive and reliable information. Since the purpose of the research is to reveal the existing situation, it is a descriptive research (Bailey, 1994; Büyüköztürk *et al.*, 2012).

2.2. Study Group

The study group of the research consists of five raters (or teachers) who scored twenty individual investigation reports prepared by the learners within the scope of IBDP internal assessment for Biology subject. The raters had between five and twenty years teaching experience in national education system, however, they had been teaching IBDP Biology for two to ten years and all had a Teaching Certificate.

Evidently, there will always be differences of interpretation of the criteria - and this may vary from person to person and from sample to sample. As an IB requirement, teachers need to meet to discuss the analytic rubric. The participants should agree on common standards at the start of the assessment and be consistent throughout. Teachers of the same science subject should mark two or three individual investigations each. They should then mark their colleagues' learners' individual investigations using the same process they used to mark their own learners' individual investigations. Afterwards, a standardization meeting should be held to determine the level of marking. Internal harmonization of marks is clearly seen as critical to obtaining reliable and valid results at the end of the assessment (IBO, 2018).

Therefore, in the line with IB guidance, in this study 2, 3 and 5 raters were chosen to assess learners' reports who had a similar internal assessment experience year in IBDP curriculum. Inter-rater reliability is applied in situations where different assessors or raters provide subjective judgment on the same target (Viera & Garret, 2005). For this reason, there should be at least two, if possible three raters, as the reliability value obtained determines how much the raters agree on the scoring of a particular target (Burry-Stock *et al.*, 1996). The reason for choosing five raters is to observe whether the increase in the number of raters significantly changes the reliability or not.

2.3. Data Collection Tools

Individual investigation is the core component of the internal assessment of the science subjects in IBDP. Learners choose one of the real-life issues in Physics, Chemistry or Biology and work on it to carry out their investigation and produce a report about it. In this research, the Biology individual investigation reports of learners who graduated from the programme in the same year were used (IBO, 2018).

2.3.1. *Analytic rubric*

In the study, analytic rubric, the basis for assessing individual investigation reports of IBDP Biology subject and developed by IBO experts and also used for the same purpose in schools that implement the programme in all countries, was used to assess the individual investigation reports of IBDP biology subject within the scope of internal assessment (IBO, 2014a).

The internal assessment requirements and analytic rubric are the same for biology, chemistry, and physics. The internal assessment, worth 20% of the final assessment, consists of one scientific investigation. The individual investigation should cover a topic that is commensurate with the level of the course of study. Learner work is internally assessed by the teacher and externally moderated by the IB examiners.

Assessment criteria should be specifically matched to any investigation that has been designed to be used to assess learners. For analytic rubric, several assessment criteria have been identified. There is a level descriptor that describes specific levels of achievement and performance, and a range of marks associated with those levels, for each assessment criterion. Teachers, or raters, are required to judge the learner's work against the level descriptors. Each of the performance levels is described with multiple indicators. There are many cases in which the indicators occur together at a specific level, but not always. In addition, not all indicators are present at all times. As a candidate's performance can fit in different levels, IB assessment models use bands of marks and recommend that teachers and examiners use a best-fit approach to deciding the appropriate mark for a particular criterion. In other words, compensation should be given for work that meets various aspects of a criterion at different levels. For a mark to be awarded, it is not necessary to meet every aspect of a level descriptor. The mark should reflect the achievement balanced against the criterion. The teacher should read each of the level descriptors until they find the one that most accurately describes the level of the work. The learner's work should be read again if it seems to fall between two descriptors and then the descriptor that more accurately describes the work of the learner should be chosen. If two or more marks are available within a level, teachers should award the higher mark if the learner's work displays the qualities described to a great extent. In other words, learners may be close to reaching a higher level. Marks should only be recorded as whole numbers; fractions or decimals are not acceptable. Teacher should not focus on the pass/fail boundary, but rather identify appropriate descriptors for each assessment criterion. Learners should be able to reach the highest-level descriptors if this is appropriate for the assessment. Teachers should not avoid using the extremes when appropriate for the assessment. If a learner achieves a high achievement level for one criterion, it does not mean that he/she will achieve high achievement levels for the other. Similarly, learners who achieve a low level of achievement for one criterion will not necessarily achieve similar levels of achievement for other criteria. The assessment of all the learners should not be assumed to result in a particular mark distribution for the teacher. Learners should be made aware of the assessment criteria. All explanations about how to use analytic rubric, criteria and descriptors should be available in IB Physics, Chemistry and Biology guides (IBO, 2014a; IBO, 2014b; IBO, 2014c). When it comes to the IB moderation, a sample of the marking of internally assessed work is remarked by a moderator to ensure that marking is accurate. During the process, assessors use statistical comparisons and linear regression techniques to determine the degree to which original teacher marks need to be adjusted to align with the set standards.

Scoring rubrics may be designed to contain both general and task specific components. The analytical rubric used in this research is a good example of this situation. The purpose of an individual investigation is to evaluate learners' scientific literacy skills and their scientific knowledge of the chosen topic. This analytic rubric used contains both a general component and a task specific component.

The IBDP analytic rubric uses five criteria with 24 points, in order to assess the final report of an individual investigation, with these raw marks and weightings assigned: personal engagement (up to 2 points/8%), exploration (up to 6 points/25%), analysis (up to 6 points/25%), evaluation (up to 6 points/25%), and communication (up to 4 points/17%). Personal engagement assesses the extent to which the learner has mastered her/his research, how she/he designed and applied it and how she/he presented it in the report. Exploration assesses the extent to which there is clear explanation of the learner's research question and supports with research and theories by reviewing the literature in this direction and completing its work in a safe, environmental, and ethical manner. Analysis assesses the extent to which some criteria such as collecting, analyzing data, and being aware of the impact of the results of the analysis on the research reflect the research situation of the learner. Evaluation assesses the

extent to which the research is supported by relevant theories, defining its strengths and weaknesses, expressing the limitations and errors, interpreting the data obtained, discussing comprehensively, and presenting suggestions based on these data. Communication assesses the extent to which the research is well structured and focuses on the research question and the clear expression of relevant information accordingly.

2.4. Data Analysis

Kappa statistical technique is the first technique applied in this study in order to determine the inter-rater reliability. Although it is often mentioned in the literature about Cohen's Kappa, Fleiss Kappa technique is preferred in cases where there are more than two raters (Cohen, 1960). In this study, Fleiss Kappa technique was used. For the Kappa statistics, SPSS syntax (stats fleiss kappa [v4]. sps)" script was used in SPSS 21.0 software. Then, for the analysis, the reliability of 2, 3 and 5 raters for five different criteria in the analytic rubrics was examined, respectively.

Krippendorff alpha technique was preferred as it can be applied to any scale level. For the Krippendorff alpha technique, "SPSS syntax (kalpha.sps)" was used in SPSS 21.0 software. Then, 2, 3 and 5 raters were calculated for both criteria and total score in the analytic rubrics to observe the consistency.

In studies based on G Theory, each rater in the rater group consisting of two, three, and five raters score each performance report in the research in a way that corresponds to the items in the analytic rubrics. In this study, the raters assessed twenty learner biology reports written for an internal assessment. In this context, the pattern used in the study is a fully crossed pattern and is expressed as (s x i x r). Accordingly, analyses were conducted in order to determine how the variance components and the percentages of these components in the total variance changed with the number of raters. EduG 6.1 software was used for statistical analysis based on G Theory in the analysis. In this context, G and Phi coefficients were determined and D study was included. In cases that occur with the change in the number of raters, the change of G coefficient is observed by conducting D coefficient study.

3. RESULTS

The findings are presented in order in which the subproblems of the research are given and interpreted. The Kappa and Krippendorff alpha statistical techniques were interpreted by calculating the inter-rater reliability values both separately for each criterion and in terms of total scores. When scoring with two, three, and five raters within the scope of the first and second sub-problems of this research, the consistency of the scores obtained was analysed by Kappa and Krippendorff alpha statistical techniques and the findings are summarized in Table 1.

When Table 1 is examined, the negative and positive values of the findings related to Kappa statistics are seen in the scores obtained from different numbers of raters. That the Kappa value (κ) is negative indicates that the agreement between two or more raters is less than expected by chance, A (–1) value for Kappa indicates no observed agreement (i.e., the raters agree on nothing), and (0) (zero) value indicates no agreement. According to Agresti (2013), negative reliability values rarely occur; however, these values were observed in this research. Even though the reliability values (Fleiss kappa coefficients) between the raters are significant, it is worth noting that these values are very low. It can be because of the fact that both low inter-rater agreement and a lack of clearly defined criteria in the rubric lead to low and negative values (Fleiss, 1971).

Table 1. *Kappa and Krippendorff's Alpha Statistical Values Regarding the Scores of Different Number of Raters.*

Number of raters	Criteria	Kappa statistical value (κ)	Krippendorff's alpha value
2	Personal Engagement	0.076*	0.026*
	Exploration	-0.026*	0.059*
	Analysis	0.133*	0.372*
	Evaluation	0.281*	0.571*
	Communication	-0.028*	-0.258*
	Total score	0.228*	0.440*
3	Personal Engagement	-0.006*	-0.073*
	Exploration	-0.049*	-0.039*
	Analysis	-0.078*	0.252*
	Evaluation	0.112*	0.337*
	Communication	0.106*	0.098*
	Total score	0.120*	0.288*
5	Personal Engagement	0.074*	0.066*
	Exploration	-0.014*	0.125*
	Analysis	0.054*	0.303*
	Evaluation	0.158*	0.503*
	Communication	0.108*	0.150*
	Total score	0.163*	0.373*

* $p<0.001$

In the condition that there are two raters, Kappa values change between -0,028 and 0,281. In this case, the lowest level of agreement is in the "communication" criterion (κ = -0.028); the highest level of agreement is estimated in the "evaluation" criterion (κ = 0.281). In the "exploration" criteria learners are expected to establish the scientific context and also they need to put a clear research question, as well as ideas or skills explored in the syllabus. Another criterion where raters scored differently from each other was "personal engagement." In this criterion, the learner is expected to reflect on the subject: why she/he chooses the subject, and how she/he uses the individual characteristics and skills she/he has while exposing the subject. These two criteria differ from one learner to another, as well as from one rater to the other. In this case, it can be thought that the criteria are perceived differently by the raters and create different expectations. In the "evaluation" criterion, the learners are expected to interpret the analysis results, make inferences, and analyse the results together with their previous knowledge. Accordingly, it is observed that learners' research is designed to meet the expectations of the raters of this section, even partially. When Table 1 is analysed, it is seen that, Kappa values are not too high or do not even get negative values. Negative values indicate low inter-rater agreement and raters make different evaluations from each other (Agresti, 2013; Fleiss, 1971). However, with the Kappa technique, no information can be obtained about the sources of errors causing no-agreement between raters. When looking at the overall inter-rater agreement across the overall score, the Kappa value (κ = 0.228) indicates low agreement (Landis & Koch, 1977).

It was determined that the mismatch regarding "personal engagement" and "analysis" criteria increased in the measurement involving three raters. This may be because the relevant items are not correctly understood by the raters, or the raters' expectations for these criteria are different. However, overall inter-rater agreement is lower than the situation where two raters

are present. The biggest difference in the negative direction was in the “analysis” criterion. "Analysis” is one of the criteria that should be prepared comprehensively by supporting various data in scientific studies (IBO, 2015). A criterion in the relevant criteria may differ from one rater to another in some way. When looking at the overall agreement among the three raters, the Kappa value indicates a low agreement (Landis & Koch, 1977). Kappa values appear to decrease as the number of raters increases.

The criteria where the five raters diverged the most were the “exploration” criterion. The “evaluation” criterion was the criteria in which raters agreed, albeit partially. However, when looking at the overall agreement between the five raters, the Kappa value indicates a low level of agreement (Landis & Koch, 1977).

According to the Krippendorff’s alpha values in Table 1, it is seen that there is a relatively high level of agreement between two raters in the “analysis” with (α= 0.372) and also “evaluation” (α= 0.571) criteria. It is much higher than other criteria. For the “personal engagement” with (α= 0.026) criterion, it is seen that the raters scored quite far from each other. The reason for this is the criterion in which these criteria reveal the individual characteristics of the learners and scores whether their studies are designed and expressed well or not (IBO, 2015).

According to the IBO (2018), scientific reports are produced at a particular time by learners. As teachers have been moderated by IBO each year, they are aware of how to use the analytic rubric in a good standard and try to standardise their assessment of learners’ work to ensure reliable results in accordance with IB guidelines. However, there are still error sources such as learners, raters, the development process of the performance task, and the analytic rubric. For example, descriptors for some of the criteria may not be sufficiently expressed in the analytic rubrics or raters struggle to use analytic rubrics though they have used them before. Therefore, it is not possible to determine these potential situations and errors with the Krippendorff alpha statistical technique (Krippendorff, 2004).

When the Krippendorff’s alpha values calculated for the three raters are examined in Table 1, it is seen that the agreement rate of the “analysis” with (α= 0.252) and “evaluation” with (α= 0.337) criteria is higher than the other criteria. All the criteria except the “communication” with (α=0.098) criterion were negatively affected by the increase in the number of raters. Accordingly, it can be stated that the scores obtained from the criteria are not reliable (Krippendorff, 2011). In addition, the divergence of Krippendorff alpha values can be based on the level of objectivity of the criteria.

According to Table 1, when the Krippendorff’s alpha values calculated for the five raters are analysed, the highest agreement can be seen at “evaluation” with (α= 0.503) criterion. In the case of five raters, no negative values were found. It can be thought that the raters do not score differently enough to reach a negative level. In the ranking of the rater reliability of the criteria; as in the two and three raters, a higher level of agreement was seen in “analysis” and “evaluation” criteria than that of the others. When looking at the overall inter-rater agreement, it was found that this ratio could not reach even fair agreement (Krippendorff, 2011).

Regarding the third research question of the study, the analyses were carried out in a fully crossed pattern (s x i x r) and the variance components estimated for the learner. However, in this part, student, and s, refers to the learner, student (s), item (i) (called as criteria) and rater (r) as given in Table 2. When the variance and total variance explanation percentages as a result of the G study in Table 2 are examined, it is seen that the variance component of the main effect of the students corresponds to 9% of the total variance. The variance component of the students gives an estimate of how students' performance studies change from one student to another. The variance component of the students is ($\sigma^2b = 0.227$) and it is expected to be at a high rate as the differentiation of the students’ characteristics affects consistency. Since performance studies

are the studies that students manage the process themselves and produce a product at the end of the research, errors arising from students, or the measured feature may interfere in measurement and evaluation practices (Brennan, 2001).

The variance component ($\sigma^2m = 1.121$) estimated for the main effect of the item has the highest variance value in the total variance with 44% of the total variance and is identified as the most important source of variability among all variance sources. In this case, it is believed that students may not be able to provide the necessary and effective performance report for each item and that the ratings of the criteria differ among raters. However, since each item in the analytic rubrics measures the skills related to performance, this rate is expected to be high (Güler & Taşdelen, 2015). It should be noted, however, that these criteria try to measure skills that are not distant from each other.

Table 2. *The Variance Components and Total Variance Percentages Obtained as a Result of the G Study of the Pattern (s x i x r).*

Variance Source	Square Total	df	Mean of Squares	Variance	Percentage of Variance (%)
s	152.952	19	8.051	0.227	9.0
i	466.656	4	116.638	1.121	44.4
r	29.472	4	7.368	0.037	1.5
si	150.480	76	1.981	0.260	10.3
sr	79.968	76	1.052	0.075	3.0
ir	51.368	16	3.210	0.126	5.0
sir	205.19.2	304	0.675	0.675	26.7
Total	1.135.992	499			100%
G = 0.90 Phi = 0.90					

It is observed that variance from the rater constitutes 1.5% of the total variance. The variance value ($\sigma^2p = 0.037$) calculated for the rater effect was found low. The variance component of the raters provides the opportunity to make an estimate of how the raters give their scores on performance studies. It shows that the raters have a low role in the differentiation of scores. The low percentage of total variance explanation of the variance component of the raters can be interpreted as independent raters make scoring consistent with each other.

(student x item) interaction provides information on whether students' performance reports differ according to the criteria in the analytic rubrics (Shavelson & Webb, 1991). As can be seen in Table 2, the student x item interaction has the highest variance value in total variance. This situation can be interpreted as students' performance reports differ from one criterion to another. It also shows that each criterion measures different skills. A student may qualify for one criterion, but not for another (IBO, 2014a; IBO, 2014b; IBO, 2014c). The criteria are composed of a range of related skills that candidates should be able to demonstrate at various levels of accomplishment. The requirement of each criterion is different from each other in the analytic rubric. The achievement level descriptors for each criterion, which describe the typical ways in which a candidate can be assessed in accordance with the criterion, are used to describe differences in candidate achievement that result in a different mark. The final mark is determined by adding up the maximum levels of achievement for each criterion. Internal consistency measures of reliability are not considered appropriate because each component (assessment tool) may deliberately contain varied forms of task, or sometimes a small number of tasks (IBO, 2018). However, as an item has a/significant effect on reliability with the higher

variance value (44%), increasing the number of items may increase the impact of "student x item" interaction (Brennan, 2001).

"item x rater" is the variance of the common interaction ($\sigma^2mp = 0.126$), which creates a 5.0% effect in the total variance. This indicates that there is no significant level of difference in the scoring consistency between the raters. As can be seen in Table 2, the (student x item x rater) variance component indicates 26.7% of the total variance. This common effect is the second-high variance in total variances. That the G and Phi coefficients are 0.90 means that the scoring reliability is high; in other words, the inter-rater agreement is high (Atılgan, 2005; Brennan 2001; Shavelson & Webb, 1991).

D study investigates the impact of variability among the scores from different numbers of raters. D study conducted within the scope of G Theory analysis allows researchers to calculate two different reliability coefficients that are effective in making both relative decisions based on students' performances and absolute decisions regarding students' performances (Shavelson & Webb, 1991). Researchers benefit from G coefficient in making relative decisions, and from Phi coefficient in making absolute decisions. The study findings carried out to examine the effect of the D study and the numbers of raters on the G and Phi coefficients are given in Table 3.

Table 3. *G and Phi Coefficients of Pattern (S X I X R) Estimated by D Study.*

Measurement pattern	Number of Items	Number of raters				
		nr=1	nr=2	nr=3	nr=4	nr=5
s x i x r	5	G=0.65 Phi=0.64	G=0.79 Phi=0.78	G=0.85 Phi=0.84	G=0.88 Phi=0.88	G=0.92 Phi=0.91

As can be seen in Table 3, the increase in the number of raters causes an increase in G and Phi coefficients. It can be concluded that the G and Phi coefficients are estimated higher, and the number of raters has a significant impact on scoring reliability in cases created using a different number of raters and the same pattern. In addition, it is clearly seen in Table 3 that the Phi coefficient, which is important for this study, was positively affected by the rater increase. In assessing performance studies, although it seems ideal in theory, it may not always be possible to reach five raters in practice. In this case, making assessments with three or four raters, if possible, can lead to more reliable results and accurate decisions about students.

According to the results of analysis based on CTT and G Theory, the fourth research problem of the study was interpreted within the scope of the findings obtained. The reliability coefficients obtained from the analysis were not consistent with each other. Kappa and Krippendorff alpha statistical techniques used for the analysis based on CTT showed a low level of agreement between raters. In both techniques, consistency between raters showed negative values on many criteria in the analytic rubric. According to the analysis based on CTT, it is not possible to determine the ideal number of raters because the values vary from two raters to five raters. However, the results which were obtained by two and five raters were close to each other. Moreover, it is not possible to determine the sources of errors causing this incompatibility with these techniques. Analyses based on G theory provide the opportunity to interpret many variables both separately and also together. In the research, it is advantageous for the researchers to observe the variances arising from the learners, the items and the raters. In this research, in line with the IBO guide (2018), learners need to have 10 hours to complete their scientific reports with the teacher guidance. They also need to produce a project plan before they carry out their investigations and experiments. Meanwhile, as an IB requirement, teachers support the learner in the line of analytic rubric. As it is a standardised process, in all schools

where this programme is implemented and their educators must follow the same stages, all environmental conditions, limitations and error sources are minimised.

4. DISCUSSION and CONCLUSION

Within the scope of the research, analyses were made by using techniques based on CTT and G Theory and the results obtained were compared in determining the inter-rater reliability levels. Values obtained from two, three, and five raters based on the Kappa statistical technique indicate a low level of agreement when examined in each criterion and in the total score. According to the results of the analysis, the highest agreement between the raters was determined as the situations where two raters were included, and the lowest level of agreement was determined as the situations where three raters were included. Negative values are seen in some criteria, in other words, inconsistency between raters indicate that raters do not make consistent assessments when scoring. According to the analysis findings based on the Kappa statistical technique, the increase in the number of raters has decreased the Kappa value relatively (Nying, 2004). This situation is thought to be an indication that Kappa statistics are affected by the increase in the number of raters. Accordingly, it can be said that it is sufficient to include two raters. These findings coincide with the finding that the increase in the number of raters in the measurement of performance of Abedi *et al.,* (1995) studies decrease reliability by increasing the level of variability in scores. However, it can be stated that the Kappa statistical technique is insufficient in determining the ideal number of raters in determining the performance-based assessment.

Based on the Krippendorff alpha statistics technique, the consistency between raters indicates a low level of agreement when the analyses obtained from different numbers of raters are analysed in each criterion and total score. The highest values indicating the compatibility between raters from the analysis made with Krippendorff alpha technique were calculated in cases related to the situation of two raters, as in the Kappa statistics. This finding coincides with the findings of Bıkmaz Bilgen and Doğan (2017), where the highest agreement was found when there were two raters. However, in both techniques, the analyses in the case of three raters indicate that the inter-rater agreement is at the lowest level. In the Krippendorff alpha technique, as in the Kappa technique, as the number of raters increased, the alpha value changed; however, this change was not as significant as in the Kappa statistic and displayed a relatively more stable structure.

The Kappa and Krippendorff alpha values, the basics of which were developed based on the CTT, calculated the levels of inter-rater reliability exceptionally low. Although there are sources of students, item, and scoring variability in this study, it can be said that these techniques based on CTT are insufficient in reaching the variable that causes negative values and incompatibility. According to the findings of the study, it is seen that Kappa and Krippendorff alpha techniques are insufficient in deciding the ideal number of raters and error sources in assessing the performance reports reflected by the learner characteristics.

Based on the G studies, the effect of the variance originating from the students' items and raters in the measurement process was calculated in the total variance. The study findings related to sources of variance from G study showed that the main source of variance across all criteria was items, while raters represented a relatively small source of variance. It shows that raters were not a significant source of error. Moreover, it means that items measured different kinds of skills and raters could not create a significant impact on the assessment process. Additionally, the common effect resulting from students and items has a high value. This indicates that students show different competencies in different items. This result also shows that each of the items measures different skills, and it is an expected result. It also means the analytic rubric is a reliable measurement tool. According to the D study, it is seen that increasing the number of

raters increases the reliability positively. Especially in cases where five raters are not reached in practice, it can be said that assessing with three or four raters increases the scoring reliability, so that accurate decisions can be made. In the analysis based on G theory, the inter-rater reliability coefficient was higher than that of the Kappa and Krippendorff alpha techniques. Moreover, it was concluded that increasing the number of raters with the D study would increase reliability (Büyükkıdık, 2012; Deliceoğlu, 2009; Güler, 2009; Öztürk, 2011). These results provide more comprehensive data against the limitations of the CTT. However, Kamış and Doğan (2017) state that the G and Phi coefficients, which are obtained in real situations where the raters are not randomly selected from the population and estimated as a result of different decision studies, differ even though they take values close to each other. The G and Phi coefficients obtained as a result of the D studies require that the relevant sources of variability (raters, items, etc.) be randomly selected from the population in the new application to be carried out. Random selection of raters from the population is possible in large-scale measurement applications, but it is practically not possible in-class measurement applications. It is not clear whether the G and Phi coefficients obtained as a result of the D study accurately predict the real situation. The analysis results in this research show that the actual values obtained and predicted in the D studies are similar but differentiated. It is recommended raters should be selected randomly from the population and determine which coefficient will be more accurate to use afterwards.

In addition to the quantitative studies, opinions were received from the raters. These views are primarily the performance reports developed by the teacher and the learner together, while the teachers try to improve themselves to provide reliable feedback to the learner while trying to apply the scientific research methods and steps in the most correct way. However, they think that it is important to elaborate on the descriptions in the criteria, in other words, to make the expressions used to measure the targeted feature clearer. In this context, it would be an appropriate decision to expand the explanations of the criteria in an analytic rubric.

Based on the findings of the research, it can be stated that it is important to use performance based assessment and evaluation approaches in order to observe the learners' characteristics in all aspects. It is seen that the individual investigation steps carried out in the field of biology science within the scope of IBDP and the measurement and evaluation practices of these studies may be examples of the studies to be carried out at the national level. Furthermore, in this study, the rater group had a similar teaching experience and background (e.g., rating experience) in IBDP. In future research, raters with different teaching and/or rating experiences and backgrounds in IBDP should be preferred. Researchers or educators, therefore, compare the relationship between rater experience and the assessment of scientific reports.

Acknowledgments

This paper was produced from part of 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/Social Sciences Institute, 17/07/2019-09-259.

Authorship Contribution Statement

Sinem Arslan Mancar: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing-original draft. **H. Deniz Gulleroglu**: Methodology, Supervision, and Validation.

Orcid

Sinem Arslan Mancar https://orcid.org/0000-0002-2031-2189
H. Deniz Gulleroglu https://orcid.org/0000-0001-6995-8223

REFERENCES

Abedi, J., Baker, E.L., & Herl, H. (1995). *Comparing reliability indices obtained by different approaches for performance assessments* (CSE Report 401). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST). https://cresst.org/wp-content/uploads/TECH401.pdf

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). John Wiley & Sons.

Airasian, P.W. (1994). *Classroom assessment* (2nd ed.). McGraw-Hill.

Aktaş, M. & Alıcı, D. (2017). Kontrol listesi, analitik rubrik ve dereceleme ölçeklerinde puanlayıcı güvenirliğinin genellenebilirlik kuramına göre incelenmesi [Examination of scoring reliability according to generalizability theory in checklist, analytic rubric, and rating scales]. *International Journal of Eurasia Social Sciences*, *8*(29), 991-1010.

Anadol, H.Ö., & Doğan, C.D. (2018). Dereceli puanlama anahtarlarının güvenirliğinin farklı deneyim yıllarına sahip puanlayıcıların kullanıldığı durumlarda incelenmesi [The examination of realiability of scoring rubrics regarding raters with different experience years]. *İlköğretim Onlin*e, 1066-1076. https://doi.org/10.17051/ilkonline.2018.419355

Ananiadou, K., & Claro, M. (2009), 21st century skills and competences for new millennium learners in OECD countries. *OECD Education Working Papers*, *41*. OECD Publishing, Paris,https://doi.org/10.1787/218525261154

Atılgan, H.E., (2005). Genellenebilirlik kuramı ve puanlayıcılar arası güvenirlik için örnek bir uygulama [Generalizability theory and a sample application for inter-rater reliability]. *Educational Sciences and Practice, 4*(7), 95-108. http://ebuline.com/pdfs/7Sayi/7_6.pdf

Atılgan, H., Kan, A., & Doğan, N. (2007). *Eğitimde ölçme ve değerlendirme* [Assessment and evaluation in an education] (2nd ed.). Anı Yayıncılık.

Bailey, D.K. (1994). *Methods of social research* (4th ed.). Free-Press.

Baykul, Y. (2015). *Eğitimde ve psikolojide ölçme: klasik test teorisi ve uygulaması* [Measurement in education and psychology: classical test theory and practice] (3rd ed.). Pegem Yayıncılık.

Bıkmaz Bilgen, Ö., & Doğan, N. (2017). Puanlayıcılar arası güvenirlik belirleme tekniklerinin karşılaştırılması [The comparison of ınterrater reliability estimating techniques]. *Journal of Measurement and Evaluation in Education and Psychology*, *8*(1), 63-78. https://doi.org/10.21031/epod.294847

Brennan, R.L. (2001). *Generalizability theory*. Springer-Verlag.

Burry-Stock, J.A., Shaw, D.G., Laurie, C., & Chissom, B.S. (1996). Rater-agreement indexes for performance assessment. *Educational and Psychological Measurement*, *56*(2), 251-262. https://doi.org/10.1177/0013164496056002006

Büyükkıdık, S. (2012). *Problem çözme becerisinin değerlendirilmesinde puanlayıcılar arası güvenirliğin klasik test kuramı ve genellenebilirlik kuramına göre karşılaştırılması.* [Comparison of interrater reliability based on the classical test theory and generalizability theory in problem solving skills assessment] [Master's Thesis, Hacettepe University]. Hacettepe University Libraries.

Büyüköztürk, Ş., Kılıç Çakmak E., Akgün Ö.E., Karadeniz Ş., & Demirel F. (2012). *Bilimsel araştırma yöntemleri* [Scientific research methods] (11th ed.). Pegem Yayıncılık.

Bybee R.W. (1997). Towards an understanding of scientific literacy. In: W. Gräber & C. Bolte. (Eds.). Scientific literacy. An international symposium (p. 37-68). Institut für die Pädagogikder Naturwissenschaften (IPN): Kiel, Germany.

Callison, D. (2000). Rubrics. *School Library Media Activities Monthly*, *17*(2), 34-6,42.

Cambridge International Examinations (2015). *Cambridge IGCSE global perspectives 0457. Syllabus for examination in 2018, 2019 and 2020*. https://www.cambridgeinternational.org/Images/252230-2018-2020-syllabus.pdf

Cohen J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Collins, R. (2014). Skills for the 21st Century: teaching higher-order thinking. *Curriculum & Leadership Journal, 12*(14). http://www.curriculum.edu.au/leader/teaching_higher_order_thinking,37431.html?issueID=12910

Coombe, C.A., Davidson, P., O'Sullivan, B., & Stoynoff, S. (Eds.). (2012). *The Cambridge guide to second language assessment*. Cambridge University Press.

Deliceoğlu, G. (2009). *Futbol yetilerine ilişkin dereceleme ölçeğinin genellenebilirlik ve klasik test kuramına dayalı güvenirliklerinin karşılaştırılması* [The comparison of the reliabilities of the soccer abilites' rating scale based on the classical test theory and generalizability]. [Doctoral dissertation, Ankara University, Ankara]. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Dietel, R.J., Herman, J.L., & Knuth, R.A. (1991). What does research say about assessment? NCREL, Oak Brook. http://www.ncrel.org/sdrs/areas/stw_esys/4assess.htm

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378- 382. https://doi.org/10.1037/h0031619

Goodrich Andrade, H. (2001) The effects of instructional rubrics on learning to write. *Current issues in Education, 4*. http://cie.asu.edu/ojs/index.php/cieatasu/article/view/1630

Güler, N. (2009). Genellenebilirlik kuramı ve SPSS ile GENOVA programlarıyla hesaplanan G ve K çalışmalarına ilişkin sonuçların karşılaştırılması [Generalizability theory and comparison of the results of G and D studies computed by SPSS and GENOVA packet programs]. *Eğitim ve Bilim*, *34*(154). http://eb.ted.org.tr/index.php/EB/article/view/551/45

Güler, N. (2011). Rasgele veriler üzerinde genellenebilirlik kuramı ve klasik test kuramına göre güvenirliğin karşılaştırılması [The comparison of reliability according to generalizability theory and classical test theory on random data]. *Eğitim ve Bilim. 36*(162), 225-234. http://egitimvebilim.ted.org.tr/index.php/EB/article/view/993

Güler, N., & Taşdelen, G. (2015). Açık uçlu maddelerde farklı yaklaşımlarla elde edilen puanlayıcılar arası güvenirliğin değerlendirilmesi [The evaluation of rater reliability of open-ended items obtained from different approaches] *Journal of Measurement and Evaluation in Education and Psychology*, *6*(1). 12-24. https://doi.org/10.21031/epod.63041

Gwet, K. (2002), Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Series: Statistical Methods for Inter-Rater Reliability Assessment*, *1*(1).1-5. https://www.agreestat.com/papers/kappa_statistic_is_not_satisfactory.pdf

Holbrook, J., & Rannikmae, M. (2009). The meaning of scientific literacy. *International Journal of Environmental & Science Education*, *4*(3), 275-288. https://files.eric.ed.gov/fulltext/EJ884397.pdf

Hurd, P. D. (1998) Scientific literacy: new minds for a changing world. *Science Education*, *82*, 407-416.

International Baccalaureate Organization (IBO). (2014a). *International Baccalaureate Diploma Programme Biology Guide First Assessment 2016*. https://internationalbaccalaureate.force.com/ibportal/IBPortalLogin?lang=en_US

International Baccalaureate Organization (IBO). (2014b). *International Baccalaureate Diploma Programme Chemistry Guide First Assessment 2016*. https://www.ibchem.com/root_pdf/Chemistry_guide_2016.pdf

International Baccalaureate Organization (IBO). (2014c). *International Baccalaureate Diploma Programme Physics Guide First Assessment 2016*. https://ibphysics.org/wp-content/uploads/2016/01/ib-physics-syllabus.pdf

International Baccalaureate Organization (IBO). (2015). *International Baccalaureate Diploma Programme: From principles into practice*. International Baccalaureate Organization.

International Baccalaureate Organization (IBO). (2018). *International Baccalaureate Organization (IBO). (2018). The IB Diploma Programme Statistical Bulletin, May 2018 Examination Session*. https://www.ibo.org/contentassets/bc850970f4e54b87828f83c7976a4db6/dp-statistical-bulletin-may-2018-en.pdf

International Baccalaureate Organization (IBO). (2018). *Assessment principles and practices-Quality assessments in a digital age*. https://www.ibo.org/contentassets/1cdf850e366447e99b5a862aab622883/assessment-principles-and-practices-2018-en.pdf

Kamış, Ö., & Doğan, C. (2017). *Genellenebilirlik kuramında gerçekleştirilen karar çalışmaları ne kadar kararlı?* [How consistent are decision studies in G theory?]. *Journal of Education and Learning, 7*(4). https://dergipark.org.tr/en/download/article-file/336342

Klucevsek, K. (2017). The intersection of information and science literacy. *Communications in Information Literacy, 11*(2), 354-365. https://files.eric.ed.gov/fulltext/EJ1166457.pdf

Krippendorff, K. (2004). Measuring the reliability of qualitative text analysis data. *Quality and Quantity*, *38*(6), 787-800. https://doi.org/10.1007/s11135-004-8107-7

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology.* Sage.

Krippendorff, K. (2011). Computing Krippendorff‟s alpha reliability. http://repository.upenn.edu/asc_papers/43

Kulieke, M., Bakker, J., Collins, C., Fennimore, T., Fine, C., Herman, J., Jones, B.F., Raack, L., & Tinzmann, M.B. (1990). Why should assessment be based on a vision of learning? [online document] NCREL, Oak Brook: IL. Available online: http://www.ncrel.org/sdrs/areas/rpl_esys/assess.htm

Kutlu, Ö., Doğan, D.C., & Karakaya, İ. (2008). *Performansa ve portfolyoya dayalı durum belirleme* [Assessment and evaluation determination based on performance and portfolio] (5th ed.). Pegem Yayıncılık.

Landis, J,R., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1) 159-174. https://doi.org/10.2307/2529310

Lane, S., & Sabers, D. (1989). Use of generalizability theory for estimating the dependability of a scoring system for sample essays, *Applied Measurement in Education*, *2*(3). 195-205. https://doi.org/10.1207/s15324818ame0203_1

Law, B., & Eckes, M. (1995). *Assessment and ESL*. Peguis publishers.

Lord F.M. (1959). Statistical inferences about true scores. *Psychometrika, 24*(1), 1–17. https://doi.org/10.1007/BF02289759 .

Maienschein, J. (1998). Scientific literacy. *Science*, *281*(5379), 917. https://www.proquest.com/openview/568e8a30ee2b1c68d787bbcb39e3f94e/1?pq-origsite=gscholar&cbl=1256

Marzano, R. (2002). A comparison of selected methods of scoring classroom assessments. *Applied Measurement in Education, 15*(3). 249-268. https://doi.org/10.1207/S15324818AME1503_2

Marzano, R.J., & Heflebower, T. (2012). *Teaching & assessing 21st century skills*. Marzano Research Laboratory.

Mertler, C.A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research and Evaluation*, *7*(25), 1-8. https://doi.org/10.7275/gcy8-0w24

Millî Eğitim Bakanlığı (MEB) (2016). *PISA 2015 Ulusal Raporu [PISA 2015: National Report for Turkey*]. Millî Eğitim Bakanlığı, Eğitimi Araştırma ve Geliştirme Dairesi Başkanlığı, Ankara. https://odsgm.meb.gov.tr/test/analizler/docs/PISA/PISA2015_Ulusal_Rapor.pdf

Moskal, B.M. (2000) Scoring rubrics: What, When, How? *Practical Assessment Research and Evaluation*, *7*(3), 1-11. https://doi.org/10.7275/a5vq-7q66

National Research Council. (2012). *Education for life and work: developing transferable knowledge and skills in the 21st century*. The National Academies Press. https://doi.org/10.17226/13398

Nbina, J., & Obomanu, B. (2010). The meaning of scientific literacy: A model of relevance in science education. *Academic Leadership: The Online Journal, 8*(4). https://scholars.fhsu.edu/alj/

Novick M.R. (1966) The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*(1), 1-18. https://doi.org/10.1016/0022-2496(66)90002-2

Nying, E. (2004). *A comparative study of interrater reliability coefficients obtained from different statistical procedures using monte carlo simulation techniques* [Doctoral dissertation, Western Michigan University]. https://scholarworks.wmich.edu/dissertations/1267

Oakleaf, M. (2009). The information literacy instruction assessment cycle: a guide for increasing student learning and improving librarian instructional skills. *Journal of Documentation*, *65*(4). 539-560. https://doi.org/10.1108/00220410910970249

Organisation for Economic Cooperation and Development (OECD). (2017). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*. OECD Publishing. https://doi.org/10.1787/9789264281820-en

Özmen Hızarcıoğlu, B. (2013). *Problem çözme sürecinde dereceli puanlama anahtarı (Rubrik) kullanımında puanlayıcı uyumunun incelenmesi* [Examining scorer's coherence of using rubric in the problem solving process] [Master's dissertation, Abant Izzet Baysal University]. https://tez.yok.gov.tr/UlusalTezMerkezi/tezDetay.jsp?id=-9VIu1xAI6tVn8H1Pmf2Mg&no=XE36zEJKy4iJQQ-bARoPnA

Öztürk, M.E. (2011). *Voleybol becerileri gözlem formu ile elde edilen puanların, genellenebilirlik ve klasik test kuramına göre karşılaştırılması* [The comparison of points of the volleyball abilities observation form (VAOF) according to the generalizability theory and the classical test theory] [Unpublished doctoral dissertation, Hacettepe University]. National Thesis Centre. https://tez.yok.gov.tr/UlusalTezMerkezi/tezDetay.jsp?id=K9erNYiV2Ks_xzov1XrfSQ&no=5OJsxJV1JE2E3hGJDkB8lQ

Partnership for 21st Century Learning. (2007). *Framework for 21st century learning.* https://files.eric.ed.gov/fulltext/ED519462.pdf

Reeves, T.C. (2000). Alternative assessment approaches for online learning environments in higher education. *Educational Computing Research, 3*(1), 101-111.

Rychen, D.S., & Salganik, L.H. (Eds.). (2003). *Key competencies for a successful life and a well functioning society*. Cambridge.

Schleicher, A. (2015), *Schools for 21st-Century Learners: Strong Leaders, Confident Teachers, Innovative Approaches*, International Summit on the Teaching Profession, OECD Publishing. https://doi.org/10.1787/9789264231191-en

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability theory: a primer*. Sage.

Simonson, M., Smaldino, S, Albright, M., & Zvacek, S. (2000). *Assessment for distance education* (ch 11). T*eaching and learning at a distance: foundations of distance education*. Prentice-Hall.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2017). *TIMSS 2019 Assessment frameworks*. http://timssandpirls.bc.edu/timss2019/frameworks/

Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times*. John Wiley & Sons

Turgut, H. (2007). Scientific literacy for all. *Ankara University Journal of Faculty of Educational Sciences (JFES), 40* (2), 233-256. https://doi.org/10.1501/Egifak_0000000176

Turgut, M.F., & Baykul, Y. (2010). *Eğitimde ölçme ve değerlendirme* [*Assessment and evaluation in an education*]. Pegem Yayınları.

Uçak, S., & Erdem, H.H. (2020). Eğitimde yeni bir yön arayışı bağlamında 21. Yüzyıl becerileri ve eğitim felsefesi [On the skills of 21st century and philosophy of education in terms of searching a new aspect in education]**.** *Uşak Üniversitesi Eğitim Araştırmaları Dergisi, 6*(1), 76-93. https://doi.org/10.29065/usakead.690205

Viere, A.J., & Garrett, J.M. (2005). Understanding interobserver agreement: The Kappa statistic. *Family Medicine, 37*(5), 360-362.

Zurkowski, P.G. (1974). *The Information Service Environment Relationships and Priorities. Related Paper No. 5.* National Commission on Libraries and Information Science, Washington, D.C. National Program for Library and Information Services. https://files.eric.ed.gov/fulltext/ED100391.pdf

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 2, 534–547
https://doi.org/10.21449/ijate.1132931
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

How many response categories are sufficient for Likert type scales? An empirical study based on the Item Response Theory

Eren Can Aybek[1,*], **Cetin Toraman**[2]

[1]Pamukkale University, Faculty of Education, Department of Educational Sciences, Denizli, Türkiye
[2]Canakkale Onsekiz Mart University, Faculty of Medicine, Department of Medicine Education, Canakkale, Türkiye

ARTICLE HISTORY
Received: Jan. 04, 2022
Revised: Apr. 10, 2022
Accepted: June 18, 2022

Keywords:
Likert-type scale
Response categories,
Item response theory.

Abstract: The current study investigates the optimum number of response categories for the Likert type of scales under the item response theory (IRT). The data was collected from university students attend to mainly the faculty of medicine and the faculty of education. A form of the "Social Gender Equity Scale" developed by Gozutok et al. (2017) was prepared, which had 3, 5 and 7-point response categories. The graded response model (GRM) was used for item calibrations. The results of the study have revealed that using a 5-point response option provides advantages over using a 3-point response category in terms of reliability and test information perspective in the scale development process. The-5 point scale also provides easier responding process for the respondents while it does not pose a major disadvantage compared to a 7-point response category in the terms of reliability. Therefore, based on the findings of the study, researchers are recommended to use a 5-point response category in their scale development process.

1. INTRODUCTION

Using scales is one of the ways to collect data in educational, behavioral and social sciences. Various ways of developing scales have been reported in the literature. The most widely used ones among these include the Thurstone scaling technique (Anastasi & Urbina, 1997; Dunn-Rankin et al., 2004; Lord, 1954; Nunnally & Bernstein, 1994; Price, 2017; Torgerson, 1958), Guttman scales (Anastasi & Urbina, 1997; Dunn-Rankin et al., 2004; Lord, 1954; Nunnally & Bernstein, 1994; Price, 2017), and Likert-type rating scales (Anastasi & Urbina, 1997; Dunn-Rankin et al., 2004; Price, 2017).

In the Likert-type item construction, there is a statement related to the psychological trait in concern and a rating showing the levels of agreement with, or approval of, this statement (Anastasi & Urbina, 1997; DeVellis, 2003). Likert scales are widely used in instruments measuring thoughts, beliefs, and attitudes (DeVellis, 2003). Likert's (1932) arguments about Likert-type scales are that a) the distances between categories can be kept equal, b) naming of categories can be arranged beforehand even if they are subjective, and c) the judgements of the prepared scale can be changed according to the item analyses to be carried out on the basis of

*CONTACT: Eren Can AYBEK ✉ erencan@aybek.net Pamukkale University, Faculty of Education, Department of Educational Sciences, Denizli, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

the responses of those taking the scale (as cited by Dunn-Rankin et al., 2004). To summarize, Likert's arguments are evaluated based on the distribution of real variables (Price, 2017). In Likert scales, response categories are so arranged that their rating distances are equal as much as possible (DeVellis, 2003). A response category may be structured to have 5 ratings in the form of "strongly disagree", "disagree", "indecisive", "agree", and "strongly agree" (Anastasi & Urbina, 1997) as well as 6 ratings in the form of "strongly disagree", "disagree", "somewhat disagree", "somewhat agree", "agree", and "strongly agree" (DeVellis, 2003). There may also be a neutral point among the ratings of a response category. There are proposals for the ratings to be used at this neutral point such as "neither agree nor disagree" or "agree and disagree equally", but debates on how this neutral point should be expressed still continue (DeVellis, 2003). Likert scales are more popular due to the ease of constructing. They are widely used in the social sciences and educational research (Joshi et al., 2015).

In Likert scales, ordinal categorical scores are generated from responses given by the respondents to the scale items. These scores correspond to a basically two-pole range from strongly disagree to strongly agree (Price, 2017). Some researchers argue that the data obtained from a Likert scale are at an ordinal scale level and statistical techniques suitable for such data should be used (Jamieson, 2004; Stevens, 1946; Thomas, 1982). Although an equal intervals assumption is generally made for Likert scales in practice (i.e., distances between the numbers in an ordinal scale), such an assumption often cannot be evidenced from the perspective of measurement essentials/basics. In the face of this dilemma, the question "Should the data be processed on an ordinal scale or an equal interval scale?" is often asked. Norman (2010) pointed out that Likert scales can be accepted at an equal interval scale level and parametric analyses can be used based on this assumption. In their simulation-based study, Wu & Leung (2017) argued that increasing the number of ratings in the response category of a Likert scale would result in a normal distribution and a similarity with an interval scale.

What ratings and denotations should be used when the response category ratings of a Likert-type scale are prepared? How many ratings should be used to exhibit better psychometric features? These and similar questions were the objects of curiosity and major motivations for conducting this study. Studies with similar objects of curiosity have already taken their places in the literature. Aiken (1983) and Wong et al. (1993) have shown that the number of ratings in a response category has no effect on the alpha coefficient. Champney & Marshall (1939) argued that widely used response categories with 5 or 7 points were not appropriate and suggested that points of response categories should be between 18 and 24. In their study, Chang (1994) tried a 9-item scale as a Likert scale with a response category of 4 and 6 points on 165 participants. The purpose of the trial was to compare the reliability values of the scales with a 4-point or 6-point response category. The results of the study showed that the 6-point scale had a decrease in both reliability and heterothrait-monomethod (THMM) correlations. The 4-point scale also had a higher reliability than the 6-point scale in a multitrait-multimethod (MTMM) covariance matrix analysis. In their study, Preston & Colman (2000) gave 149 participants a scale (with ratings between 2 and 11) to evaluate the service of a restaurant they have visited recently. The best psychometric characteristics were exhibited by the scale with a 7-point response category. The test-retest reliability tended to decline in scales with more than 10-point response category. Dawes (2008) investigated how the use of a Likert scale with 5, 7 and 10-point response categories affected the data obtained with respect to arithmetic means and distribution metrics. Three groups of 300, 250 and 185 participants were administered a scale with 5, 7 and 10-point response categories for 8 questions. Each group were given a scale with a different response category. The 10-point format tended to produce lower arithmetic means than the 5 and 7-point formats (the 5 and 7-point formats were converted to be able compare them with the 10-point format). The skewness and kurtosis values of the scales were very close to each other. In a study by Adelson and McCoach (2010), the same mathematics attitude scale with either a 4-point

response category or a 5-point version including a neutral choice was administered to the 3rd and 6th grade students. The study result showed that the 3rd and 6th grade students had the ability to discriminate the 5-point response option. The participants were also found to like the 4-point response option more than the 5-point response option.

Leung (2011) prepared the Rosenberg Self-Esteem Scale in the form of a Likert scale having 4, 5, 6 and 11-point response categories and administered it to 1217 students. A significant difference was not found in the arithmetic means, standard deviations, item correlations, Cronbach Alpha values and factor loadings of the data obtained from these scales of different rating types. The values obtained from the response category with the largest number of ratings (11-point) were found to reduce skewness and kurtosis and produce data close to normal distribution. In the Kolmogorov-Smirnov and Shapiro-Wilk normal distribution tests applied to the study data, 6 and 11-point scales were found to show a normal distribution. In a study conducted by Wakita et.al. (2012), a scale with the same items was administered to 722 undergraduate students in the form of a Likert scale with 4, 5 and 7-point response categories. The analyses in that study were carried out based on the item response theory. The study result showed that the number of points in the scale influenced the psychological distance between the choices, particularly in the 7-point scale. In a study carried out by Bora (2013), the data obtained from the same Likert scale with 5, 7, 9 and 11-point response categories were compared with respect to arithmetic mean, standard deviation, skewness, and kurtosis. The study was conducted with 413 university students. According to the results, increasing number of choices in the response category resulted in decreasing arithmetic means. When the 5-point response category was used, the skewness value was closest to the normal distribution while the kurtosis value was closest to the normal distribution in the 11-point response category.

In summary, the studies in the literature investigated the number of response categories for the Likert type of scales according to reliability, covariance matrices, descriptive statistics, discrimination of neutral category, its effect on factor loadings, and normal distribution based on CTT. Only a study revealed that psychological distance was affected by number of response categories based on IRT. Current study would contribute to the literature by investigating how response categories work under the IRT.

In the present study, a form of the "Social Gender Equity Scale" developed by Gozutok et al. (2017) was prepared, which had 3, 5 and 7-point response categories. The purpose of the study is to investigate the psychometric characteristics of the data obtained from the scale having 3, 5 and 7-point response categories on the basis of the item response theory (IRT).

2. METHOD

2.1. Participants

The participants are students from 11 different universities. The 3-point, 5-point and 7-point Likert forms of the same scale were administered separately group by group to 512, 514 and 498 students, respectively. The number of students who received all of the forms was 153. The distribution of the participants by gender and faculty and their mean ages have been presented in Table 1.

According to Table 1, it is seen that the forms were mostly answered by female students. The students at a medical school and a faculty of education also outnumbered others in each of the three forms. The median age in all three forms was 20.

Table 1. *Descriptive statistics of the participants with respect to their genders, faculties, and ages.*

	3-Point Likert	5-Point Likert	7-Point Likert
Gender	%	%	%
Female	69.53	73.35	73.04
Male	27.54	24.51	25.15
Unknown	2.93	2.14	1.81
Faculty	%	%	%
Medicine	45.90	40.07	38.83
Education	26.17	34.63	29.18
Other	27.93	25.30	31.99
	Median	Median	Median
Age	20	20	20

2.2. Instrument

The data collection tool used in this study was the "Social Gender Equity Scale (SGES)" developed by Gözütok et al. (2017). The scale was administered to two groups of high school students as it was being developed. The first group included 396 high school students. The data obtained from this group were used for the exploratory factor analysis (EFA) and the calculation of Cronbach Alpha reliability coefficient. The second group included 265 high school students, and the data obtained from this group were used for a confirmatory factor analysis. The exploratory factor analysis showed that the scale consisted of 13 items and 2 subfactors, the first of which was "Male Dominance Mentality (MDM)." This factor had 8 items, none of which was reverse scored. The second factor is "Women's Dependence on Men Mentality (WDMM)" which had 5 items and none of them was reverse scored. The 2-factor SGES explains 53% of the total variance about the characteristic in concern (perceived social gender equity). The level of reliability was .882 for the first subscale, .701 for the second subscale and .889 for the whole SGES. The factor construct obtained was validated by a confirmatory factor analysis.

The SGES was used for university students in a study conducted by Toraman and Ozen (2019). A confirmatory factor analysis (CFA) was carried out to see whether the same factor structure explored based on the data obtained from the high school students will be valid for the university students. The result of the CFA confirmed the factor structure of the scale in the university students as well.

2.3. Data Collection

The 3, 5 and 7-point response category forms of the instrument were sent online to the participants within 2-week intervals. Although the primary goal of the data collection process was to have all the participants who could be contacted answer all of the forms, this goal could not be achieved, and the number of participants who received all forms turned out to be 153. However, assuming that all participants had received the three forms, the data collection process was planned in a way to prevent a sequence effect. Accordingly, participants at different universities received the forms in a different sequence. While some participants first took the 3-point form, then the 5-point and finally the 7-point, some others took the 7-point form first, then the 5-point and 3-point forms. In this way, the forms were administered in 6 different sequences. The forms were administered within 2-week intervals. In order to match data from different forms, a nickname, last 4 digits of their phone numbers and last 4 digits of their student numbers were collected from the participants.

2.4. Data Analysis

The data collected from the participants were analysed on R 4.1.0 (R Core Team, 2021) using the mirt 1.35.1 (Chalmers, 2012) and psych 2.1.6 (Revelle, 2021). The MVN 5.9 (Korkmaz, et al., 2014) package was used to see if the data exhibited a multivariate normal distribution. In the analysis of data, first multivariate normality was tested, then unidimensionality was checked using factor analytic techniques. The local independence was tested using Yen's Q3 statistics, and item-model fit was examined based on the S_χ^2 statistics. Finally, item calibration was performed based on the IRT.

A Henze-Zirkler test was performed to test multivariate normality assumption and the data of the three forms were observed not to meet the multivariate normality assumption ($p < .05$). In the exploratory factor analysis (EFA) to test unidimensionality, the Principal Axis Factoring technique was used as the factor exclusion method. When testing unidimensionality, the analysis was limited with a single factor and the Eigenvalues of the first and second factors were evaluated. The item discrimination indices were evaluated using the item-rest correlation and the internal reliability using McDonald's ω. The statistics obtained for each of the three forms are given in Table 2.

Table 2. *EFA results, summary of item statistics and reliability coefficient.*

	3-Point Likert	5-Point Likert	7-Point Likert
Eigenvalues			
First factor	4.046	5.572	5.393
Second factor	.623	.503	.514
Variance explained	31.1%	42.9%	41.5%
McDonald's ω	.852	.906	.900
r_{jx} minimum	.353	.454	.383
r_{jx} maximum	.586	.710	.713

It was seen that the items in all three forms could be combined under a single factor. With a single-factor analysis, 31.1%, 42.9% and 41.5% of the variance in the items of the three forms could be explained. The internal consistency coefficients of the items were over .80, and the item discrimination indices over .30 in all the forms. As such, all three forms were agreed to satisfy the unidimensionality assumption.

Yen's Q3 statistics was used to find out whether or not the items satisfied the local independence assumption, and it was seen that local independence was satisfied in all three forms. At this point, .37 was used as a benchmark for the Q3 statistics. Then, item-model fit was tested based on the S_χ^2 statistics. The RMSEA values of the S_χ^2 statistics ranged between .000 and .063 in all three forms. Thus, it was concluded that the items provided fit to the one factor model in all three forms.

After completing the prerequisite examinations, item calibrations were performed based on the The Graded Response Model (GRM). After calibrating items, item correlations, option characteristic curves (ORF), item information functions, test information function, and reliability functions were obtained.

3. RESULT

In accordance with the aim of the study, the three forms were calibrated based on the GRM. The item parameters and the RMSEA values of the S_χ^2 statistics showing item-model fit are given in Table 3 and Table 4.

A review of Table 3 and 4 reveals that although there are mathematical differences between the three forms in terms of item discrimination (*a*) parameters, the confidence intervals of the *a* parameters in the three different forms are seen to intersect. Therefore, the number of categories in the scale does not change the *a* parameters of the items. Since the GRM was used as an IRT model, item difficulty (*b*) parameters show the theta level that corresponds to the point where the likelihood of choosing category 1 versus 2 and 3, 1 and 2 versus 3 was equal. In all three forms, the *b* parameters showed increase when moving from the first response category to the last response category.

Table 3. *Item parameters for items 1-7.*

		i1	i2	i3	i4	i5	i6	i7
3-Point Likert	*a*	1.462 (.187)	.892 (.139)	1.399 (.184)	2.151 (.305)	1.773 (.241)	2.060 (.368)	2.329 (.380)
	b_1	.617 (.096)	.909 (.161)	.802 (.109)	1.349 (.119)	1.232 (.121)	2.182 (.223)	1.711 (.153)
	b_2	1.845 (.192)	2,308 (.330)	2.361 (.253)	2.319 (.219)	2.062 (.201)	2.738 (.311)	2.389 (.234)
	$RMSEA_{S\_\chi^2}$	.000	.025	.025	.029	.027	.015	.036
5-Point Likert	*a*	1.666 (.149)	1.255 (.124)	1.711 (.155)	2.723 (.248)	2.402 (.212)	3.840 (.451)	4.129 (.463)
	b_1	-.146 (.081)	-.257 (.097)	-.048 (.078)	.346 (.066)	.233 (.068)	.907 (.069)	.870 (.067)
	b_2	.952 (.095)	.963 (.113)	.788 (.088)	1.388 (.095)	1.251 (.092)	1.956 (.132)	1.535 (.093)
	b_3	1.922 (.157)	1.860 (.178)	1.854 (.151)	2.108 (.153)	1.807 (.129)	2.624 (.259)	1.876 (.120)
	b_4	3.576 (.394)	4.214 (.502)	2.842 (.258)	2.861 (.282)	2,562 (.217)	2.926 (.365)	2.603 (.259)
	$RMSEA_{S\_\chi^2}$	.024	.039	.026	.044	.037	.037	.022
7-Point Likert	*a*	1.818 (.161)	1.126 (.120)	1.787 (.163)	2.976 (.282)	2.419 (.218)	2.936 (.346)	4.192 (.483)
	b_1	.018 (.077)	-.142 (.103)	.071 (.077)	.478 (.066)	.274 (.069)	1.092 (.083)	.935 (.069)
	b_2	.843 (.087)	.848 (.117)	.764 (.086)	1.262 (.089)	1.149 (.089)	2.005 (.147)	1.298 (.082)
	b_3	1.082 (.097)	1.163 (.137)	1.060 (.099)	1.414 (.097)	1.320 (.098)	2.078 (.156)	1.443 (.090)
	b_4	1.484 (.120)	1.640 (.175)	1.404 (.120)	1.711 (.117)	1.679 (.122)	2.304 (.188)	1.607 (.101)
	b_5	2.253 (.185)	2.645 (.277)	2.243 (.187)	2.031 (.147)	2.104 (.160)	2.831 (.309)	2.059 (.145)
	b_6	3.630 (.428)	4.286 (.515)	3.364 (.353)	2.710 (.257)	3.418 (.422)	3.062 (.394)	3.460 (.687)
	$RMSEA_{S\_\chi^2}$	.000	.029	.026	.030	.007	.027	.041

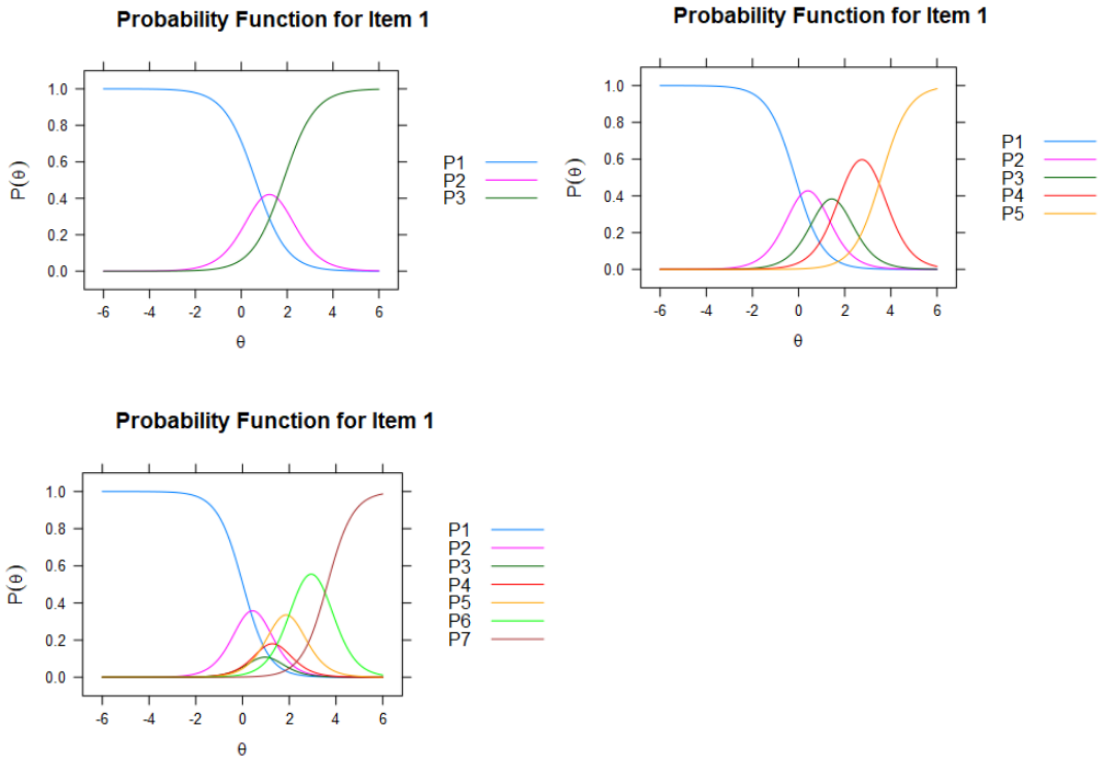
Table 4. *Item parameters for items 8-13.*

		i8	i9	i10	i11	i12	i13
3-Point Likert	*a*	2.112	1.473	3.294	2.352	2.253	1.375
		(.300)	(.215)	(.672)	(.386)	(.364)	(.198)
	b_1	1.335	1.602	2.043	1.862	1.848	1.374
		(.120)	(.174)	(.166)	(.164)	(.165)	(.158)
	b_2	2.374	2.574	2.799	2.728	2.490	2.261
		(.226)	(.292)	(.283)	(.280)	(.250)	(.259)
	$RMSEA_{S\_\chi^2}$	.011	.021	.063	.000	.035	.027
5-Point Likert	*a*	2.484	1.953	3.871	3.832	2.539	2.027
		(.230)	(.187)	(.495)	(.425)	(.274)	(.637)
	b_1	.405	.500	1.136	.870	.970	.637
		(.068)	(.076)	(.076)	(.068)	(.079)	(.077)
	b_2	1.426	1.448	2.036	1.641	1.699	1.509
		(.102)	(.113)	(.145)	(.102)	(.120)	(.116)
	b_3	2.009	2.032	2.670	2.022	2.162	2.066
		(.148)	(.162)	(.276)	(.138)	(.169)	(.164)
	b_4	3.004	3.124	NA	2.944	2.888	3.063
		(.304)	(.318)		(.364)	(.307)	(.310)
	$RMSEA_{S\_\chi^2}$	.014	.017	.040	.043	.040	.013
7-Point Likert	*a*	2.668	2.179	3.514	3.680	2.710	2.140
		(.249)	(.209)	(.450)	(.426)	(.286)	(.214)
	b_1	.489	.558	1.129	.962	.935	.648
		(.068)	(.074)	(.081)	(.072)	(.078)	(.076)
	b_2	1.403	1.288	1.939	1.527	1.672	1.241
		(.098)	(.100)	(.139)	(.099)	(.117)	(.101)
	b_3	1.531	1.398	2.111	1.697	1.743	1.361
		(.106)	(.107)	(.161)	(.112)	(.123)	(.109)
	b_4	1.766	1.637	2.190	1.900	1.952	1.697
		(.123)	(.124)	(.172)	(.133)	(.144)	(.135)
	b_5	2.320	2.129	2.512	2.475	2.513	2.365
		(.183)	(.167)	(.236)	(.226)	(.221)	(.203)
	b_6	3.790	3.015	2.837	2.634	3.774	3.009
		(.612)	(.302)	(.347)	(.264)	(.615)	(.311)
	$RMSEA_{S\_\chi^2}$	.025	.031	.038	.020	.023	.035

The option response functions (ORF) were studied to better understand how the number of categories influenced the response behavior. The ORFs of 3, 5 and 7 response categories for all items are presented in Appendix. The ORFs of the first items of each form are given in Figure 1.

When the ORFs of first items were reviewed, it was seen that each category was differentiated from each other in the forms with 3 and 5 response categories, whereas only 5 of the categories were differentiated in the form with 7 categories. The probability of choosing the third (somewhat disagree) and the fourth (neither agree nor disagree) categories in particular remained lower than the others. This means that when the first item of the scale is presented with 3 and 5 categories, every response category works, whereas when presented with 7 categories, only five categories work. A similar situation can be seen in Appendix.

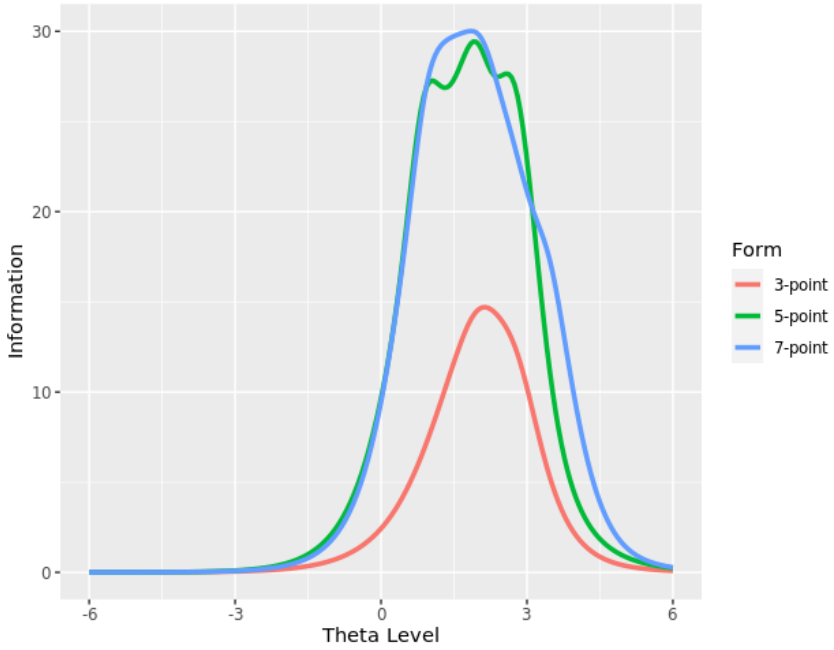
Figure 1. *ORFs of the first items of the forms with 3, 5 and 7 response categories.*



While the middle category (neither agree nor disagree) was not discriminated in items 2, 5, 6, 9, 12 and 13 of the form with 3 response categories, items 2, 5, 6, 7, 9, 11, 12 and 13 in the form with 5 response categories worked as in the 4-category form. The middle category (neither agree nor disagree) was not differentiated from the other categories in the above items except item 6. In item 6, the 'agree' category was not differentiated from the others. As for the 7-response category form, the response categories were not differentiated from each other in all the items. For each item, 4 or 5 categories could be differentiated from each other. This means that the 7-category form was perceived as a 5-category. A similar situation occurred in the test information functions. The test information functions of the three forms are given in Figure 2.

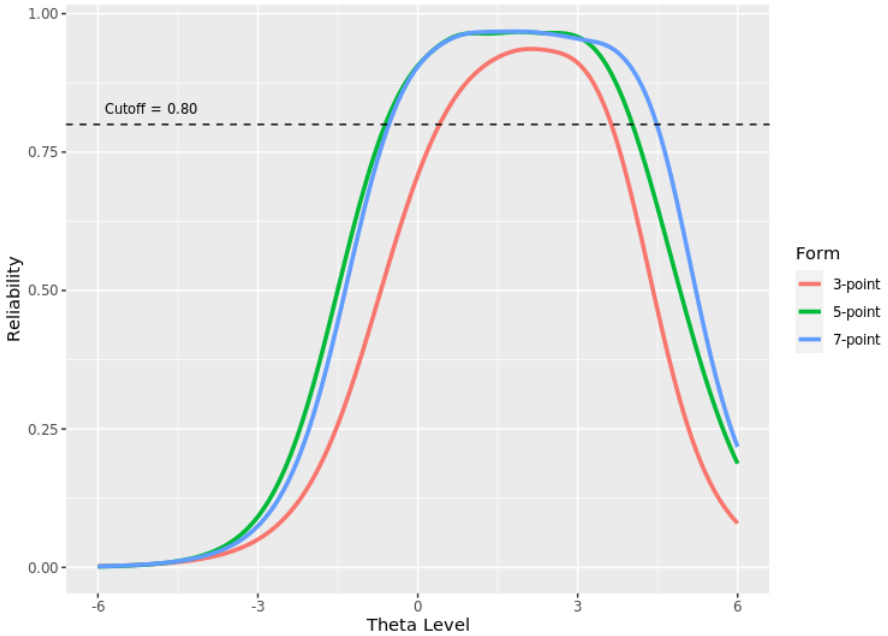
When the test information functions were reviewed, it was understood that the form with 3 response categories provided the least information at a smallest range. The forms with 5 and 7 response categories, on the other hand, provided more information at a much broader range than the form with 3 response categories. The information functions of the forms with 5 and 7 response categories were observed to be quite similar to each other.

Figure 2. *Test information functions of the three forms.*



After reviewing the test information functions, the reliability functions obtained for each of the three forms were also compared and these functions are presented in Figure 3.

Figure 3. *Reliability functions of the three forms.*



Supporting the results provided by the test information functions, a review of the reliability functions revealed that the form with 3 categories was able to make measurements with a higher internal consistency at a smaller theta range [.45-3.59]. The forms with 5 and 7 response categories could make measurements with a high internal consistency at a broader range; -.57 to 4.01 for 5-category and -.51 to 4.43 for 7-category. The reliability functions of the forms with 5 and 7 response categories, however, were similar to each other. Nevertheless, these two forms can make reliable measurements for even individuals with fewer peculiarities as compared to the form with 3 response categories.

4. DISCUSSION and CONCLUSION

In conclusion, although there is no difference between the three forms in terms of the *a* parameters, the forms with 5 and 7 response categories are more advantageous in terms of test information and reliability functions. Additionally, the 7 response category could not be differentiated by the participants as shown by the ORFs. The test information and reliability functions showed that using 7 response categories did not provide a significanrt advantage over using 5 response categories.

The number of ratings that is necessary for a response category in Likert-type scales has been debated in the literature. The studies seem to deal with this matter from different points of view. Chang (1994), Preston & Colman (2000) studied it in relation how reliability changed depending on the use of Likert-type scales with response categories having different ratings. Chang (1994) tested a 9-item Likert scale in its 4 and 6-point versions on 165 participants. They found that the 4-point scale had a higher reliability than the 6-point scale. Preston & Colman (2000) obtained the best psychometric outcomes with a scale having a 7-point response category. Studies investigating test-retest reliability have found that test-retest reliability tends to decline in scales with more than 10 ratings. Leung (2011) administered 4, 5, 6 and 11-point versions of a Likert scale to 1217 students. Their study results revealed that there was not a major difference in their Cronbach Alpha values and factor loadings. In the present study, the "McDonald's ω" reliability coefficients of the 3, 5 and 7-point forms of SGES were calculated and the values of .852, .906 and .900 were obtained respectively (see Table 2). Response categories consisting of more categories achieved higher reliability values. In the studies of Chang (1994), Preston & Colman (2000), however, scales with a response category having fewer ratings produced higher reliability values.

Some studies in the literature have focused on what kind of tendency the use of Likert scales with response categories having various ratings created in statistics such as arithmetic mean, normal distribution, skewness and kurtosis. Dawes (2008) investigated in their study how the use of a Likert scale with 5, 7 and 10-point response categories affected the data with respect to arithmetic means and distribution metrics. They found that the 10-point format tended to produce lower arithmetic means than the 5 and 7-point formats (the 5 and 7-point formats were converted to be able compare them with the 10-point format). They obtained very close values between the scales in terms of skewness and kurtosis. Leung (2011) administered a Likert scale with 4, 5, 6 and 11-point response categories to 1217 students. No major difference was found in the arithmetic means, standard deviations, item correlations, Cronbach Alpha values and factor loadings of the data obtained from these scales of different rating types. The values obtained from the response category with the largest number of ratings (11-point) were found to reduce skewness and kurtosis and produce data close to normal distribution. In the Kolmogorov-Smirnov and Shapiro-Wilk normal distribution tests applied to the study data, 6 and 11-point scales were found to show a normal distribution. In a study conducted by Bora (2013), the data obtained from the 5, 7, 9 and 11-point versions of a Likert scale were compared with respect to arithmetic mean, standard deviation, skewness and kurtosis. According to the result of the study, increasing number of choices in the response category resulted in decreasing arithmetic means. As per the skewness value, the scale closest to normal distribution was the 5-point one. As per the kurtosis value, the scale closest to normal distribution was the 11-category scale. In the present study, a multivariate normal distribution could not be obtained in the data obtained from the administration of the 3, 5 and 7-point forms of SGES. In this respect, the results of the present study are not similar to those in the literature.

Some studies in the literature have focused on how participants understand the choices or ratings when Likert scales prepared with response categories having different ratings were used. In a study of Adelson and McCoach (2010), an attitude scale with either a 4-point response

category or a 5-point version including a neutral choice was administered to the 3rd and 6th grade students. The result showed that the 3rd and 6th grade students had the ability to distinguish between the 5-point response option. The participants were also found to favour the 4-point response option more than the 5-point response option. The results of the present study are similar to those of the study carried out by Adelson and McCoach (2010). According to the results, the form where the choices worked best was the 5-point scale form.

Some studies in the literature have focused on how the use of Likert scales prepared with response categories having different ratings affected the choices or ratings based on the Item Response Theory (IRT). In a study conducted by Wakita et.al. (2012), a scale with the same items was administered to 722 undergraduate students in the form of a Likert scale with 4, 5 and 7-point response categories. The analyses in that study were carried out based on the item response theory. The study result showed that the number of ratings of the scale influenced the psychological distance between the choices, particularly in the 7-point scale. As in the study of Wakita et.al. (2012), the present study also used an IRT-based approach. In conclusion, the forms with 5 and 7 response categories are more advantageous in terms of test information and reliability functions than those with 3 response categories. The 7-response category could not be discriminated by the participants.

Although it was originally intended to administer all the forms to exactly the same participants, the number of participants who received all of the three forms remained limited to 153 due to the difficulties in contacting the participants during the pandemic. This may be considered as a limitation of the study, but since the IRT item parameters are group independence, the study was completed as it is. A study where all participants receive all the forms may be designed in further studies. In scale development studies, using a 5-point response option provides advantages over using a 3-point response category, but does not pose a major disadvantage compared to a 7-point response category. Therefore, researchers are recommended to use a 5-point response category, also considering the ease of responding. It is important to conduct similar studies using different scales under IRT so that the generalizability of results can be tested.

Declaration of Conflicting Interests and Ethics

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

Authorship Contribution Statement

Author 1: Finding the problem, literature review, designing the research, data collection, data analysis, and reporting. **Author 2**: Literature review, designing the research, data collection, data analysis, and reporting.

Orcid

Eren Can Aybek https://orcid.org/0000-0003-3040-2337
Cetin Toraman https://orcid.org/0000-0001-5319-0731

REFERENCES

Adelson, J.L., & McCoach, D.B. (2010). Measuring the mathematical attitudes of elementary students: The effects of a 4-point or 5-point Likert-Type scale. *Educational and Psychological Measurement, 70*(5) 796-807. https://doi.org/10.1177/0013164410366694

Aiken, L.R. (1983). Number of response categories and statistics on a teacher rating scale. *Educational and Psychological Measurement, 43*, 397-401.

Anastasi, A., & Urbina, S. (1997). *Psychological testing*. The USA: Prentice-Hall International, Inc.

Bora, B. (2013). *Pazarlama araştırmalarında kullanılan likert türü ölçeklerin uygulanabilirliğinin incelenmesi* [*A Study on The Applicability of The Likert Type Scales in Marketing*] [Unpublished doctoral dissertation]. Sakarya University.

Chalmers, R.P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Champney, H., & Marshall, H. (1939). Optimal refinement of the rating scale. *Journal of Applied Psychology, 23*, 323-331.

Chang, L. (1994). A psychometric evaluation of 4-point and 6-point Likert-type scales in relation to reliability and validity. *Applied Psychological Measurement, 18*(3), 205-215. https://doi.org/10.1177/014662169401800302

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. International Journal of Market Research, *50*(1), 61-104. https://doi.org/10.1177/147078530805000106

DeVellis, R.F. (2003). Scale development, theory and applications. SAGE Publications.

Dunn-Rankin, P., Knezek, G.A., Wallace, S., & Zhang, S. (2004). *Scaling methods*. Lawrence Erlbaum Associates, Inc.

Gozutok, F.D., Toraman, C. ve Acar Erdol, T. (2017). Toplumsal cinsiyet eşitliği ölçeğinin (TCEÖ) geliştirilmesi [Development of gender equality scale]. *İlköğretim Online Dergisi (Elementary Education Online), 16*(3), 1036-1048. http://dx.doi.org/10.17051/ilkonline.2017.330240

Jamieson, S. (2004). Likert Scales: How to (Ab)use them. *Medical Education, 38*, 1217-1218.

Joshi, A., Kale, S., Chandel, S., & Pal, D.K. (2015). Likert scale: Explored and explained. *British Journal of Applied Science & Technology (BJAST), 7*(4), 396-403. https://doi.org/10.9734/BJAST/2015/14975

Leung, S.O. (2011). A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point Likert Scales. Journal of Social Service Research, 37, 412-421. https://doi.org/10.1080/01488376.2011.580697

Lord, F.M. (1954). Chapter II: Scaling. *Review of Educational Research, 24*(5), 375-392. https://doi.org/10.3102/00346543024005375

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Adv in Health Sci Educ 15*, 625-632. https://doi.org/10.1007/s10459-010-9222-y

Nunnally, J.C., & Bernstein, I. H. (1994). *Psychometric theory*. McGraw-Hill, Inc.

Preston, C.C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica 104*, 1-15. https://doi.org/10.1016/s0001-6918(99)00050-5

Price, L.R. (2017). *Psychometric methods, theory into practice*. New York: The Guilford Press

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Revelle, W. (2021) *psych: Procedures for Personality and Psychological Research*, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych version=2.1.6.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677-680.

Thomas, H. (1982). IQ interval scales, and normal distributions. *Psychological Bulletin, 91*, 198-202.

Toraman, C. & Ozen, F. (2019). An investigation of the effectiveness of the gender equality course with a specific focus on faculties of education. *Educational Policy Analysis and Strategic Research, 14*(2), 6-28. https://doi.org/10.29329/epasr.2019.201.1

Torgerson, W.S. (1958). *Theory and methods of scaling*. New York: John Willey & Sons, Inc.

Wong, C.-S., Chuen, K.-C., & Fung, M.-Y. (1993). Differences between odd and even number of response scales: Some empirical evidence. *Chinese Journal of Psychology, 35*, 75-86.

Wu, H., & Leung, S.O. (2017) Can Likert Scales be treated as interval scales? A simulation study. *Journal of Social Service Research, 43*(4), 527-532. https://doi.org/10.1080/01488376.2017.1329775

APPENDIX

Appendix – 1: Option Response Functions for 3, 5, and 7-point Likert Type Items

Figure 4. *Option response functions for 3-point Likert Type Items*

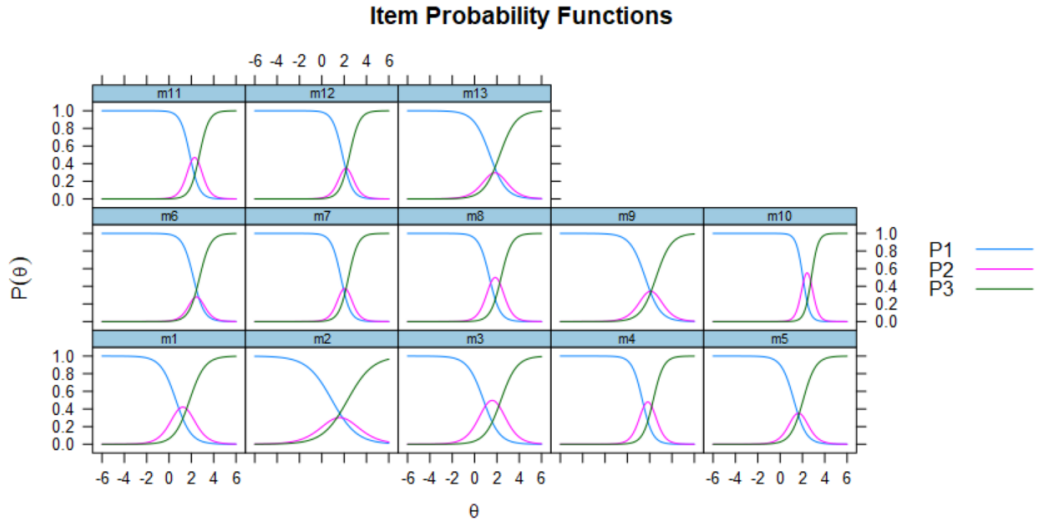


Figure 5. *Option response functions for 5-point Likert Type Items*

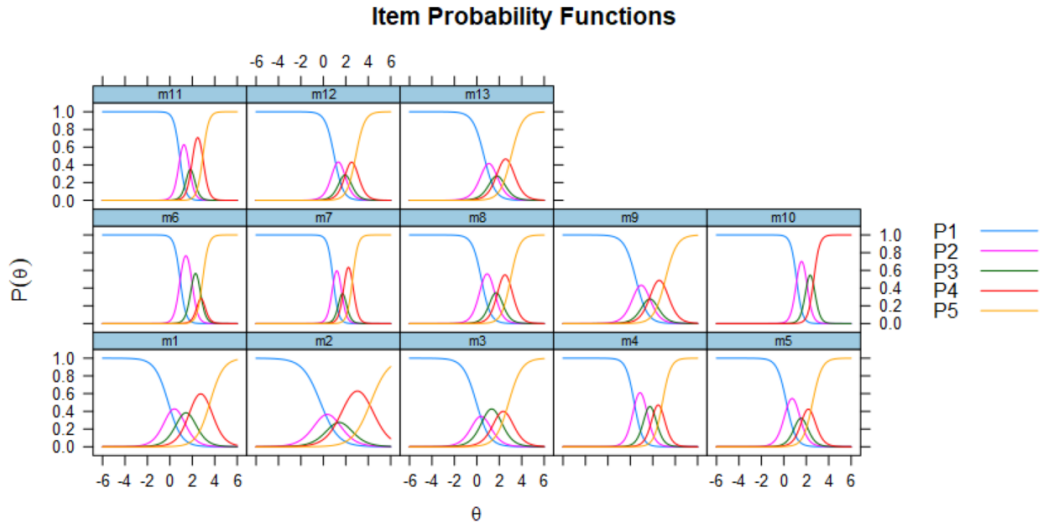


Figure 6. *Option response functions for 7-point Likert Type Items*