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

International Journal of Assessment Tools in Education

Volume: 9 Issue: 4 December 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 4 2022

Editor : Dr. Hakan KOGAR
Address : Akdeniz University, Faculty of Education,
Dumlupinar Bulvari, 07058, Kampus, Antalya,Türkiye
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, Faculty of Education,
Kinikli Yerleskesi, 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 : Dr. 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

Editor
Dr. Hakan KOGAR, *Akdeniz University*, Türkiye

Section Editors
Dr. Safiye BILICAN DEMIR, *Kocaeli University*, Türkiye
Dr. Selma SENEL, *Balikesir University*, Türkiye
Dr. Esin YILMAZ KOGAR, *Nigde Omer Halisdemir University*, Türkiye
Dr. Sumeyra SOYSAL, *Necmettin Erbakan University*, Türkiye

Editorial Board
Dr. Beyza AKSU DUNYA, *Bartın University,* Türkiye
Dr. Stanislav AVSEC, *University of Ljubljana*, Slovenia
Dr. Kelly D. BRADLEY, *University of Kentucky*, United States
Dr. Okan BULUT, *University of Alberta*, Canada
Dr. Javier Fombona CADAVIECO, *University of Oviedo*, Spain
Dr. William W. COBERN, *Western Michigan University*, United States
Dr. R. Nukhet CIKRIKCI, *İstanbul Aydın University*, Türkiye
Dr. Nuri DOGAN, *Hacettepe University*, Türkiye
Dr. Selahattin GELBAL, *Hacettepe University*, Türkiye
Dr. Anne Corinne HUGGINS-MANLEY, *University of Florida*, United States
Dr. Francisco Andres JIMENEZ, *Shadow Health, Inc*., United States
Dr. Nicole KAMINSKI-OZTURK, *The University of Illinois at Chicago*, United States
Dr. Tugba KARADAVUT, *Izmir Democracy Unıversity*, Türkiye
Dr. Orhan KARAMUSTAFAOGLU, *Amasya University*, Türkiye
Dr. Yasemin KAYA, *Atatürk University*, Türkiye
Dr. Hulya KELECIOGLU, *Hacettepe University*, Türkiye
Dr. Omer KUTLU, *Ankara University*, Türkiye
Dr. Seongyong LEE, *BNU-HKBU United International College*, China
Dr. Sunbok LEE, *University of Houston*, United States
Dr. Froilan D. MOBO, *Ama University*, Philippines
Dr. Hamzeh MORADI, *Sun Yat-sen Universit*, China
Dr. Nesrin OZTURK, *Izmir Democracy University*, Türkiye
Dr. Turan PAKER, *Pamukkale University*, Türkiye
Dr. Murat Dogan SAHIN, *Anadolu University*, Türkiye
Dr. Hossein SALARIAN, *University of Tehran*, Iran
Dr. Halil İbrahim SARI, *Kilis 7 Aralik University*, Türkiye
Dr. Ragip TERZI, *Harran University*, Türkiye
Dr. Turgut TURKDOGAN, *Pamukkale University*, Türkiye
Dr. Ozen YILDIRIM, *Pamukkale University*, Türkiye

English Language Editors
Dr. R. Sahin ARSLAN, *Pamukkale University*, Türkiye
Dr. Hatice ALTUN, *Pamukkale University*, Türkiye
Dr. Arzu KANAT MUTLUOGLU, *Ted University*, Türkiye
Ahmet KUTUK, *Akdeniz University*, Türkiye

Editorial Assistant
Dr. Ebru BALTA, *Agri Ibrahim Cecen University*, Türkiye
Dr. Neslihan Tugce OZYETER, *Kocaeli University*, Türkiye
PhDc. Ibrahim Hakki TEZCI, Akdeniz University, Türkiye

Technical Assistant
Dr. Eren Can AYBEK, *Pamukkale University*, Türkiye

CONTENTS

Research Articles

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 4, 808–830
https://doi.org/10.21449/ijate.1059200
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

The comparison of the dimensionality results provided by the automated item selection procedure and DETECT analysis

Ezgi Mor[1,*], **Seval Kula-Kartal**[2]

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

ARTICLE HISTORY
Received: Jan. 17, 2022
Revised: Aug. 23, 2022
Accepted: Dec. 13, 2022

Keywords:
Dimensionality,
Mokken scaling DETECT,
Automated item selection procedure,
Non-parametric item response theory.

Abstract: The dimensionality is one of the most investigated concepts in the psychological assessment, and there are many ways to determine the dimensionality of a measured construct. The Automated Item Selection Procedure (AISP) and the DETECT are non-parametric methods aiming to determine the factorial structure of a data set. In the current study, dimensionality results provided by the two methods were compared based on the original factorial structure defined by the scale developers. For the comparison of the two methods, the data was obtained by implementing a scale measuring academic dishonesty levels of bachelor students. The scale was conducted on junior students studying at a public and a private university. The dataset was analyzed by using the AISP and DETECT analyses. The "mokken" and "sirt" packages on the R program were utilized for the AISP and DETECT analyses, respectively. The similarities and differences between the findings provided by the methods were analyzed depending on the original factor structure of the scale verified by the scale developers.

1. INTRODUCTION

In social sciences, the traits mostly studied are complex, and have an abstract structure that is generally composed of several different components. Researchers frequently employ the exploratory techniques to explore the assessed constructs, and they endeavor to find out the relationships between the constructs and theories. Discovering these associations provides evidence to confirm or invalidate theoretical propositions (Antino et al., 2018). The researchers analyze structures of the latent constructs in detail by employing different dimensionality approaches. Therefore, the investigation of the dimensionality analyses has been an essential part of examining a psychological construct.

The dimensionality has been defined as the minimum number of latent traits which is required to describe the statistical dependency in the data (Zhang & Stout, 1999). If the structure of the data can be explained by only one latent trait, then the dimensionality turns into the unidimensionality. Unidimensionality means that a set of items composing a scale measure only

*CONTACT: Ezgi Mor ezgimor@gmail.com Kastamonu University, Faculty of Education, Department of Educational Sciences, Turkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

one psychological trait (Hattie, 1985). It refers to the existence of only one underlying dimension accounting for the variation in examinee responses. The items of a unidimensional scale purport to measure a single attribute (Sick, 2010). Hence the interpretation of the total score becomes easier and more meaningful. However, unidimensionality may not be valid for each data set. Most of the latent traits targeted by the measurement tools tend to be multidimensional due to the complex nature of psychological constructs (Hemker et al., 1995). Since the targeted traits generally have complex structures, it is very likely to observe multidimensionality in a given dataset.

Multidimensionality in a dataset might be introduced in several ways, because there are many factors affecting respondents' performances on a test apart from the assessed latent trait. These factors might be the personal ones, such as the level of motivation, anxiety, and fatigue etc., or testing factors such as local dependence of the items. However, if the test assesses one dominant dimension, the mentioned factors affect the respondents' performance as minor factors. The dominant dimension reflects the targeted trait with the test, and it determines the success levels of the respondents on the test, hence the test is accepted as unidimensional (Stout, 1999). Considering the complex structure of the dimensionality and unidimensionality issues, it is an undeniable fact that intensive analyses should be employed by the researchers to determine the dimensionality of the traits correctly.

Messick (1975) stated that to assess the meaningfulness of the inferences made from test scores, test developers should confirm what the test score itself actually exhibits. Hence, to make meaningful, appropriate, and useful inferences from the test scores, the construct validity of the scores should be examined meticulously (Kane, 2006; Lissitz, 2009; Sireci, 2009; Zumbo, 2009). Investigation of dimensionality of the measured trait or the structure of the phenomenon is an inevitable part of the construct validity (Slocum-Gori & Zumbo, 2011). Based on discussions on dimensionality, evaluation of dimensionality is a required stage in gathering evidence to support the validity of inferences made from total scores (Yu et al., 2007).

Many methods have been proposed by researchers to investigate the dimensionality of a dataset. For the last 30 years, the two notable reviews of methods and indices of the unidimensionality have been conducted. One of the first studies was conducted by Hattie (1985), in this research the researcher reviewed numerous approaches, and revealed weak sides of these approaches. Tate (2003) expanded the findings of Hattie's (1985) study and included a review of methods and indices applied to discrete variables. In addition, the researcher stated that the most of the available methods perform effectively "within the assumptions". It can be stated that the parametric dimensionality techniques such as the factor analytic methods have strict assumptions to be met to provide accurate results concerning dimensional structure of a dataset. Hence, to assess the dimensionality of the data, there has been an increasing interest in the use of nonparametric techniques and there is increasing number of studies comparing these techniques. To investigate the internal structure of the scales composed of dichotomous items, several researchers have suggested using the Mokken scale analysis (MSA) (Hemker, Sijtsma, and Molenaar 1995; Mokken 1971; van der Eijk and Rose 2015; van Schurr 2003). In addition to these researches proposing MSA, there are several research stuides in which the parametric and nonparametric techniques are compared and the advantages of the drawbacks are analyzed (Finch 2010, 2011; Kuijpers, van der Ark, and Croon 2013; van Abswoude, van der Ark, and Sijtsma 2004; Wismeijer et al. 2008). Wismeijer et al. (2008) compared the results of PCA and MSA with the real data set gathered by Self-Concealment Scale. They proposed the MSA as a complementary tool to PCA to determine the dimensionality of a data set. The scalability coefficients produced by the MSA and the different cutoff values, c values, were cited as the advantages of the MSA over the PCA. They recommended the usage of the MSA in addition to the PCA especially in the decision of the items' retaining or discarding from the scale.

One of the latest researches conducted by Antino et al. (2018) compared MSA with factorial analysis models under conditions of multidimensionality. The researchers compared the nonparametric techniques MSA, item factor analysis (IFA) and Normal Ogive Harmonic Analysis Robust Method (NOHARM). The results of the study proved that MSA should be used as a tool to allocate the items after the unidimensionality is ensured with other methods. The MSA results indicated that items from different but correlated latent dimensions may be grouped as in the same dimension. Eijk and Rose (2015) also stated that the application of MSA is recommended only when the latent structure is refined well.

The popularity of the nonparametric methods is not surprising because they are generally based on less restrictive assumptions than parametric methods. In addition, these methods allow researchers to analyze the dimensionality of datasets obtained from smaller samples (Stout et al., 2002). In line with these advantages of the nonparametric methods, many studies have examined alternative nonparametric methods to analyze the dimensionality of a dataset. Some methods suggested in the related studies have been widely accepted and used by researchers. They are the DIMTEST (Stout, 1987; 1990), the DETECT (Kim, 1994; Zhang & Stout, 1999), and the Hierarchical Cluster Analysis with Proximity Matrix (Roussos et al., 1998). These three methods are all nonparametric statistical analyses. One of the more recently proposed nonparametric methods to analyze the dimensionality of a dataset is the Automated Item Selection Procedure (AISP) of the Nonparametric Item Response Theory (NIRT) approach. The AISP is also known as Mokken Scale Analysis (MSA) (Sijtsma et al., 2011).

The comparative research studies that investigate the performances of different nonparametric dimensionality assessment methods were mostly conducted on simulated data sets. Several studies reported that the performance of the AISP is inferior to the alternative nonparametric techniques in demonstrating the correct dimensionality of the data set (Mroch & Bolt, 2006; van Abswoude et al., 2004). Specifically, it was found that if the components of the latent constructs are correlated, the MSA may produce more erroneous results, and item may load on more than one dimension at the same time (Mroch & Bolt, 2006; van Abswoude et al., 2004). It should be noted that these results were obtained from the simulated datasets, and despite the stated drawbacks, the MSA and AISP methods have been kept using in determination of dimensionality of the assessed traits (Emons et al., 2010; Koster et al., 2009; Meijer et al., 2011; Ordon˜ez et al., 2009; Roorda et al., 2011; Sousa et al., 2010). In Stout et al.'s study (1996), the results obtained from the AISP were compared with the results obtained from the DIMTEST. The researchers found that the AISP has the advantage that it agrees with measurement practice in personality measurement to form facet scales. In addition, it has been still recommended to be used in the dimensionality analyses (Sijtsma et al., 2011). Therefore, the researchers have concluded that there is still a need to investigate the performance of the AISP, especially on empirical data sets. Therefore, the current study aimed to analyze dimensionality results provided by the two nonparametric techniques, the AISP and DETECT for a real dataset. More detailed information for the AISP and DETECT analyses were given in the following.

1.1. The Automated Item Selection Procedure (AISP)

The AISP is a technique that provides a way to investigate the dimensionality assumption in the context of the NIRT approach. This procedure is primarily based on the inter-item covariances, and the strength of the relationship between items and the assessed trait(s). This procedure reveals homogenous subscales of a scale based on the item covariances and item discrimination indexes called as scalability coefficient in the NIRT. While determining item discriminations, it also allows discarding the low-quality items out of the analysis. It results in clustering of test items with reasonable discrimination power that measure the same latent trait, and it composes a unidimensional scale from a large item pool. From this point of view, it can

be used to analyze the dimensionality of scales and investigate the psychometric properties of scales (Meijer & Baneke, 2004; Sijtsma & Molenaar, 2002).

The AISP takes the raw dataset as input and reveals the dimensionality structure of the dataset. While doing this, the AISP uses the scalability coefficients of H (Loevinger, 1948; Molenaar, 1991). Scalability coefficients have crucial importance for MSA that works by pursuing unidimensional scales based on the Loevinger's definition of homogeneity and H coefficients. Scalability coefficients are related with homogeneity which is denoting the unidimensionality of a measure and MSA employs these coefficients to compose unidimensional scales. The H coefficients are defined at three levels: the item (Hi), item pair (Hij), and the whole scale level (H). These coefficients can be expressed as ratios of observed covariance and maximum possible covariance (Meijer, et al., 2015). The first step of MSA is testing the hypothesis about the scalability coefficients. These hypotheses are 1) For each item pairs, item pair scalability coefficients are calculated, and these coefficients show the covariance between two ordered variables. This index expresses the degree to which two items may belong to the same dimension. 2) Like item pair scalability coefficients, item level scalability coefficients are estimated that articulating how much an item is correlated to the sum score based on the remaining items. 3) The last hypothesis is based on the whole scale, as a complete set of the items, there is a test scalability coefficient showing the degree to which the total scores rank the test-taker on the assessed trait accurately. This index reaches a value of 1 when the scale is perfectly unidimensional (van Schuur, 2003). It has conventially been accepted to be higher than 0.30 (Mokken & Lewis, 1982).

Within the AISP, these coefficients are compared with a suitably chosen positive constant lower bound value, which is represented by the c. These coefficients are evaluated according to the lower bound value-constant (c) suggested by Mokken (1971, p.185). The c value is often accepted as 0.3, and items having Hi coefficients higher than 0.3 are included in the scale. For interpretation of all kinds of H values, the guidelines defined by Mokken (1971) are generally accepted. These guidelines are:

$.30 \leq H < .40$: items form a weak scale,

$.40 \leq H < .50$: items form a medium scale,

$.50 \leq H \leq 1.00$: items form a strong scale in terms of discrimination power.

The H coefficient of the scale is estimated from the Hi coefficients of the items. Therefore, power of a scale to discriminate among test-takers is dependent on whether scale items have high scalability coefficients or not. The power of the scale to measure the intended trait and provide an accurate ordering of individuals is determined based on some benchmark values. However, as Meijer et al. (2015) stated, there is no satisfactory level of studies explaining the meaning of these benchmarks. For that reason, the researchers have been advised to select different c values to control the quality of the scale.

There are also some problematic issues about the c values. In practice, higher values of scalability coefficients imply better item discrimination, the researchers may want to higher positive lower bound c. However, it doesn't always mean that high scalability coefficients will compose a discriminating unidimensional scale. In case of multiple latent variables models, the values of the Hij indexes may change according to two types of relationships. If the two items belong to the same latent dimension, the Hij index will show the impact of the factorial loading between each item and the common latent variable. In the second situation, if two items belong to the different latent dimensions, the Hij index will show the factorial loading of each item with its respective dimension, which is calculated as the multiplication of the correlation between two latent dimensions. This may cause a problem especially when the items are highly correlated with each other, and their discrimination indexes are high. They get higher Hij values as a product of correlations between each other, and even if they belong to the different latent

dimension they may be grouped as in the same dimension. Hence based on AISP, the multidimensional scale may be erroneously accepted as a unidimensional scale (Antino et. al, 2018).

The AISP is a "bottom-up" procedure starting from selecting the pair of items of which a) the inter-item covariance, Hij, is higher than 0 significantly, and b) the Hij is the largest among the coefficients for all possible item pairs. Then, the third item is selected from the remaining items based on the levels of Hi coefficients. For the third item, (c) the Hi should be significantly higher than the 0, (d) it should be positively correlated with the first selected item-pair, and (e) the Hi coefficient should be higher than the selected benchmark for the scalability coefficients (c values). Thus, this process continues as long as items meeting specified conditions (c, d, e) are available. At the end of the process, the results might reveal more homogenous item clusters measuring different latent traits or latent trait composites (Meijer & Baneke, 2004). The interpretation of the clusters can be done based on the content of the items composing the same cluster. Lastly, a unidimensional scale is composed which provides a reasonable and reliable ranking of individuals on the latent trait by using their total scale scores (Sijtsma & Molenaar, 2002).

Suppose one wants to reach a scale with high reliability especially for a specified trait range. In that case, it is necessary to select highly discriminative items with item difficulties that span the desired range on the trait continuum. It might be very difficult to measure the whole trait continuum with the same level of precision; therefore, researchers may want to focus on one or more trait level. Sijtsma and Molenaar (2002) showed that items selected in the bottom-up procedure used in the AISP discriminate well across a wide range of item difficulties.

The other advantage of the AISP is that if multidimensionality is suspected in an empirical data set, well-chosen lower bound values will provide critical information about the dimensionality structure of the trait (Hemker et al., 1995). They suggested running the algorithm more than once with different lower bounds, c values, varying between 0.0 and 0.55. For a multidimensional structure, the AISP with varying lower bounds might result with the expected patterns such as: a) the most or all items belonging to one scale, b) items belonging to the two or more unidimensional scales, c) two or more scales including fewer items, and some items that need to be discarded from the procedure. Hemker et al. (1995) stated that the second step should be accepted as a result. As for unidimensional structure, the algorithm provides three sequential steps in case of the varying lower bounds. Firstly, most items are included in one scale; secondly, one smaller scale is detected with the increase in the lower bound. Lastly, one or several scales are determined, and some items are rejected. In this case, the result of the first step should be accepted as final. These findings revealed that the AISP may be used for unidimensional and multidimensional traits considering the different lower bounds for scalability coefficients. In addition, this feature of the AISP may provide a way to scale items that do not fit to any of the parametric IRT models (Reise & Waller, 2003). Hence, it can be concluded that using the AISP makes it possible to compose scales without conceding the content validity.

The AISP provides information for the psychometric qualities of items, therefore using the results of the AISP for building an item bank with already known psychometric properties is more suitable than utilizing the AISP in the context of construction of a scale based on the raw dataset. In addition, researchers are strongly advised to predict the dimensional structure of their item set based on the related theoretical foundation or the content of items. This makes easier to interpret the results of this procedure, and especially when the item set is not unidimensional, the findings can be put into better perspective (Sijtsma & Molenaar, 2002). Based on this suggestion, in the current study, a simulated data set was not generated, instead, a

multidimensional scale whose psychometric qualities were already examined by the scale developers was preferred to evaluate performance of the AISP more efficiently.

1.2. The DETECT Analysis

The other method used to compare the results of the AISP is the DETECT technique. The DETECT provides information regarding the dimensionality of a dataset by enabling evidence for amount of multidimensionality. The main principles of the analysis are to specify the magnitude of dimensionality, test structure and the number of the dominant latent dimensions accounting for the inter-item covariances. It reveals whether an approximate simple structure underlies the item response data. The DETECT provides an index that is defined as the average of all signed conditional covariances calculated for item pairs. Suppose there is only one latent dimension influencing the item responses. In that case, the conditional covariances obtained from some item pairs will be positive while they will be negative for some item pairs. This will result with a low DETECT index since it is calculated based on the average of all signed covariances. However, if more than one dimension is underlying the test data, positive conditional covariances for the items within the same clusters, and negative conditional covariances for the items in distinct clusters will be explored. This will result with a higher DETECT index, which shows that the item response data departs from the unidimensionality and simple structure (Ackerman et al., 2003; Stout et al., 1996).

The DETECT aims at determining cluster partition providing the highest index. To reveal that partition, it calculates the index for different cluster partitions. The DETECT index is designed to be higher when calculated based on a cluster partition that is close to approximate simple structure. It uses different algorithms such as hierarchical cluster analysis to define cluster partition that produces the highest index. The partition giving the highest index determines the maximum value of the DETECT index. When this maximum value is equal or less than 0.10, it shows that one dominant dimension underlies the dataset. A maximum value between 0.10 and 0.50 indicates a weak amount of dimensionality; an index between 0.51 and 1.00 indicates a moderate amount of dimensionality. When the DETECT index is higher than 1.00, it can be accepted that strong amount of dimensionality exists in the data (Roussos & Özbek, 2006; Stout et al., 1996; Tate, 2003)

If the DETECT index reveals that the data differ from the (essential) unidimensionality, then, determining the dimensional structure gains importance. Another index, r, which is also estimated by the DETECT analysis provides information for the structure. This index is computed by dividing the maximum index by the sum of the absolute values of conditional covariances, which are calculated based on the cluster partition that gives the highest DETECT index. An r index between 0.80 and 1.00 indicates that the data is close to approximate simple structure, which means that test items form dimensionally homogenous clusters that are distinct from other clusters. The indexes produced by the DETECT provide answers to the three significant questions regarding the dimensionality of a dataset: Does the item response data hold (essential) the unidimensionality assumption? What is the amount of multidimensionality observed in the data? How many dominant dimensions account for the variation existed in the data? The analysis reveals the amount of multidimensionality exists in the data. Furthermore, if it is concluded that there is more than one dominant dimension accounting for item covariances, the DETECT provides a way to explore the dimensional structure and the number of dominant dimensions (Nandakumar, & Ackerman, 2004; Yu, & Nandakumar, 2001).

When the properties of the AISP and the DETECT methods are examined, it is clear that both techniques aim to discover the dimensionality of a dataset. Both techniques are nonparametric and require fewer assumptions than the parametric methods. The parametric techniques, such as the explanatory (EFA) and confirmatory factor analysis (CFA), are widely known and used by the researchers. However, despite the popularity of these methods, the factor analytic

methods may sometimes perform inadequately, because they may confound the variation caused by item difficulty. As a result, the true number of latent factors is generally overestimated, hence, the findings may cause researchers to make erroneous inferences while interpreting individuals' total test scores (Stout et al., 1996). This situation is valid especially for dichotomous data. When test items are dichotomously scored, the Pearson matrix should be replaced with the tetrachoric matrix. However, the usage of this matrix for the factor analysis may not create common factors unless normality assumptions are met (Lord & Novick, 1968). In addition, if the sample size is less than 200, and item difficulties vary too much, the results of the tetrachoric matrix may not be dependable (Roznowski et al., 1994).

The parametric techniques may not always be suitable for analyzing a dataset's dimensionality of a dataset due to the difficulties in meeting the required assumptions. Furthermore, the parametric methods may result with the erroneous factorial solutions for the data if the researcher insists on using the parametric method although the data fail in meeting the necessary assumptions of the analysis. Therefore, it may be more accurate to utilize both the nonparametric and parametric methods to analyze the dimensionality of a dataset to lessen the possibility of obtaining erroneous results concerning the dimensional structure of the data. If findings obtained from the parametric and nonparametric methods are compatible, this will provide stronger evidence for the dimensional structure of the data. In the current study, the dimensional structure of a psychological trait, which was previously examined based on a parametric dimensionality technique (the exploratory or confirmatory factor analysis) will be determined based on the two nonparametric techniques: the AISP and DETECT procedures.

Theoretically, determining dimensional structure of a psychological trait is one of the most important steps of the test construction and analysis process. However, there is a very limited number of studies empirically investigating dimensionality of a dataset based on the nonparametric methods, especially the AISP procedure (Antino et al. 2018; Hemker et al. 1995; van Abswoude et al. 2004,). Therefore, it is envisaged that the present study will guide researchers to analyze the dimensionality of their data based on the nonparametric approach, which is expected to be great importance to researchers in educational and psychological measurement community and test developers in many fields. Since empirical studies examining the findings of dimensionality provided by nonparametric techniques are very rare, it is expected that findings of the study will contribute to the related empirical knowledge. Accordingly, the current study aims to analyze dimensionality of the dataset obtained from the implementation of the Academic Dishonesty Tendency Scale based on the AISP and DETECT methods. In addition, the CFA was carried out to validate the data gathered by the scale. Hence, the secondary purpose of the study is to compare the results provided by the nonparametric techniques with the results of the parametric one (confirmatory factor analysis) to examine whether the factorial solutions provided by the methods based on different approaches vary significantly. It is expected that revealing the differences and similarities among the dimensionality results provided by these techniques and providing detailed explanation and guidance on how to apply these techniques on the data and interpret the results of them will provide important information to the researchers interested in dimensionality analyzes.

2. METHOD

2.1. Research Model

This is a quantitative research study validating the factorial structure of a scale measuring the academic dishonesty of the undergraduate students based on the three methods, the CFA, AISP and DETECT. Considering the main goal of this study, it is suitable to define the study as a basic study.

2.2. Study Group

To gather the data of the study, undergraduate students of a public and a private university in Türkiye were included in the study group. It was not aimed to generalize the findings of the current study to population; therefore, instead of composing a random sample, convenient sampling was utilized based on the purposive sampling method. The scale aims to assess the academic dishonesty. The researchers thought that only the students who took the methods of scientific research course before may be aware of the concept of academic dishonesty. Therefore, the study group included junior students who had taken and succeeded the methods of scientific research course. The study group consisted of 212 junior students aged 19 to 21. The 44% of the students were male, and the 56% of them were female. The participants were informed about the purpose of the study, and they participated the study voluntarily by signing the consent form.

2.3. Data Collection Tools and Procedure

The Academic Dishonesty Tendency Scale developed by Eminoğlu and Nartgün (2009) was utilized to collect the data. The scale consists of 22 items measuring 4 latent dimensions. The first dimension named as "tendency towards cheating" includes 5 items, the second one, "dishonesty tendency at studies as homework" includes 7 items; the third dimension named as "dishonesty tendency at research and process of write up" has 4 items, and the last dimension, "dishonest tendency towards reference" consists of 6 items. The main reason of selecting this scale was that the issue of academic dishonesty had been investigated in detail in the scientific research courses lectured by the researchers. Another reason of preferring this scale within the context of the study was that the scale developers followed the main principles of the scale-development process neatly and provided the required reliability and validity evidence for the scale.

The scale development process began with literature review and analyzed undergraduate students' views towards academic dishonesty in terms of essays. At first draft of the scale, 40 items were written. The half of the items was negatively worded, while the other half of the items was positively worded. The items were presented to experts to get their ideas regarding the quality of the items, and based on the experts' suggestions, 15 items were discarded from the scale. The trial form of the scale was composed of 25 items. The respondents gave answers to the items on a 5 point-Likert scale (from 1 meaning "completely disagree" to 5 meaning "completely agree"). The trial form was administered to a sample including 300 participants. The psychometric qualities of the items and the whole scale were investigated based on different statistical techniques. The item-total correlations obtained for the items ranged from 0.27 to 0.68. The items were also analyzed based on the scores of the low and high group differences, and these differences were found significant for all items. The scale's construct validity was tested based on the exploratory and confirmatory factor analysis. The EFA was performed with the Principal Component Analysis and the Varimax method. The number of factors was determined based on the variance ratio and Kaiser criterion. The EFA revealed that the scale was composed of four dimensions, and the item loadings were between 0,558 and 0,743, with 53% explained variance ratio. (Eminoğlu & Nartgün, 2009).

The CFA was performed to be able to provide more evidence for the construct validity of the scale by the test developers. In the CFA, the t values of three items were found insignificant and discarded from the scale. The X2/sd ratio was found as 1.85, which provided evidence for a good model-data fit. All fit indexes were estimated as good levels, and the model-data fit was accepted as moderate and good level. Lastly, the reliability of the scale was investigated based on internal consistency. The test-retest and Cronbach Alpha coefficients were estimated for reliability of the scale, and both coefficients were found above 0.70. Based on these findings, the developers stated that the scale could be used to assess the academic dishonesty tendency of university students in a valid and reliable way (Eminoğlu & Nartgün, 2009).

In the current study, the scale was conducted on the study group during the two weeks of the fall semester of the 2018-2019 academic-year. The participation of the study group was voluntary. They were free to withdraw their consent for participation for any reason. In addition, they were informed about the goals of the study before the implementation of the scale.

2.4. Data Analysis

To gather evidence of validity, the CFA was performed to check whether the original factor structure of the scale was preserved in the present study or not. Firstly, the data were examined in terms of the assumptions of the CFA such as normality, multi-collinearity and singularity, linearity, missing and extreme values. The Maximum Likelihood Estimation method was preferred while carrying out the CFA because the normality assumption of the total score was met. Several fit statistics were also estimated to evaluate the model data fit. The Relative Chi Square Test, Root Mean Square Error of Approximation (RMSEA), Root Mean Square Residual (RMR), Normed Fit Index (NFI), Non-Normed Fit Index (NNFI), Relative Fit Index (CFI), Relative Fit Index (RFI), Goodness of Fit Index (GFI), and Adjusted Goodness of Fit Index (AGFI) were considered while examining the model data fit. The related literature proposes various cut-off values for the result of the chi-square test, and the X2/df ratio. For example, Kline (2005) suggests that the values below 3 indicate perfect fit; the ones between 3 and 5 indicate a moderate fit. According to Brown's (2006) suggestions, the values $\leq .08$ are accepted good for the RMSEA, RMR and SRMR. The recommended thresholds indicating moderate values are mostly above 0.90 for the fit indices.

Secondly, the dimensionality of the data was analyzed based on the AISP method. At the first phase of the analysis, the exploratory Mokken scale analysis (Mokken, 1971) was used to examine the scalability and dimensionality of the scale. Furthermore, the scalability coefficients were estimated at this phase. The scalability coefficients were calculated at three levels: the item Hi, item-pair Hij, and scale level, H. Several lower bound values (c) for item level scalability coefficients (c=0.2 and c=0.3) have been proposed by researchers as lower bound values (Loevinger, 1948; Sijtsma & Molenaar, 2002). In the exploratory MSA, Hemker's procedure (Hemker et al., 1994) was adopted, and the AISP was used to select items to form scales. This procedure follows an iterative process. The homogeneous item clusters are composed at each step based on the scalability coefficients of the items, and the steps are repeated until no item satisfying the lower bound determined by the researchers remained. The H values start at 0 in the exploratory analysis and rise to 0.6 in 0.05 increments. In the current study, both the exploratory and confirmatory analyses were performed, and the AISP analyses were carried out on the R program by using the "mokken" package.

In addition to the AISP, the DETECT was also conducted to analyze dimensionality of the data. The exploratory and confirmatory DETECT analyses were carried out on the R program by using the "sirt" package. The confirmatory analysis was conducted based on the original structure explored by the scale developers. The index values (D, ASSI and Ratio) provided by the analyses for different item partitions were evaluated based on the criteria generally accepted for those index values. The D index value over 1 means that strong multidimensionality exists in the data. Index value between 0.40 and 1 indicates existence of medium level multidimensionality. Index value between 0.20 and 0.40 shows that weak dimensionality is observed in the data. Index values lower than 0.20 evidence that the data has an approximate simple structure. The ASSI (Approximate Simple Structure Index) and the Ratio index values could be accepted as the standardized forms of the DETECT index (Zhang, 2007). Similarly, the ASSI value over 0.25 and the ratio value over 0.36 indicate that the dataset shows significant deviation from the simple structure.

3. RESULTS

3.1. The Results Provided by the Confirmatory Factor Analysis

Firstly, the data were reviewed regarding the assumptions of the CFA. The Missing Completely at Random (MCAR) test was used to examine the missing values in the data. The results of the test yielded that the missing values occurred randomly. The 5 cases including missing values were discarded from the data set, and the CFA was performed on the 209 cases, which may be seen small for CFA. However, there are several studies proving that the sample size would be enough for the analysis. Some studies on the necessary sample size for the CFA noted considering the effects of the number of factors, the number of variables per factor and the size of communalities. The common conclusion of the related studies is that there cannot be a rule of thumb that can fit to every situation when deciding the sample size in the CFA. However, Monte Carlo simulation studies provided some guiding results on this issue. Mundrom et al. (2006) revealed that with a variables-to-factors ratio of at least 7, the minimum necessary sample size for excellent agreement is never greater than 180 and, in most cases, less than 150. Similarly, Wolf et al. (2013) revealed that if the number of variables per factor is equal to or higher than 6 necessary sample size does not exceed 200, even for the condition of low communalities. The scale utilized in the current study includes 4 factors having high variables-to-factors ratios. The numbers of the factor included by the 4 factors are 5, 7, 4 and 6, respectively. In addition, most scale items have loadings above 0.55, which indicates high communalities among items belonging to the same factors. Therefore, based on the findings of the related studies, the sample size of 209 can be accepted as enough for conducting CFA on the dataset.

The CFA was conducted to check whether the original four-dimensional structure of the scale was preserved in the current study or not. The results of the CFA revealed that the data obtained in the present study confirmed the original factorial structure of the scale. All fit indices indicated that the proposed model (four-dimensional model) yielded excellent or good model data fit [$\chi^2_{(203)}$ =428.98, p=.34; χ^2/df= 2.09; RMSEA=0.057 (0.049, 0.064; 90% CI); CFI=0.96; RFI=0.92; NFI=0.96; NNFI=0.96; GFI=0.90; AGFI=0.87; SRMR=0.058].

The standardized coefficients of the proposed model ranged from 0.40 to 0.82, above the lower bound value, 0.4 (Crocker & Algina, 1986). When the t-values of the items were analyzed, all of them were found significant, which evidence that all observed variables can be predicted by their latent variables. In addition to the item coefficients, the whole model was found significant in the assessment of academic dishonesty tendency of undergraduate students. Hence, the original factorial structure of the scale was preserved in the current study.

3.2. The Results Provided by the AISP

The exploratory MSA was preferred, and the scalability coefficients were calculated at the item, item pair and scale levels to investigate the suitability of the items to the Mokken scaling. The H_{ij} values were calculated for all item pairs, and it was revealed that all H_{ij} values were positive, and significantly higher than 0, which is the first requirement of the Mokken scaling. In the second step of the analysis, the item level scalability coefficients, H_i, were analyzed, and the Hi values estimated for the items were presented in Table 1. The item level scalability (H_i) coefficients given in Table 1 revealed that only three items (9, 12, and 15) had higher H_i coefficients than the lower bound value, c= 0.3. The low item scalability coefficients indicated that these items do not fit to a unidimensional structure. The scale level scalability coefficient (H) was found as 0.26, which supported that the scale is too weak to be scaled as a unidimensional scale. Upon estimating the scalability coefficients, the significances of these coefficients were analyzed, and all coefficients were found significant. Even though the items

have low scalability values, the significance of the coefficients indicated that the MSA procedure may be applied.

Table 1. *The item level scalability coefficients - H_i value.*

Items	H_i coefficients	Standard error of H_i	Items	H_i coefficients	Standard Error of H_i
1	0.285	0.029	12	0.313	0.029
2	0.290	0.029	13	0.232	0.030
3	0.262	0.027	14	0.298	0.030
4	0.182	0.032	15	0.307	0.031
5	0.325	0.028	16	0.247	0.030
6	0.172	0.035	17	0.214	0.030
7	0.237	0.032	18	0.262	0.031
8	0.218	0.031	19	0.231	0.031
9	0.309	0.029	20	0.296	0.030
10	0.284	0.028	21	0.151	0.031
11	0.228	0.033	22	0.251	0.031
H value =	0.26				

The AISP procedure was started with the lowest value, $c = 0.0$. The AISP results obtained based on the c value of 0.0 revealed that all items were grouped into the same cluster as stated by Hemker et al. (1995), which was an expected finding. However, the c value of 0.0 should be accepted as a starting value, increasing gradually. It is suggested to try different lower bound values while scaling the items based on the AISP (Hemker et al., 1993; Meijer & Baneke, 2004). Depending on this suggestion, in the second step, the cut-off value for the AISP analysis was accepted as 0.2 and the obtained results were given in Table 2.

Table 2. *The results of the AISP.*

Items	Dimension Number	Items	Dimension Number
1	1	12	1
2	1	13	1
3	1	14	1
4	0	15	1
5	1	16	1
6	0	17	1
7	1	18	1
8	1	19	1
9	1	20	1
10	1	21	0
11	1	22	1

Table 2 indicated that the number of the dimensions for most items was defined as 1 by the second AISP analysis. This finding revealed that all items could compose a unidimensional scale, except for the three items (item 4, 6, and 21). These results evidenced that the scale could be accepted as unidimensional if the three items were excluded from the scale. However, the c value of 0.2 may lead to a weak scale because the lower scalability values will result in higher Guttman errors. Molenaar and Sijtsma (2000) proposed that the H values should be higher than 0.3 to get a reliable scale. In addition, the original factorial structure of the scale was multidimensional, and the CFA analysis of the data of the current study also confirmed the original four-dimensional structure. Therefore, the AISP was reiterated several times with higher cut-off values, c=2.25, 2.50, 2.75 and 3.00. The c values of 2.25, 2.50 and 2.75 provided similar results with each other, and the results were presented in Table 3.

Table 3. *The classifications of items according to AISP results.*

Items	Dimension Number			Items	Dimension Number		
	c values				*c* values		
	0.225	0.250	0.275		0.225	0.250	0.275
1	1	1	1	12	1	1	1
2	1	1	1	13	1	1	1
3	1	1	1	14	1	1	1
4	0	0	0	15	1	1	1
5	1	1	1	16	1	1	1
6	0	0	0	17	2	2	2
7	1	1	0	18	1	1	2
8	1	0	0	19	2	2	2
9	1	1	1	20	1	1	1
10	1	1	1	21	2	2	0
11	1	1	1	22	1	2	2

In Table 3, the numbers (0, 1, 2, and 3) indicated the number of possible dimensions of the scale. In addition, the numbers indicated the order of the dimensions, that is, the dimension number 1 meant that the items having this value belonged to the first dimension of the scale. Similarly, items having dimension numbers of 2 and 3 formed the second and the third dimension of the scale, respectively. The number 0, however, meant that these items had very low scalability coefficients, and the scalability criterion was not met for these items. It was found that for the *c* value of 0.225, 17 out of 22 items form a unidimensional scale, while 15 items constituted a unidimensional scale for the *c* value of 0.25. Lastly, 13 items out of 22 items formed a unidimensional scale for the c value of 0.275. The results also revealed that the number of the items that should be omitted from the scale increased as the *c* values got higher. In addition, the number of items included in the second scale increased based on the *c* values. These findings indicated that the scale has a multidimensional structure. The AISP was carried out again with different *c* values (0.300,0.325, 0.350 and 0.375), and the results were given in Table 4.

Table 4. *The classifications of items according to the second AISP results.*

Items	Dimension Number				Items	Dimension Number			
	c values					*c* values			
	0.300	0.325	0.350	0.375		0.300	0.325	0.350	0.375
1	1	1	1	1	12	1	1	1	2
2	1	1	1	1	13	0	0	2	2
3	1	1	1	1	14	1	1	2	2
4	0	0	0	0	15	1	1	2	2
5	1	1	1	1	16	0	2	3	3
6	3	4	0	0	17	2	3	4	4
7	0	0	0	0	18	2	3	4	4
8	0	0	0	0	19	2	3	4	4
9	1	1	3	3	20	1	1	2	2
10	1	2	3	3	21	0	0	0	0
11	3	4	0	0	22	2	3	4	0

Table 4 indicated that the dimensionality results obtained for the *c* values of 0.300, 0.325, 3.50, and 0.375 provided different results than the results obtained from the previous analyses carried out for the *c* values lower than 0.300. For example, the three dimensions were detected even for the *c* value of 0.300. The findings also revealed that the items grouped in the first scale were almost same for all *c* values. The items grouped in the first scale for the *c* values of 0.300 and 0.325 included item 1, 2, 3, 5, 9, 14, 15 and 20. In addition, the items 4, 7, 8, 11, 13 and 21

were detected as unscalable for more than one *c* value. The second, third and fourth dimensions included the items varied for each *c* value. These results confirmed that the scale has a multidimensional structure. However, the cluster partitions of the items are not consistent across the *c* values. Because of these inconsistences, the AISP was reiterated for the c values of 0.4, 0.425, and 0.450. The results suggested a seven-dimensional structure including fewer items, which is not applicable for the scale. Therefore, it was concluded that the results obtained from the analyses carried out for the c values of 0.350 and 0.375 were more similar to the scale's original factorial structure.

When the results obtained from the AISP were compared with the original factor solution achieved by the scale developers, it was seen that the item allocations were so different from the original scale at the all *c*-levels. The results obtained for the *c* value of 0.350 were accepted as the final result by taking into consideration Hemker et al.'s (1995) recommendations. For this c value, the four-factor solution was detected more balanced item distribution of scale's dimensions than the other c values. This item distribution pattern produced the most similar results with the original factor structure of the scale. In this solution, several items (item 4, 6, 7, 9, and 21) were not grouped under any factor. Based on these results, it was decided to discard these items from the scale. To summarize, the stepwise applications of the AISP indicated that the scale has a multidimensional structure, and the factor solution obtained for the c value of 0.350 can be accepted as the result of the AISP. However, it should be noted that this solution is not the same with the original factor solution proposed by the scale developers. It is the most similar one with the four-factor solution, but it proposed to discard 5 items from the scale, which resulting in the biggest difference from the original factor scale. When the items' distribution was analyzed, it was detected six items (I12, I15, I20, I9, I10 and I22) were allocated to the different factors from the original solutions. The other 10 items were estimated at the right factors as proposed by the original scale. This is the best solution created by the AISP; hence these results were accepted as the final solution for this technique.

3.3. The Results Provided by the DETECT Analysis

The exploratory DETECT analyses were carried out to analyze whether the dataset has simple structure or not. The index values estimated by the exploratory analysis for different item partitions were given in Table 5.

Table 5. *The results obtained from the exploratory DETECT analysis.*

The Number of Clusters	The D index	The ASSI	The Ratio
2	2.589	0.030	0.242
3	7.664	0.506	0.717
4	8.076	0.524	0.756
5	8.729	0.610	0.817
6	8.406	0.593	0.787
7	8.392	0.593	0.785

Table 5 indicated that the highest *D* index was estimated for the five-dimensional structure. The *D* index gives information regarding the structure of the data and the amount of multi-dimensionality observed in the data. A low index value means that inter-item covariances estimated conditioned on total scores are not high. This finding indicates that one dominant dimension explains inter-item relations and the dataset has a simple structure. A high index value shows that the dataset has a multidimensional structure. The *D* index value over 1 means strong multidimensionality exists in the data. According to the values given in Table 5, all of the *D* index values estimated for different item partitions were over 1. When the dataset was not unidimensional, obtaining a high *D* index value was expected since high conditional covariances among items belonging to the same item cluster. Therefore, the high DETECT,

ASSI and Ratio index values given in the Table 5 revealed that conditional covariances among items were high. There were more than one dominant dimension explaining inter-item covariances, and the dataset showed significant differences from the unidimensional structure.

If the *D* index value evidences that the dataset has a multidimensional structure, it is necessary to specify the number of dimensions explaining the variance observed in the data and to explore how the items spread into different item clusters. The DETECT analysis estimated the highest *D* index for the five-dimensional structure. However, the original scale had a four-dimensional structure, and also the CFA results of the current study confirmed the original structure. Therefore, the confirmatory DETECT analysis was carried out based on the original structure defined by the scale developers. The index values provided by the analyses were given in Table 6.

Table 6. *The index values estimated by the exploratory and confirmatory DETECT.*

DETECT Analyses	Number of item cluster	D Index	ASSI	Ratio
Exploratory	5	8.729	0.610	0.817
Confirmatory	4	8.466	0.593	0.792

According to indices given in Table 6, the exploratory DETECT analysis indicated that the dataset obtained from applying the scale on the study group had five-dimensional structure. As stated before, the highest index values were estimated for the five-dimensional structure. The values calculated for the five-dimensional structure by the exploratory DETECT analysis were used as criterion to compare the results provided by the confirmatory DETECT analysis. The *D*, ASSI and Ratio index values estimated for the four-dimensional structure by the confirmatory analysis were high. The high values produced by the confirmatory analysis supported the results provided by the exploratory analysis. The results of both analyses indicated that the dataset has a multidimensional structure. When the index values were analyzed, it could be seen that the values obtained for the four-dimensional structure were very close to the values calculated for the five-dimensional structure. The cluster solution provided by the exploratory DETECT for the four-dimensional structure was given in Figure 1.

Figure 1. *The cluster solution provided by the exploratory DETECT.*

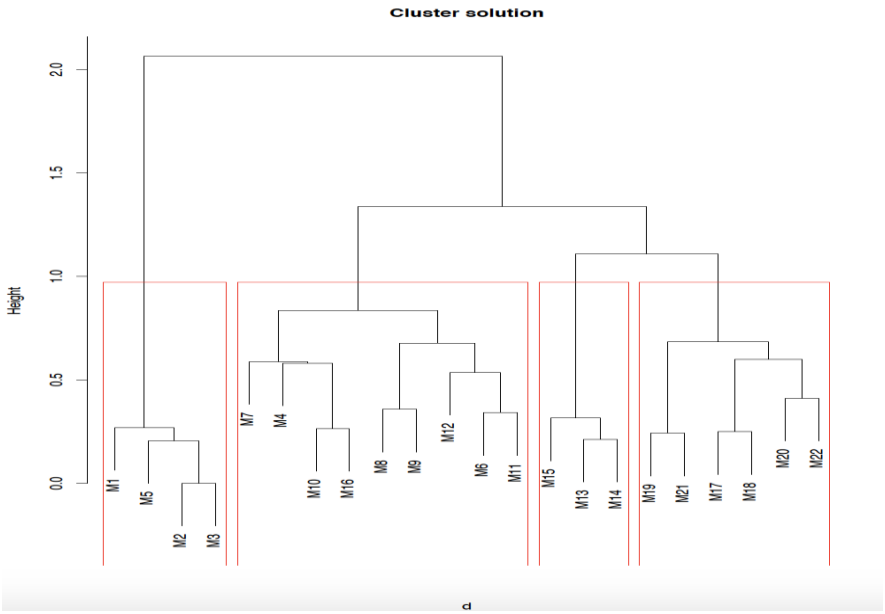


Figure 1 represents the cluster solution of the DETECT analysis. According to Figure 1, the first dimension included four items (1, 2, 3 and 5). The second dimension consisted of 9 items

(4, 6, 7, 8, 9, 10, 11, 12, 16), the third dimension included three items (13, 14, 15), and the fourth dimension consisted of six items (17, 18, 19, 20, 21, 22). To summarize, the results of both exploratory and confirmatory DETECT analyses indicated that the dataset is multidimensional, and items comprise homogenous item clusters. To make the results clearer and understand the proposed dimensionality structure for the scale, the obtained results from both techniques were given in Table 7.

Table 7. *The dimensionality structures proposed by the AISP and DETECT and the original scale.*

	Dimension1	Dimension2	Dimension3	Dimension4
AISP	1, 2, 3, 5, 12	9, 10, 16	13, 14, 15, 20	17, 18, 19, 22
DETECT	1, 2, 3, 5	4, 6, 7, 8, 9, 10, 11, 12, 16	13, 14, 15	17, 18, 19, 20, 21, 22
Original Scale	1, 2, 3, 4, 5	6, 7, 8, 9, 10, 11, 12	13, 14, 15, 16	17, 18, 19, 10, 21, 22

In Table 7, it is possible to see the items' allocation to the dimensions according to both techniques. Compared the results of the techniques with the original scale structure, it is clear that DETECT produced nearly the same factorial structure with the original scale. Only two items were placed to a different dimension, the other 20 items were However, as for AISP, the results were found so different from the original scale. Firstly, six of the 22 items were discarded from the scale based on the results of the AISP. The other dimensions suggested by the AISP were found similar with the other techniques, but the second dimension were found so different. Based on these results, it can be deduced that DETECT produced more suitable results with the original scale structures than the AISP.

4. DISCUSSION and CONCLUSION

The AISP analyses proposed several different factorial solutions. Firstly, the scalability coefficients were analyzed, and all coefficients were found low, but significant. In the MSA, the scalability coefficients have critical importance, and have been described as a method for evaluating a variety of measurement properties such as unidimensionality and local independence (Lind, 2017; Meijer et al., 2015). Despite of the recommended interpretations of scalability coefficients, there have been ongoing discussions regarding the usage of scalability coefficients in dimensionality analyses (Smits et al., 2012). That is why it may not be suitable to decide on the dimensionality of the scale based on the weak scalability coefficients. Therefore, the dimensionality of the data was examined by the AISP method. The related studies on scalability coefficients criticize that the accepted benchmarks (0.3) for the coefficients are so high that it is difficult to obtain these values for many scales, and items (Mokken & Lewis, 1982). For that reason, as stated by Hemker et al. (2015), the AISP analyses were reiterated for various *c* values to have more reliable evidence regarding the factorial structure of the scale.

Various *c* values ranging from 0.2 to 0.450 were utilized to reach the original factor solution of the data. When the c value of 0.2 was accepted as a cut-off value, it was found that the scale could be accepted as unidimensional except for 3 items. The *H* value was estimated as 0.26, which indicates a high Guttman error. Therefore, this solution was not acceptable for the scale. The AISP analyses were reiterated for the *c* values of 0.3, 0.325, 0.350, and 0.375. In addition, the results obtained from the AISP analyses carried out for the *c* values of 0.4, 0.425, and 0.450 were examined. However, it was concluded that the results of these analyses are too ambiguous to interpret. Furthermore, the results of these analyses suggested to discard several items from the scale, which might affect the content validity of the scale negatively. Although the results provided by the analyses are somehow inconsistent, it is still easy to infer from the results that the scale has a multidimensional structure.

The complex factor solutions in which items are mixed across the factors are generated by the AISP, when factors of scales are correlated with each other (Meijer & Baneke, 2004). In the

current study, the AISP proposed several different and complex factorial structures for the scale with some unscalable items. In addition, the results of the AISP varied across the different c values. Because of the inconsistency among the results, it was concluded that the AISP may not be able to provide correct factor solutions in case that the scale has a multidimensional structure, and the correlations among these dimensions are medium or high levels (in this study, the inter-factors correlations ranged from 0.42 to 0.58).

In addition to the AISP, the dimensionality of the data was also examined based on the DETECT analysis. Similar with the AISP and the CFA results, the exploratory DETECT analysis supported the multidimensional structure of the scale. However, the highest index value was obtained for the five-dimensional structure by the exploratory DETECT analysis, while the CFA and AISP provided four and two-dimensional solutions, respectively. The exploratory DETECT analysis provided similar findings in terms of detecting the existence of the multidimensionality with the two methods, but the methods resulted with different factorial solutions. However, the exploratory DETECT analysis provided very similar cluster solution with the CFA. Only two items (4 and 16) were defined in different clusters by the two methods. While the DETECT revealed that item 4 belonged to the second dimension, the same item belonged to the first dimension in both the original-factorial structure and the structure defined in the current study. Similarly, the DETECT defined that item 16 belonged to the second dimension, while this item belonged to the third dimension in both the original and current study. The exploratory DETECT analysis provided results supporting the validity of four-dimensional structure explored by the CFA.

Similar with the AISP, both the exploratory and confirmatory DETECT analyses supported the existence of multidimensionality in the data. However, it is not possible to state that the AISP and DETECT analysis provided similar results regarding the factor numbers. The AISP defined four dimensions, while the DETECT analyses defined five factors underlying the scale items. In addition, the two methods provided very different item cluster solutions. The results of the analyses revealed that the DETECT provided more similar results with the CFA. The findings provided by the AISP were not in line with the factor solution proposed by the scale developers.

The results of the AISP analyses indicated that the scale is not suitable to be scaled based on the NIRT approach, which requires unidimensionality. It can be scaled based on the NIRT only if several items are excluded from the scale, but this situation may create new validity problems. Therefore, it is possible to state that the results obtained from the AISP did not support the original results of the scale. However, the AISP enabled to reveal multidimensionality observed in the data. The inconsistency between the factorial solution provided by the AISP and the original factorial structure might be caused by high correlations among dimensions of the scale. In the study conducted by Antion et al. (2018), the AISP correctly identified the dimensionality of the data, but in that study, the latent dimensions were uncorrelated. van Schuur (2003) mentioned the same drawback of the AISP. The researcher stated that in multidimensional scenario, only if the latent dimensions are uncorrelated, the AISP provides the accurate dimensionality. In addition, the results of the related studies (Antino et al., 2018; van Abswoude et al., 2004) confirmed van Schuur's (2003) claims. The findings of these studies revealed that correlations among latent dimensions result with relatively high Hij values for the items belonging to different dimensions, and the AISP erroneously tend to group all items in a single scale. The Hij values estimated in the current study ranged from 0.45 to 0.75, which indicates that there are medium and high correlations among the dimensions. As stated by Antino et al. (2018), the erroneous grouping effect often tends to occur wherever intermediate or high loading items are found together with moderately correlated latent dimensions. In addition, these situations may occur commonly in practice, therefore, the Mokken scale analysis may not be an adequate technique to explore the dimensionality of scales whose latent structure tend to

be multidimensional. The results obtained from the AISP were consistent with the inferences of the study conducted by Antino et at. (2018). The scale utilized in the current study has a multidimensional structure, therefore the AISP could not be able to provide consistent results regarding the factorial structure of the scale. On the condition that the c value was accepted lower than the required level, the findings were found similar the findings reported van Abswoude et al. (2004) and Antino et al. (2018). They observed a tendency to lump items together in a single scale as in the findings of this study. Accordingly, it was concluded that it is necessary to utilize other dimensionality techniques together with the AISP when there is any suspicion regarding the existence of multidimensionality in the data.

Upon considering the related literature, it has been deduced that there is very limited number of studies investigating the usage of the AISP in the determination of the dimensionality. Wind (2017) stated that even though the AISP has been applied as a technique for evaluating the dimensionality and selecting items in affective domains, the usage of this procedure has not been fully explored especially in educational testing. The first study was conducted by Cavalini (1992), and the researcher compared the findings of factor analysis with the AISP. He used different lower bounds of scalability coefficients, and the results suggested that either three or four scales may be accepted. In the explanatory factor analysis, the four-factor solution was accepted as the best one. Hence, it may be accepted that the decisions about the number of dimensions should be made by considering reliability of the per scale score, the number of items in the per scale, and the interpretation of the meaning of the scales. Comparing the results of the EFA and AISP, the researcher deduced that the AISP can be used instead of the EFA in scale development process.

Another related study (Hemker et al., 1993) showed that results of the AISP may be affected by several factors such as the number of factors and correlations among factors. The number of items in separate factors may lead different solutions of the AISP. Considering these results, they proposed applying the AISP in the beginning of the scale development process. In addition, the researchers suggested that new studies should be done to compare the results obtained from empirical data sets and simulated data set together. To summarize, the results provided by the AISP in the current study, and the findings of the related research revealed that it is necessary to investigate the AISP method more to be able decide whether it is an effective dimensionality method or not.

The findings of the AISP did not provide the same factor solution proposed by the scale developers. However, both non-parametric methods (the DETECT and the AISP) revealed that the scale is multidimensional. Therefore, it is not appropriate to analyze the dataset based on the unidimensional IRT models. The results of the study indicated that both the DETECT and AISP succeeded to reveal the multidimensional structure of the scale. However, to determine the correct number of dimensions may not be the only goal in scale construction process. In this process, scale developers may want to create multidimensional scale of which factors are highly correlated. The AISP can provide strong evidence for the construct validity if the researchers select high cut-off values for the scalability coefficients.

The current study makes contributions from the methodological standpoint. In the first place, to the best of our knowledge, the present study is the first to compare the AISP and the DETECT with the CFA. On the other hand, our results obtained from the AISP are congruent with the findings reported by the related studies (Abswoude et al., 2004; Antino et al., 2018; Hemker et al., 1995). The researchers showed that the AISP may present misleading results when items and dimensions of scales have intermediate and high correlations among each other. In addition, we build on the existing work by showing the superiority of the parametric factorial techniques like the CFA compared to the non-parametric ones, such as the AISP and DETECT in the detection of the number of factors. Beyond the contribution made by the current study to the

methodological literature, there are some practical implications of our findings for the researchers interested in social sciences. Our results revealed that the application of certain techniques under inadequate conditions may lead to erroneous results. Using only non-parametric techniques to examine dimensionality may cause researchers to make inaccurate decisions on the latent structures of the scale. To update the recommendations made by the related studies (Antino et al., 2018; van Abswoude et al., 2004; van der Eijk & Rose, 2015; van Schuur, 2003), social scientists are recommended to prefer the AISP only when the factorial structure is defined as unidimensional, or to develop a unidimensional scale. Another suggestion to the researchers regarding the AISP is to try out different lower bounds based on the item scalability coefficients. In a study by Meijer and Baneke (2004), conducting the AISP with a wide range of c values, it was found that if the item scalability coefficients are too low than the 0.3, high c values like 0.4 and higher may not produce meaningful results. For higher c values, the AISP generated so inconsistent results that the factorial solutions are almost impossible to interpret. In addition, researchers are advised to use the AISP method in dimensionality analysis only if the item scalability coefficients are higher than the lower bound values. As stated before, the AISP uses scalability coefficients based on the inter-item covariances, and if these coefficients are low, the AISP may generate inconsistent and unreliable results. Lastly, the usage of the DETECT analysis in combination with a parametric technique will provide more powerful and reliable results in examination of the dimensionality

Despite the theoretical and practical contributions of the current study, it is affected by several limitations that are discussed here together with the related future research. Firstly, the initial and whole item pool of the scale was not used in the dimensionality analyses process, since the scale was already developed, and the final version of it was available to use. This situation may have affected the results of the current study. Therefore, in the future studies, the researchers are recommended to use the DETECT and AISP techniques to analyze dimensionality by using the whole item pool, which may lead to different and more accurate results in terms of the factorial structure of the scale.

The second limitation of the study is that the correlations among dimensions were not manipulated, hence it might have altered the results as it was stated by the researchers (Antino et al., 2018; Hemker et al., 1993). In the future studies, correlations among dimensions may be controlled, and the effects of the correlations among factors on the dimensionality results can be observed. Thirdly, the item characteristics, such as item difficulty and discrimination indexes were not considered because the scale was already developed. Especially, item covariances may result with different factorial solutions in the AISP method, hence in the future studies, item covariances should be considered. Fourth, the data considered in the study was polytomous based on Likert response formats. However, dichotomous items are also used very frequently in educational settings. Therefore, researchers may examine dimensionality of the data obtained from dichotomous items. Lastly, the study group of the current study was relatively small, which may have affected the variances of the total scale scores, therefore, in the future research, the sample size can be modified to examine the factorial structure more neatly. For these reasons, in the future studies, these limitations should be addressed, and the researchers might apply several methods while deciding the number of factors. In that case, the results provided by the techniques may be more comparable, and both item characteristics and contents may be considered together in the process of the deciding the number of factors and items included in factors.

Declaration of Conflicting Interests and Ethics

The data of this study were gathered before February 2020. Hence, there is no ethical committee approval of the study. However, all ethical issues were taken into consideration during the data collection process by the authors. In addition, within the context of this study, the data were not

used to make any decisions about the participants, only the theoretical comparisons were administered based on the results.

Authorship Contribution Statement

Ezgi MOR: Investigation, Resources, Analysis based on the automated item selection procedure, and Writing-original draft. **Seval KULA KARTAL**: Investigation, Analysis based on the DETECT, and Writing-original draft.

Orcid

Ezgi MOR https://orcid.org/0000-0003-0250-327X
Seval KULA KARTAL https://orcid.org/0000-0002-3018-6972

REFERENCES

Ackerman, T.A., Gierl, M.A., & Walker, C.M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*(1), 37-53. https://doi.org/10.1111/j.1745-3992.2003.tb00136.x

Antino, M., Alvarado, J.M., Asún, R.A., & Bliese, P. (2020). Rethinking the exploration of dichotomous data: Mokken scale analysis versus factorial analysis. *Sociological Methods Research, 49*(4), 839-867. https://doi.org/10.1177/0049124118769090

Cavalini, P.M. (1992). *It's an ill wind that brings no good. Studies on odour annoyance and the dispersion of odorant concentrations from industries* [Unpublished doctoral dissertation]. University of Groningen, The Netherlands.

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

Finch, H. (2010). Item parameter estimation for the MIRT model bias and precision of confirmatory factor analysis based models. *Applied Psychological Measurement 34*(1), 10-26. https://doi.org/10.1177/0146621609336112

Finch, H. (2011). Multidimensional item response theory parameter estimation with non-simple structure items. *Applied Psychological Measurement, 35*(1), 67-82. https://doi.org/10.1177/0146621610367787

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*(1), 255-282.

Guttman, L. (1950). *The basis for scalogram analysis*. In S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, & J.A. Clausen (Eds.), Measurement and prediction (pp. 60-90). Princeton University Press.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 139-164. https://doi.org/10.1177/014662168500900204

Hattie, J., Krakowski, K., Jane Rogers, H., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement, 20*(1), 1-14. https://doi.org/10.1177/014662169602000101

Hemker, B.T., & Sijtsma, K. (1993). A practical comparison between the weighted and the unweighted scalability coefficient of the Mokken model. *Kwantitatieve Methoden, 14*(44), 59-73.

Hemker, B.T., Sijtsma, K., & Molenaar, I.W. (1995). Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. *Applied Psychological Measurement, 19*(4), 337-352. https://doi.org/10.1177/014662169501900404

Junker, B.W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics, 21*(1), 1359-1378. https://doi.org/10.1214/aos/1176349262

Kane, M.T. (2006). Validation. *Educational Measurement, 4*(2), 17-64. https://doi.org/10.1111/j.1745-3992.1985.tb00874.x

Kim, H.R. (1994). *New techniques for the dimensionality assessment of standardized test data* [Doctoral dissertation, University of Illinois at Urbana Champaign]. ProQuest Dissertations and Theses Global.

Kuijpers, R.E., van der Ark, L.A. & Croon, M.A. (2013). Standard errors and confidence intervals for scalability coefficients in Mokken scale analysis using marginal models. *Sociological Methodology, 43*(1), 42-69. https://doi.org/10.1177/0081175013481958

Lissitz, R.W. (Ed.). (2009). *The concept of validity: Revisions, new directions and applications*. Information Age Publishing.

Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects ofscale analysis and factor analysis. *Psychological Bulletin, 45*(1), 507-530. https://doi.org/10.1037/h0055827

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Addison Wesley.

Meijer, R.R., Sijtsma, K., & Smid, N. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement, 14*(1), 283-298. https://doi.org/10.1177/014662169001400306

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*(10), 955 966. https://doi.org/10.1037/0003 066X.30.10.955

Mokken, R.J. (1971). *A theory and procedure of scale analysis*. De Gruyter.

Mokken, R.J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6*(1), 417-430. https://doi.org/10.1177/014662168200600404

Mokken, R.J., Lewis, C., & Sijtsma, K. (1986). Rejoinder to The Mokken scale: A critical discussion. *Applied Psychological Measurement, 10*(1), 279-285. https://doi.org/10.1177/014662168601000306

Molenaar, I.W. (1982). Mokken scaling revisited. *Kwantitatieve Methoden, 3*(8), 145-164.

Molenaar, I.W. (1991). A weighted Loevinger H-coefficient extending Mokken scaling to multicategory items. *Kwantitatieve Methoden, 12*(37), 97-117.

Molenaar, I.W. (in press). *Nonparametric models for polytomous responses*. In W.J. van der Linden, & R.K. Hambleton (Eds.), Handbook of modern psychometrics (pp. 361-373). Springer.

Molenaar, I.W., Debets, P., Sijtsma, K., & Hemker, B. T. (1994). *User's manual for the computer program MSP* (Ver. 3.0). ProGAMMA.

Mroch, A.A., & Bolt, D.M. (2006). A simulation comparison of parametric and nonparametric dimensionality detection procedures. *Applied Measurement in Education, 19*(1), 67-91. https://doi.org/10.1207/s15324818ame1901_4

Nandakumar, R., & Ackerman, T. (2004). *Test modeling*. In D. Kaplan (Ed.), The SAGE handbook of quantitative methodology for the social sciences (pp. 93 107). SAGE Publications.

Reise, S.P., & Waller, N.G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods, 8*(2), 164-184. https://doi.org/10.1037/1082-989X.8.2.164

Roussos, L.A., & Özbek, Ö.Y. (2006). Formulation of the DETECT population parameter and evaluation of detect estimator bias. *Journal of Educational Measurement, 43*(3), 215-243. https://doi.org/10.1111/j.1745-3984.2006.00014.x

Roussos, L.A., Stout, W.F., & Marden, J.I. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement, 35*(1), 1-30. https://doi.org/10.1111/j.1745-3984.1998.tb00525.x

Roznowski, M., Humphreys, L.G., & Davey, T. (1994). A simplex fitting approach to dimensionality assessment of binary data matrices. *Educational and Psychological Measurement, 54*(2), 263-283. https://doi.org/10.1177/0013164494054002002

Sick, J. (2010). *Assumptions and requirements of Rasch measurement*. SHIKEN: JALT Testing & Evaluation SIG Newsletter, 14(2), 23-29.

Sijtsma, K., & Molenaar, I.W. (2002). *Introduction to nonparametric item response theory*. SAGE Publications.

Sireci, S.G. (2009). *Packing and unpacking sources of validity evidence: History repeats itself again*. In R.W., Lissitz (Ed.), The concept of validity: Revisions, new directions and applications (pp. 19–37). IAP Information Age Publishing.

Slocum-Gori, S.L., & Zumbo, B.D. (2011). Assessing the unidimensionality of psychological scales: Using multiple criteria from factor analysis. *Social Indicators Research, 102*(3), 443-461. https://doi.org/10.1007/s11205-010-9682-8

Stochl J, Jones P.B., & Croudace, T.J. (2012). Mokken scale analysis of mental health and well-being questionnaire item responses: A non-parametric IRT method in empirical research for applied health researchers. *BMC Medical Research Methodology, 12*(1), 1-16. https://doi.org/10.1186/1471-2288-12-74

Stocking, M.L., Swanson, L., & Pearlman, M. (1991). *Automated item selection using item response theory*. Educational Testing Service.

Stocking, M.L., Swanson, L., & Pearlman, M. (1993). Application of an automated item selection method to real data. *Applied Psychological Measurement, 17*(2), 167-176. https://doi.org/10.1177/014662169301700206

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*(4), 589-617. https://doi.org/10.1007/BF02294821

Stout, W. F. (1990). A new item response theory modelling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 49*(1), 293-325. https://doi.org/10.1007/BF02295289

Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika, 67*(4), 485-518. https://doi.org/10.1007/BF02295128

Stout, W., Froelich, A.G., & Gao, F. (2001). *Using resampling methods to produce an improved DIMTEST procedure*. In A. Boomsma, M.A.J. van Duijn, & T.A.B., Snijders (Eds.), Essays on item response theory (pp. 357-375). Springer.

Stout, W., Habing, B., Douglas, J., Kim, H.R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20*(4), 331-354. https://doi.org/10.1177/014662169602000403

Stout, W., Nandakumar, R., & Habing, B. (1996). Analysis of latent dimensionality of dichotomously and polytomously scored test data. *Psychometrika, 23*(1), 37-65. https://doi.org/10.2333/bhmk.23.37

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement, 27*(1),159-203.

van Abswoude, A.A., van der Ark, L.A., & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. Applied Psychological Measurement, 28(1), 3-24. https://doi.org/10.1177/0146621603259277

van der Ark, L.A. (2007) Mokken scale analysis in R. *Journal of Statistical Software, 20*(11), 1-19. https://doi.org/10.18637/jss.v020.i11

van der Eijk, C. & Jonathan, R. (2015). Risky business: Factor analysis of survey data-assessing the probability of incorrect dimensionalisation. *PLoS One, 10*(3), 1-35. https://doi.org/10.1371/journal.pone.0118900

Wismeijer, A.A.J, Sijtsma, K., van Assen M.A.L.M & Vingerhoets, J.J.M. (2008). A comparative study of the dimensionality of the self-concealment scale using principal

components analysis and mokken scale analysis. *Journal of Personality Assessment, 90*(4), 323-334. https://doi.org/10.1080/00223890802107875

Yu, C.H., Osborn Popp, S., DiGangi, S., & Jannasch Pennell, A. (2007). Assessing unidimensionality: A comparison of Rasch modeling, parallel analysis, and TETRAD. *Practical Assessment, Research, and Evaluation, 12*(1), 1-19. https://doi.org/10.7275/q7g0-vt50

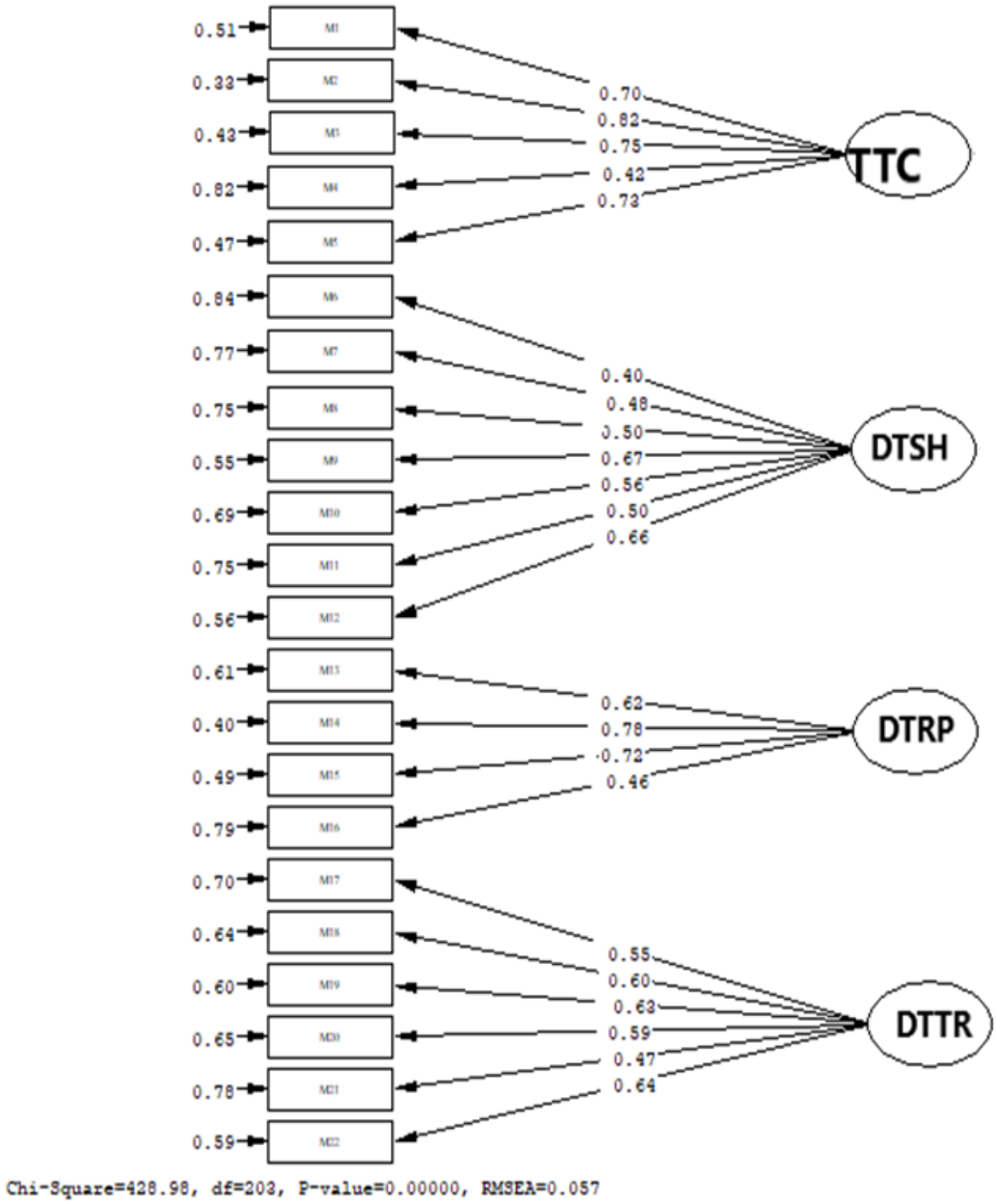
Yu, F., & Nandakumar, R. (2001). Poly Detect for quantifying the degree of multidimensionality of item response data. *Journal of Educational Measurement, 38*(2), 99-120. https://doi.org/10.1111/j.1745-3984.2001.tb01118.x

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*(2), 213-249. https://doi.org/10.1007/ß

Zumbo, B.D. (2009). *Validity as contextualized and pragmatic explanation, and its implications for validation practice*. In R.W., Lissitz (Ed.), The Concept of Validity: Revisions, New Directions and Applications. IAP Information Age Publishing.

APPENDIX

The Path Diagram Provided by the CFA



TTP: Tendency towards cheating

DTSH: Dishonesty tendency at studies as homework

DTRP: Dishonesty tendency at research and process of write up

DTTR: Dishonesty tendency towards reference

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 4, 831–847
https://doi.org/10.21449/ijate.1084501
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Views of academician examiners on the testing accommodations of the measurement, selection and placement center for disabled test takers

Mustafa Ilhan[1,*], **Melek Gulsah Sahin**[2], **Bayram Cetin**[3]

[1]Dicle University, Ziya Gökalp Faculty of Education, Department Mathematics and Science Education, Diyarbakır, Türkiye
[2]Gazi University, Gazi Faculty of Education, Department of Educational Sciences, Ankara, Türkiye
[3]Gaziantep University, Gaziantep Faculty of Education, Department of Educational Sciences, Gaziantep, Türkiye

ARTICLE HISTORY
Received: Mar. 8, 2022
Revised: Dec. 17, 2022
Accepted: Nov. 3, 2022

Keywords:
High–stake tests,
Testing accommodations for disabled examinees,
Testing accommodations of MSPC,
Academicians' views.

Abstract: The Measurement, Selection and Placement Center–MSPC (original acronym: Ölçme, Seçme ve Yerleştirme Merkezi–ÖSYM) administers many of the high-stake examinations applied in Türkiye. In order to support equality of opportunity in education and create a fair evaluation system, MSPC actualizes various testing accommodations by adjusting the standardization protocol for disabled test takers. In this research, we examined the views of academicians who served in the halls where disabled candidates take the test in the examinations held by MSPC about the testing accommodations for the disabled. The study design was in the basic qualitative research model. The participants consisted of 12 academicians working at a state university in Türkiye, who had served at least three times in the examination halls reserved for disabled test takers by MSPC. We collected the data via an interview form which included four items and administered it to the participants according to the drop-off and pick-up later method. The research results revealed that academician examiners have various positive opinions about MSPC's testing accommodations for disabled test takers. However, the participants also expressed that current accommodations have certain limitations that should be revised.

1. INTRODUCTION

Educational placement and admissions to institutions such as identification of personnel to be employed in the public sector and selection of students for certain types of high schools (e.g., science high schools) and universities are made through high-stake examinations in Türkiye as in many other countries of the world. The main issue in high-stake tests, which are a common part of educational systems, is to standardize the administration procedures as much as possible (Engelhard et al., 2010). Standardization refers to the administration and scoring of tests under uniform conditions for all examinees (Geisinger, 1994). However, some aspects of standardization make the administration of these tests unfair to certain groups, especially to individuals with disabilities (Sireci et al. 2005). More clearly, for some subgroups the validity of inferences from standardized test results may be doubtful because certain characteristics of

*CONTACT: Mustafa ILHAN ✉ mustafailhan21@gmail.com Dicle University, Ziya Gökalp Faculty of Education, Department Mathematics and Science Education, Diyarbakır, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

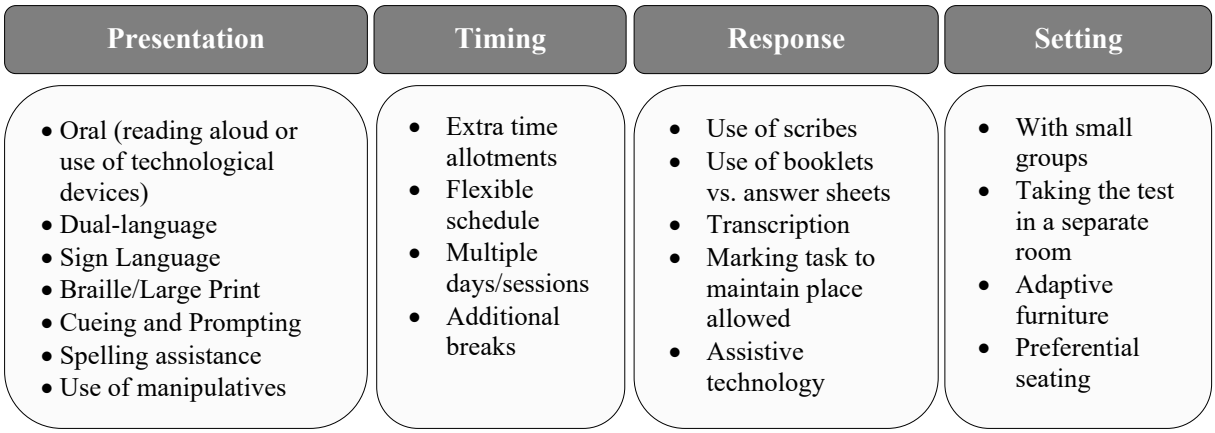
the individuals in these groups can impede their performance on the test and the scores do not correctly reflect the outcomes that the test purports to measure (Elliott et al., 2001; Schulte, et al., 2001). As a matter of fact, Sireci (2008) noted that strict standardization brought along a lack of fairness in the measurement process for certain test takers, which derived a favorable ground for construct-irrelevant variance to disseminate. Therefore, fairness as a fundamental validity issue requires attention in high-stake tests.

According to *Standards for Educational and Psychological Testing* published by American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) the broad target of fairness in testing is assuring equality of opportunity in the society. From a psychometric perspective the objective of fairness is maximizing, to the extent possible, the opportunity for examinees to demonstrate what they know on the trait the test is intended to measure and also minimizing the situations that are likely generate advantages or disadvantages for some test takers due to the characteristics irrelevant to the intended construct (AERA, APA & NCME, 2014). In this sense, test administration conditions must be regulated for disadvantaged subgroups by eliminating construct-irrelevant obstacles in order to establish fairness. In particular, as various disabilities may compromise examinees' opportunity to fully display their knowledge and skills in areas measured by the test, and thus unfairly disadvantage these individuals, assessment accommodations must be enabled for disabled test takers (Saka et al., 2022).

1.1. Testing Accommodations

The Standards for Educational and Psychological Testing uses the term accommodations to specify the changes to the presentation and/or format of the test and the way of administration or response procedures that maintain the nature of the target construct and result in scores comparable to those on the original test (AERA, APA & NCME, 2014). Bolt and Thurlow (2004) pointed out that although the terms of *test modifications* and *test accommodations* are often used interchangeably, these terms actually have different meanings. While modification remarks the alterations that change the test construct in some way, accommodation denotes the changes that aid in the measurement of a given construct. That is to say, testing accommodation implies altering established standardization protocol and test administration procedure without modifying test construct for curtailing the effect of the examinee's disability on his/her test result (Huynh & Barton, 2006; Sireci et al., 2003). The relevant changes could be related to how the test will be presented, how it will be responded, how the responses will be recorded, where the test will be administered, the type of equipment that will be allowed, and timing or scheduling of the test (Thurlow et al., 1993). Figure 1 summarizes the principal testing accommodations for disabled examinees in high-stake tests.

Figure 1. *Principal testing accommodations in high-stake tests (Fuchs et al. 2005; Prater, 2018; Sireci et al., 2003; Thurlow et al. 1993; Weis et al., 2014).*



From Figure 1 we see that the assistive technologies can be utilized both in the presentation of the test and in recording the responses. Oral accommodation can be performed as the presentation of the test direction and items by reading aloud or by means of technological devices such audiotape, videotape, and screen-reading software. Similarly, the responses can be dictated to a scribe by examinees or recorded through the speech-to-text software. Besides, examinees may take the advantage of technological devices (e.g., calculator, magnifier, zoomtext software) while answering the test items. In the relevant literature it has been reported that granting extended time is the most common test accommodation (Wightman, 1993). Reading a test aloud to the examinee, provision of a scribe to note the examinee's oral responses, presenting large print or braille booklets, and administering the test in a separate room are other test accommodations most frequently authorized (Bolt & Thurlow, 2004). Examinees might or might not consider the test as easier or more pleasant with listed accommodations, but either way the accommodations should spark off more accurate estimates of test takers' levels of target skills (Lovett & Leja, 2013).

1.2. Testing Accommodations in Türkiye

In Türkiye, various legal arrangements were adopted in order to accommodate the measurement and assessment practice to the disabled individuals' special needs. In Official Gazette of the Republic of Türkiye dated 2018 and numbered 30472, the following items were included regarding the measurement and evaluation processes of the individuals with disabilities: *(i)* students with visual impairment can be tested with other questions equivalent to these questions instead of questions containing pictures, figures and graphics, *(ii)* students with motor skills deficiency can be exempted from the applied parts of the courses requiring motor skills, and *(iii)* students with hearing impairment, intellectual disability or autism can also be exempted from foreign language exams. Such accommodations are considered in both classroom assessments and high-stake tests. Accordingly, not only teachers but also institutions conducting high-stake tests implement different accommodation policies according to the test takers' special needs in their examinations. In this direction, the Measurement, Selection and Placement Center–MSPC (original acronym: Ölçme Seçme ve Yerleştirme Merkezi–ÖSYM), which conducts most parts of the high-stake tests in Türkiye, executes certain testing accommodations for disabled examinees. MSPC published a special edict in 2018 and explained the accommodations provided to test takers with disabilities as follows: (on the condition that the examinee submits the petition stating his/her disability status, a certified copy of his/her health board report, completed health/disability information form, and a copy of the examinee application registration information to MSPC):

- Depending on the disability/health condition (Cerebral palsy patients who cannot control their body movements because the motor system mechanism in their bodies is not sufficiently developed, those who are visually impaired, those with pervasive developmental disorders, and those with specific learning difficulties), the examinee is provided with marker and/or reader assistance. While reader reads the direction and items to the examinee aloud and verbatim, marker transcribes response to the answer sheet once examinee completes an item. Two proctors in the same examination hall serve alternately as readers and markers.
- These test takers are allowed additional time to a predetermined extent according to the exam duration and the number of questions in the exam.
- Questions containing complex expressions and/or visual data such as figures, graphics, tables, pictures are not asked to the visually impaired examinees who request reader assistance.
- Additional time is given to examinees who can read the questions themselves (not requesting readers) but have special needs and vision impairment above 25%.
- Examinees with low vision but can read the questions by themselves are given a question booklet written in 9 or 14 font sizes upon their request, and marker assistance is provided.

- Examinees with pervasive developmental disorder in the unclassifiable group, mental retardation, specific learning difficulties, and those with deaf/mutes/hearing impaired making involuntary sounds can take the test in single-person halls if they wish, even if they do not receive reader and marker assistance.
- In electronic exams (e–exam), visually impaired examinees can take the test with screen reader software or screen magnifier software upon their request.
- Examinees with physical disabilities are assigned to the examination halls suitable for their status (to the exam buildings with working elevators or to the examination halls on the ground floors of the buildings) taking into account the information they provide in the health status/disability information form (can climb stairs, has difficulty in climbing stairs and cannot climb stairs).
- Examinees are allowed to bring drugs, equipment, devices, and materials related to their current disability/health status to the exam hall. According to this;
    - Examinees with hearing impairment who use hearing aids/bionic ears and whose condition is written in their exam entry documents are taken to the exam buildings with the relevant devices. However, these examinees should leave these devices at the place indicated by staff who serve in the hall to receive them after the exam is completed. Examinees who want to wear the aforementioned devices during the test should mark the relevant field in the health status/disability information form. Examinees who fill in the related form are taken to the test with their relevant devices in the examination halls prepared by MSPC, where all wired/wireless communication is cut off.
    - Examinees with diabetes are allowed to bring insulin pump, glucometer, supplementary food, etc. to the examination hall.
    - Examinees taking drug due to a chronic illness are allowed to bring the drug with them.
    - Examinees with temporary health problems or special conditions such as pregnancy are tolerated to meet their needs such as additional food and toilet. In addition, these examinees are permitted to bring the materials (drug, bandage, crutch, walking stick, neck brace, plaster, seat squab, etc.) they need for their health problems to the examination hall. These test takers are provided with marker assistance in line with their requests by applying the normal test duration.

As can be understood from these listed principles, the number and combination of testing accommodations implemented by MSPC are vast and diverse. The said accommodations are mainly based on altering the way the test is administered (e.g., the duration of the test, altering the format of presentation) without changing the content of it except for the items to which the examinee is exempt. Thus, MSPC intends to obtain a more accurate picture of the abilities of disabled examinees.

1.3. Purpose and Significance of the Research

The institutions that carry out high-stake tests need to pay attention to balance the individual rights of the disabled examinees against the obligation to maintain the integrity of the testing enterprise when planning testing accommodation policies (Phillips, 1994). Furthermore, these institutions should not overlook that there are two sides of the same coin when it comes to testing accommodations. Specifically, testing accommodations have the potential to eliminate construct-irrelevant variance and promote validity by removing barriers that prevent disabled examinees from demonstrating their actual abilities. But the flipside of the coin is that an accommodation may also inadvertently introduce construct-irrelevant variance if it alters the trait tested (Sireci 2008). Therefore, in order to ascertain how well the testing accommodation put in the practice in the pursuit of fairness serves its goal, it is important to reveal the positive and limitation aspects of the existing accommodations. We believe that it is especially important to scrutinize the views of examiners (i.e., the proctors/readers/scribes who serve in the

examination halls where disabled candidates take the test) on the subject, since they can directly observe the effective and limited aspects of the actualized accommodations for disabled test takers. That is to say, academician examiners may introduce important data on how much the accommodations set forth by MSPC are implemented and what kind of problems there are in practice. For example, a vision impairment examinee may not be able to form a view on how accurately the questions he/she is exempted from are determined because he/she cannot see the exam booklet. Nonetheless, the academicians who serve as a reader in the hall where the disabled examinee take the test can see both the booklet and how much the candidate can understand the orally presented items. Correspondingly, they can provide important information about how accurately the items that the candidate is exempted from are determined. In this context, we aimed to investigate the views of academicians who serve as proctors, readers, scribes, etc. in the high-stake tests administered by MSPC for disabled examinees about the testing accommodations implemented in these exams and sought answers to the following research questions:

1. What are the participants' views about the positive aspects of the available testing accommodations of MSPC?
2. What are the participants' views on the limited aspects of current testing accommodations of MSPC?
3. What are the challenges that the participants encounter while serving in the examination halls reserved by MSPC for disabled test takers?
4. What are the participants' suggestions for improving the MSPC's existing testing accommodations?

When we review the related literature, we see that there are studies about the testing accommodations for the disabled examinees in high-stake tests in Türkiye. For example, Şenel (2015) examined the experiences of visually impaired students in the university entrance exam while Tavşancıl et al. (2012) conducted a study to research the problems faced by visually impaired students in the university entrance exam and to offer solutions in this direction. In addition, Karabay (2016) investigated the effect of live reader and computer assisted reading on test score of visually impaired students. Şenel (2017) also tried to determine the suitability of computer adaptive tests for visually impaired students. On the other hand, Ozarkan et al. (2017) tested whether the items in the mathematics subtest administered in the scope of transition from basic education to secondary education in the first semester of 2015–2016 academic year show the differential item function in terms of the examinees' visual impairment status. Furthermore, Çobanoğlu-Aktan et al. (2018) compared the high-stake exams for disabled students in Türkiye and the USA in terms of legal responsibilities, administration methods and validity while Yılmaz (2019) analyzed the central common tests held in order to select students for high schools in terms of item bias according to the disability status of the examinees. In another study Dogus et al. (2020) examined the views of visually impaired students on the accommodations in high stake tests. Şenel (2021), on the other hand, explored the measurement invariance of the central examination applied in order to select students for secondary education institutions in Türkiye according to participants' disability status. However, in the relevant literature, there is no study that investigates the accommodations implemented in high-stake tests for examinees with disabilities directly based on the views of academicians taking office in these examinations. Therefore, the study is thought to contribute to the literature.

2. METHOD

2.1. Research Model

We carried out the study according to the basic qualitative research. Basic qualitative research, most common type of qualitative study found in education and most likely in other fields of

practice, imparts rich descriptive accounts aimed to understanding a phenomenon, an experience or a process from the perspective of the participants (Ary et al., 2019; Meriam, 2009). This specific research focuses on how events, processes, and activities are viewed by those involved in the study and also on purposes to describing recurrent themes or patterns in the data obtained (Ary et al., 2019).

2.2. Participants

Considering the aim and design of our study, we determined our participants according to convenience and criterion sampling, which are among the purposive sampling methods. We selected the participants from the academicians in our close circle, who we know take office in the examination halls where disabled candidates take the test, and we adhered to the criterion of having served in the examination halls allocated for disabilities at least three times. As such, we were able to reach 15 academicians, 12 of which were participants in our study (The other three academicians, to whom we forwarded the data collection tool, did not get back). We coded the participants as P1, P2,..., P12 within the scope of the study. All of the participants notified that they took office in the examination halls reserved for disabled test takers in the Higher Education Institutions Exam and Disabled Public Personnel Selection Exam. Those participants with codes P1, P4, P5, P10, and P11 reported that they had served in the examination halls reserved for the disabilities also in the Academic Staff and Graduate Education Exam in addition to the aforementioned two exams. The participants' missions in the halls where disabled examinees take the test were as follows: All 12 participants remarked that they served as reader/marker/scribe in the halls where visually impaired examinees take the test. In addition, the participant coded with P11 expressed that he took office as a marker/scribe in the hall where an examinee with cerebral palsy took the test. P2 and P5 coded participants, on the other hand, stated that they had served as a reader and marker in the halls where examinees with a special learning disability take the test.

2.3. Instrument

We collected the study data through an interview form consisting of four items. We prepared the items in the interview form in line with the research problems. Accordingly, we asked the participants to remark their opinions about the positive aspects and limitations of the existing accommodations in the first and second items, respectively. The third item was about the difficulties encountered during the task and the fourth one was regarding the suggestions for improving the current testing accommodations. After we created the draft form for the interview form, we received opinions from two measurement and evaluation experts. We asked the experts to judge the items in the instrument in terms of suitability for the purpose and sub-problems of the research and clarity. The experts stated that the interview form served the purpose of the study and that no changes were necessary. Then, we sought the opinion of a Turkish language expert to review the interview form in terms of spelling and grammar rules. The Turkish language expert stated that the language used in the instrument was understandable, but she made some suggestions in terms of punctuation marks. We made the necessary changes in the form in line with these suggestions related to punctuation marks. Subsequently, we received opinions from two academicians who had previously taken office as a disabled hall staff in the exams by MSPC order to get feedback on the applicability of the interview form. The feedback we received showed that the interview form was ready for administration, thus we started the data collection process.

2.4. Data Collection Process

Before starting the data collection process, we obtained ethics committee approval regarding the compliance of the research with scientific ethics. Following this, we started the data collection process. As known, interview forms can be administered to the participants orally as

well as in written format. In the written format, the data collection tool can be administered in person (i.e., face-to-face), electronically (via mail or internet-based program) or by a researcher dropping off the instrument to intended participants so that they can complete and then return it at a later date (Manchaiah et al., 2022). We adapted dropping-off and pick-up method in our study. In this direction, we left the interview form to the potential participants and gave them instruction about the research purpose and a brief description related to the instrument. Besides, we stated that the participation in the study was on a voluntary basis and emphasized that the data would remain anonymous and would not be shared with any other person or institution. We dropped-off the data collection tool to the participants on the first working day of the week and picked it up on the last working day of the week. We delivered the measurement tool to 15 academicians and 12 of them returned it.

2.5. Data Analysis, Dependability, Credibility, and Transferability

We used content analysis while analyzing the participants' responses. The main purpose of content analysis is to reach the concepts and relationships that can explain the collected data (Yıldırım & Şimşek, 2016); in other words, the major aim is to reveal the patterns hidden in the data. In the study, we first identified four themes, each corresponding to a research question and thus to an item in the instrument. In the second stage, we analyzed the data in line with each theme and detected the words, phrases, and sentences that had close meanings. We created sub-categories based on the words/sentences we determined to have close meanings.

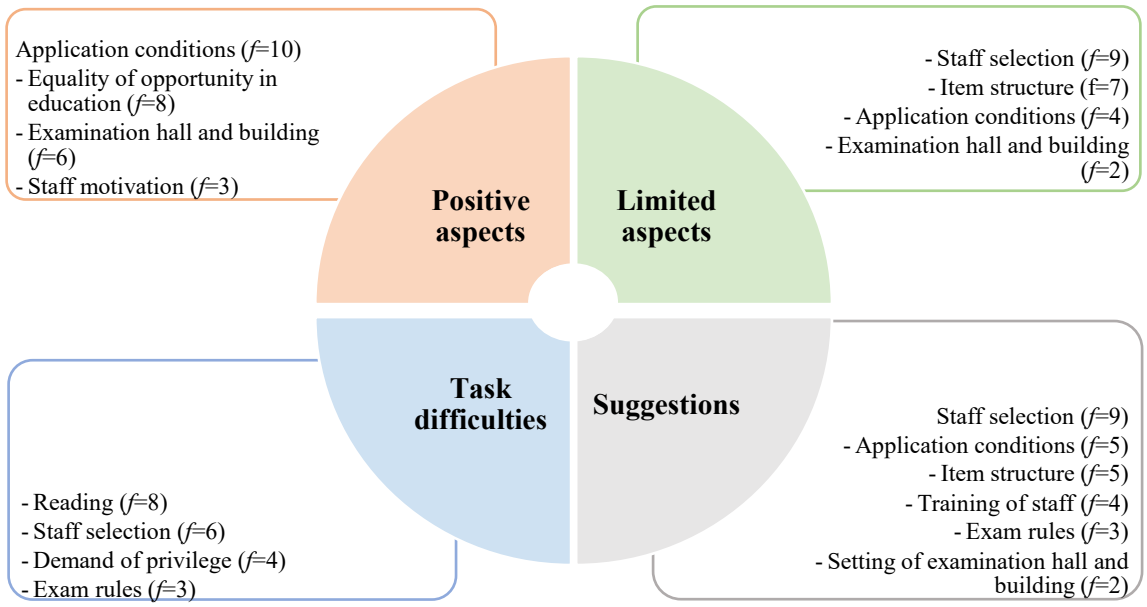
To ensure dependability of the study two independent researchers analyzed the data separately. Then we checked the consistency between the encodings using the formula of *"(number of agreements)/(total number of agreements + disagreements)*" proposed by Miles and Huberman (1994). We calculated the intercoder agreement as .90, .80, 1.00 and .96 for each theme, respectively. Later, in order to discuss differences of opinions among the coders and to achieve complete consistency, we sought the opinion of an assessment and evaluation expert who had previously served in the hall reserved for disabled test takers in the examinations by MSPC, and who was different from the researchers who coded the data. We held a Zoom meeting and got the opinion of the relevant expert about the coding of the data.

Brown and Rodgers (2022) specified that credibility and transferability will be enhanced if the researcher has a clear, complete, and detailed description of the study. Correspondingly, we took every care to describe the whole research process in elaboration for credibility and transferability of our study. We told in detail the steps we followed while collecting the data and explained one by one in which exams the research participants took office in the halls reserved for disabled candidates and their position in these examinations (i.e., reader, marker, etc.). Moreover, we included direct quotations from participants' views when presenting the results. We gave priority to the expressions that best reflected the situation while presenting direct quotations. Another procedure we performed for credibility and transferability was to obtain participant confirmation. Within this framework, after analyzing and reporting the collected data, we presented draft form of the results to the participants along with the instrument they answered, and we received feedback from the participants themselves on how accurately we interpreted their opinions.

3. RESULT

We grouped the participants' views about the testing accommodations for disabled examinees under four themes based on our research questions and the items in our instrument. We present these themes in Figure 2 along with the sub-categories under each theme.

Figure 2. *Themes and categories related to participants' views.*



As seen in Figure 2, the first theme was related to the positive aspects of MSPC's test accommodations for disabled test takers. This theme consisted of the categories of application conditions, equality of opportunity in education, examination hall and building, and staff motivation, respectively, according to the frequencies of occurrence. Table 1 displays direct quotations from the participants' views for each of these categories.

Table 1. *Direct quotations from participants' views for the positive aspects theme.*

Category	Examples of participants' expressions
Administration conditions	• *Preparing a booklet according to the examinee's disability* (P1) • *Additional time is provided, which makes the examinee feel more comfortable* (P3, P4) • *Taking drug or special equipment to the hall with the examinee* (P4)
Equality of opportunity in education	• *Providing some privileges to disabled examinees who cannot take the exam on equal terms* (P2) • *Subject candidates to test in the most appropriate and equal conditions possible* (P6) • *I find the test accommodations of MSPC for disabled examinees positive in the context of equality of opportunity in education* (P7)
Examination hall and building	• *Allotments of single-person classes for examinees with disabilities* (P5) • *Recording it with a camera ensures the reliability of the examination for the test taker* (P3) • *Allowing the exam hall door to be closed when necessary (e.g. when noise occurs due to reading aloud)* (P9) • *In general, all of the disabled examinees take the test in the same building and a dedicated coordinator is sent to these buildings by MSPC* (P9)
Staff motivation	• *Considering that it is more difficult compared to other duties, higher wages are paid to staff serving in examination halls allocated for disabled test takers compared to the proctors taking office in the other examination halls* (P9) • *While being proctor in normal examination halls may be boring, reading the questions in the halls where visually impaired candidates take the exam makes the time pass faster* (P3)

While the participants found MSPC's test accommodations for disabled examinees positive in various aspects, they also expressed that existing accommodations should be advanced in some respects. We named the theme, which includes the participants' opinions on the respects that should be improved in the test accommodations for disabilities, as limited aspects. Under this theme, there were four categories labeled as staff selection, item structure, application

conditions, and examination hall and building. Views were expressed most in the category of staff selection, and the least in the category of examination hall and building (see Figure 2). Table 2 illustrates direct quotations from the participants' views for the categories under the theme of limited aspects.

Table 2. *Direct quotations from participants' views for the theme of limited aspects.*

Category	Examples of participants' expressions
Staff selection	• *Reading clarity and fluency may differ from one reader to another and these differences can lead to unfairness among visually impaired examinees* (P9) • *It is troublesome for the reader that his/her field does not coincide with the test field he/she reads. This situation is a disadvantage for also examinees. For example, an academician from the verbal field may have problems especially when reading math questions. In one examination, the other personnel in the hall was from the verbal field and he/she read a question about factorial subject in mathematics as "5 with an exclamation point next to it". Unfortunately, such situations can happen. In yet another exam, the staff started the paragraph question by reading the paragraph directly. When the examinee asked her "do you read the item stem first", she replied as "what do you refer with item stem?"* (P9) • *We cannot interfere with the sudden changes in the health status of the disabled examinees who take the test with special equipment. ……. In addition, we are not asked whether we have first aid knowledge in the staff operating system* (P12).
Item structure	• *Sometimes we understand from the operations that visually impaired examinee asks us to write, that he/she is capable of solving a mathematical problem in question. However, we see that he/she could not give the correct answer as he/she could not do the operations required by the question himself/herself using paper-pencil and had to complete it in his/her mind after a point. We cannot provide support to the examinee in calculating the results of mathematical operations in these processes. Hence calculator support can be given at these points* (P9) • *Not exempting examinees from some questions (especially for visually impaired candidates)* (P1) • *In verbal ability questions requiring creating a paragraph by ordering the sentences presented, we usually encounter the examinees' "Let's skip this question" discourse. Examinees avoid answering such questions* (P3) • *The test takers are not exempted from some items that cannot be followed by listening (For instance, an item like "When a meaningful paragraph is formed from the five sentences given, which is the fourth sentence from the beginning?)"* (P9)
Application conditions	• *Conditions such as traffic and passenger density on arrival and departure to the examination buildings force the disabled examinees. That's why, even if it is not possible for candidates to take the exam at their home, it should be possible to take the exam on different dates. If no facilities are available, special services should be available for these examinees. It is not right to force the examinees to come and go to the exam buildings by public transportation or in their own vehicles* (P12) • *It is better to conduct such exams by means of computers than with the staff who will be assigned* (P5)
Examination hall and building	• *Buildings and halls are being tried to be suitable for disabled candidates, but inspections are insufficient* (P12) • *Sometimes halls outside the ground floors are allocated for disabled candidates* (P1)

Another theme that emerged as a result of the content analysis of the participants' views was about the difficulties experienced by the staff in the examination halls where disabled examinees take the test. So, we labelled this theme as task difficulties. There are four categories under this theme. These categories are respectively "reading, staff selection, demand of privilege, and exam rules", according to the frequency of expression by the participants. Table 3 exhibits direct quotations from the participants' views for each of the listed categories.

Table 3. *Direct quotations from participants' views for the theme of difficulties faced by examiners.*

Category	Examples of participants' expressions
Reading	• *Since the test duration is long, we sometimes have a sore throat while reading the questions* (P4) • *Although I did not encounter much challenges, there were times when I had difficulty reading the paragraph questions to the examinee* (P1) • *It is a big problem to read the questions in the booklets in a way that examinees can understand because a common language structure may not be ensured while reading the formulas, abbreviations etc. in some questions* (P12)
Staff selection	• *As a result of the assignment of staff not related to the test content that the examinee is responsible for, I had to carry out the task alone (The other staff in the examination hall did not have the mathematical knowledge to read the questions on the math test)* (P11) • *Sometimes the superintendent of the examination building does not have information about the accommodations for disabled candidates* (P9)
Demand of privilege	• *Some examinees ask staff for help in answering the questions* (P3, P4) • *Sometimes the examinee requests for help* (P12)
Exam rules	• *We sometimes have problems because examinees do not have enough knowledge about the exam rules. For example, some examinees think that they can read the questions themselves, even though they request a reader. When we say that MSPC does not allow examinees who request reader assistance to see the booklet, they react* (P9) • *We try not to show the booklet as visually impaired candidates are not allowed to read the questions themselves. But still, we sometimes have concerns like: "Does the examinee see the booklet, does the camera record, will we be punished?"* (P10)

We entitled the last theme that arose as a result of the content analysis as suggestions. This theme includes participants' views on what can be done to reduce the difficulties faced by the hall staff during their duty and to improve the test accommodations for disabled examinees. The suggestions expressed by the participants were grouped under six categories: staff selection, application conditions, item structure, staff training, exam rules, and setting of examination hall and building. Table 4 presents direct quotations from the participants' views for the theme of suggestions.

Table 4. *Direct quotations from participants' views for suggestions theme.*

Category	Examples of participants' expressions
Staff selection	• *The staff in the examination halls (i.e., proctor, scribe and especially reader) should be selected based on their expertise field* (P5) • *In particular, the staff assigned as readers should be put through a trial application at the MSPC centers in the provinces before the exam, and how well/intelligible they can read should be tested. The staff to be assigned should be selected according to the results of this test* (P9) • *Readers need to be given a professional education* (P4) • *The assignments of staff to the examination halls must be made as one female and one male so that they can chaperone the disabled examinee who is allowed to go to the toilet during the test* (P12)
Application conditions	• *Examination systems should be expanded in electronic environment and examinees should be allowed to take the test without leaving home* (P12) • *In order to prevent the effects arising from reader differences, applications for the effectiveness of computer-assisted reading can be considered instead of live readers* (P9) • *Before the examination, recordings can be taken where the questions are read by professional individuals. Thus, the test can be applied in a computer environment and the examinee can progress by pressing simple arrow keys* (P10) • *Examinees with disabilities can be offered a shuttle service to and from the examination building. A health worker must be present in these shuttles* (P12)

Item structure	• *In tests such as Turkish, History, and Geography, paragraph questions should be kept a little shorter for disabled examinees… It is difficult to keep in mind by listening to the paragraph items that take up almost half of the page with their answer options. Such long items create a situation to the detriment of disabled candidates* (P6) • *Since mentally handicapped examinees don't understand most of the items and they usually answer randomly, the items administered to these examinees should be different from the items of other disabled examinees* (P3) • *Visually impaired candidates should be exempted not only from questions containing figures/graphics, but also from lengthy questions that cannot be answered by listening* (P9)
Training of staff	• *Staff assigned to these examinations should receive a training, albeit a short one, before the exam* (P9) • *Individuals who want to serve in the examinations of disabled examinees should have at least one training/seminar on the sensitivities of disabled person* (P8)
Exam rules	• *We cannot take phones to the exam hall. An emergency response button should be sent to each hall in order to notify the superintendent of the exam building for emergency health problems* (P12) • *Examination staff chaperon the candidates in the exam building. It will be better if the chaperonage services are provided by the candidate's relative* (P12)
Setting of examination hall and building	• *A standard desk-table may not be the solution. There should be special exam centers where physically disabled people can easily take tests* (P10) • *Many details such as washbasins, emergency exits, routing tapes on the floor, elevators, and ramps need to be examined meticulously in the examination buildings* (P12) • *Ground floors should be allocated for disabled examinees or elevators should be working* (P1) • *There should be an appropriate desk for the physically disabled examinees who can read the questions themselves and solve them with pen and paper* (P10)

4. DISCUSSION and CONCLUSION

The present study was designed to set out the views of the academician examiners about the testing accommodations of MSPC for disabled test takers. Academicians expressed various positive aspects such as the application conditions, equality of opportunity in education, examination hall/building, and staff motivation for the accommodations of MSPC. Preparing a booklet suitable for the disability of the examinee, provision of additional time to the examinee, exempting the examinee from certain questions according to her/his disability, offering the examinee the facility to take the test in a single person-hall depending on the his/her disability, and paying higher wages to the staff served in the halls allocated for the disabled are among the academicians' positive views related to these categories. Providing constant conditions for all examinees taking the tests is not enough to ensure fairness and to get valid measurements. In order to increase validity and talk about fairness in the real sense, it is necessary to accept that disabled examinees differ from other candidates due to their special conditions and to offer positive privileges to these examinees. From this point of view, testing accommodations are required because standard assessment formats and procedures can present obstacles to disabled students, which means they may not be able to display their abilities under normal assessment conditions (Douglas et al., 2015). Briefly, it is not enough that the rules of the game are equal. Fundamentally, the game must be fair (Şişman, 2014) and the playing field must be leveled for all players (Jarvis, 1996; Sireci, 2008). As a matter of fact, National Council on Measurement in Education (NCME) states establishing a fair and equitable assessment system as one of the basic principles of measurement and evaluation (https://www.ncme.org/home). The views expressed by the participants reflect that MSPC is trying to provide fairness in the examinations for all individuals with the accommodations it offers for disabled examinees.

As a result of the research, we detected that the academicians' views on the limited aspects of MSPC's testing accommodations for disabled examinees were collected in the staff selection, item structure, application conditions, and examination hall and building categories. It was

among the opinions expressed in the category of staff selection that the reader differences can cause unjustness among the visually impaired candidates and that the reader may be insufficient in reading some test items due to his/her field. These views are in line with the results obtained in the study of Şenel (2015) and Doğuş et al. (2020). Şenel (2015) analyzed the experiences of visually impaired students in university entrance exam and in this study, the participants stated that some readers had difficulties in reading especially mathematics questions due to their branches. Doğuş et al. (2020) investigated the opinions of individuals with visual impairment on the exam accommodations in high-stakes tests and reported that disabled test takers have problems in the exams due to the reading characteristics of the readers (such as diction, pronunciation, spelling, and intonation) and their lack of sufficient field knowledge. Similarly, in the study by Tavşancıl et al. (2012), visually impaired test takers who took the university entrance exam remarked reader related problems as one of the factors that cause difficulties for them in the examination.

In the item structure category, another category under the theme of limited aspects, the participants of our study emphasized that the examinees should have the opportunity to utilize a calculator in the questions that require four operations that cannot be done mentally. In addition, they drew attention to the fact that visually impaired candidates are not exempt from long questions that they cannot answer by listening. In parallel with this result, in the research conducted by Şenel (2015), visually impaired students stated that they experienced concentration problems in long questions (items with long paragraphs), and that it is debatable how the items they were exempted from were determined. Actually, MSPC (2018) exempts disabled test takers who request reader assistance in their examinations from the items containing tables, graphics, figures, and complex expressions. However, when the views of the participants are considered together with the results of the existing studies in the literature, it is understood that it is not sufficient to exempt the disabled candidates from the questions containing only visuals or complex expressions. Thus, we can allege that the items to which the test takers will be exempted should be determined as a result of a more detailed expert examination.

The third category under the theme of limited aspects was related to the application conditions of the exam. In this category, opinions were expressed that disabled examinees had difficulties in transportation to the exam building. Furthermore, it was stated that it would be better if the examinations for disabled individuals were computer-based instead of live readers/markers. This view regarding the tests for disabled examinees with the help of computers is supported by the study results of Çobanoğlu-Aktan et al. (2018). Çobanoglu-Aktan et al. (2018) sought the opinions of assessment and evaluation experts, and personnel specialized in visually impaired individuals on what can be done to improve the exams accommodations for disabled people. Experts stated that it would be more appropriate to conduct the tests in a computer environment and present the items to the examinees in the form of pre-recorded audio files. In the same vein, in Şenel's (2015) research, some of the visually impaired examinees stated that if they made a choice, they would prefer to take the test with computerized technologies instead of live readers.

The fourth and last category under the theme of limited aspects was related to the characteristics of the examination hall and building. The participants worded that efforts are shown to make the buildings suitable for the candidates with disabilities, but inspections are insufficient. Indeed, Tavşancıl et al. (2012) explored the problems faced by visually impaired students in the university entrance exam and found that some of the problems experienced were related to the hall in which the examination was held. Essentially, in the report published by MSPC (2018) on the subject, a framework has been drawn for disabled examinees to take the test in halls suitable for their special circumstances. However, the opinions expressed by the academicians

about the examination buildings signal that there are some problems in the practice of envisaged accommodations.

Another remarkable point about the limited aspects theme is the absence of a category related to the timing of the test, or put it another way, the research participants did not express any negative opinions regarding MSPC's accommodations of test timing. In the studies in the literature, it is stated that the most common practice among the test accommodations for the disabled examinees is the provision of extra time (Gregg & Nelson, 2015; Lovett, 2010). Therefore, we can say that MSPC effectively operates this accommodation, which is the most frequently fulfilled testing accommodation for disabled examinees in different countries.

The third theme that surfaced when we analyzed participants' opinions was the task difficulties. Participants stated that they sometimes had difficulties during their tasks due to such reasons as examinees' lack of knowledge about the exam rules and demanding privileges, the selection of readers who are not compatible with the test content, the lack of knowledge of some staff about the exam rules, and the wearying reading questions. When we probed these views, the following point draws our attention: Only the selection of the readers from these opinions is the responsibility of MSPC. Other opinions expressed are related to the staff/examinees not reading the exam rules well enough before the test and the examinees' manner during the test. To put it more clearly, a significant part of the difficulties experienced by exam staff during their task is pertinent to the other staff and examinees rather than the accommodations of MSPC.

When we look at the suggestions of the participants for the improvements of MSPC's test accommodations for disabled examinees, there appeared opinions such as more careful selection of the exam staff, providing training to disabled hall staff before their duty, using computer-assisted reading instead of employing live readers, determining the items to which the examinees will be exempted from a more detailed perusal, providing shuttle vehicle to the disabled examinees, and even switching to electronic tests where examinees can take tests at their home. These views generally overlap with the results obtained in the current studies in the literature. We can summarize this overlap as follows: In the study conducted by Çobanoğlu-Aktan et al. (2018), experts suggested that the tests for disabled examinees should be carried out on the computer environment. The suggestion of transferring the tests for disabled examinees to electronic environment was also expressed in the research by Tavşancıl et al. (2012) and Şenel (2015). Şenel (2015) also mentioned that the items to be exempted from the examinees should be determined more carefully based on expert opinions. Additionally, participants' views on the more careful selection of staff to be employed in halls for the disabled examinees and the provision of training to these personnel are in line with the results reported and recommendations made in the study of Doğuş et al. (2020).

To summarize, the academicians in the study group found the accommodations implemented by MSPC for disabled candidates positive in various aspects. Nevertheless, they expressed their opinions that the current accommodations are limited in some aspects, and therefore, the testing accommodations for disabled people should be developed by taking these limited aspects into account. It was stated that there is a need for precautions to reduce the limitations of existing accommodations, especially in terms of technology assistance, selection, and training of proctor/reader/scribe and setting of examination halls and buildings. The findings we have reached are analogous with such results obtained in previous studies on the subject. We anticipate that the conclusions we report will be beneficial for MSPC and the Ministry of National Education in their further accommodations to improve the examination practices for disabled individuals. Nonetheless, as the participants of this study were mostly academicians serving in the examination halls where visually impaired candidates take the test, the results obtained were able to provide limited information about what revisions should be made in the test accommodations for candidates in different disability groups. Thereby, we can recommend

carrying out similar studies with the people who serve in the examination halls where candidates from different disability groups take the test.

Acknowledgments

This study was orally presented at the 7th International Congress on Measurement and Evaluation in Education and Psychology.

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 and Protocol No:** Dicle University Social and Humanities Sciences Ethics Committee, Approval letter dated 18.03.2020 and numbered 34061.

Authorship Contribution Statement

Mustafa Ilhan: Investigation, creating of the instrument, data collection and analysis, resources, visualization, and writing-original draft. **Melek Gulsah Sahin**: Methodology, resources, data analysis, and writing-original draft. **Bayram Cetin:** Investigation, creating of the instrument, supervision, writing the original draft.

Orcid

Mustafa Ilhan https://orcid.org/0000-0003-1804-002X
Melek Gulsah Sahin https://orcid.org/0000-0001-5139-9777
Bayram Cetin https://orcid.org/0000-0001-5321-8028

REFERENCES

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

Ary, D., Jacobs, L.C., Sorensen Irvine, C.K., & Walker, D.A. (2019). *Introduction to research in education* (10th ed.). Cengage.

Bolt, S.E., & Thurlow, M.L. (2004). Five of the most frequently allowed testing accommodations in state policy: Synthesis of research. *Remedial and Special Education*, *25*(3), 141–152. https://doi.org/10.1177/07419325040250030

Brown, J.D., & Rodgers, T.S. (2022). *Doing second language research*. Oxford University.

Çobanoğlu-Aktan, D., Aksu, G., & Eser, M.T. (2018). Türkiye ve Amerika'da engelli öğrenciler için yapılan geniş ölçekli sınavların yasal sorumluluklar, uygulama yöntemleri ve geçerlik açısından incelenmesi [Investigation of legal responsibilities, practices and validity of the large-scale exams for students with disabilities in Türkiye and the US]. *Mersin University Journal of the Faculty of Education, 14*(1), 69-83. http://dx.doi.org/10.17860/mersinefd.322551

Doğuş, M., Aslan, C., & Çakmak, S. (2020). Görme engelli bireylerin merkezi sınav düzenlemelerine ilişkin görüşleri [The opinions of individuals with visual impairment on the exam accommodations in high-stakes tests]. *Journal of Research in Education and Society, 7*(1), 219–247. https://dergipark.org.tr/tr/pub/etad/issue/55359/697087

Douglas, G., McLinden, M., Robertson, C., Travers, J., & Smith, E. (2015). Including pupils with special educational needs and disability in national assessment: Comparison of three country case studies through an inclusive assessment framework. *International Journal of Disability, Development and Education, 63*(1), 98-121. http://dx.doi.org/10.1080/1034912x.2015.1111306

Elliott, S.N., Kratochwill, T.R., & McKevitt, B.C. (2001). Experimental analysis of the effects of testing accommodations on the scores of students with and without disabilities. *Journal of School Psychology*, *39*(1), 3–24. https://doi.org/10.1016/S0022-4405(00)00056-X

Engelhard G. Jr, Fincher M., & Domaleski C.S. (2010). Mathematics performance of students with and without disabilities under accommodated conditions using resource guides and calculators on high stakes tests. *Applied Measurement in Education*, *24*(1), 22–38. https://doi.org/10.1080/08957347.2010.485975

Fuchs, L.S., Fuchs, D., & Capizzi, A.M. (2005). Identifying appropriate test accommodations for students with learning disabilities. *Focus on Exceptional Children*, *37*(6), https://doi.org/10.17161/foec.v37i6.6812

Geisinger, K.F. (1994). Psychometric issues in testing students with disabilities. *Applied Measurement in Education*, *7*(2), 121-140. https://doi.org/10.1207/s15324818ame0702_2

Gregg, N., & Nelson, J.M. (2015). Meta-analysis on the effectiveness of extra time as a test accommodation for transitioning adolescents with learning disabilities: More questions than answers. *Journal of Learning Disabilities*, *45*(2), 128-138. https://doi.org/10.1177/0022219409355484

Huynh, H., & Barton, K.E. (2006). Performance of students with disabilities under regular and oral administrations of a high-stakes reading examination. *Applied Measurement in Education*, *19*(1), 21–39. https://doi.org/10.1207/s15324818ame1901_2

Jarvis, K.A. (1996). *Leveling the playing field: A comparison of scores of college students with and without learning disabilities on classroom tests* [Doctoral dissertation, Florida State University]. https://www.proquest.com/pagepdf/304232975?accountid=15780

Karabay, E. (2016). *Canlı okuyucu ve bilgisayar destekli okumanın görme engelli öğrencilerin test başarıları üzerindeki etkilerinin karşılaştırılması [Comparing the effects of the live-reader with computer-aided reading on test achievement of students with visual impairment]* [Doctoral dissertation, Ankara University]. National Thesis Center of Higher Education Board. https://tez.yok.gov.tr/UlusalTezMerkezi/

Lovett, B.J. (2010). Extended time testing accommodations for students with disabilities: Answers to five fundamental questions. *Review of Educational Research*, *80*(4), 611–638. https://doi.org/10.3102/0034654310364063

Lovett, B.J., & Leja, A.M. (2013). Students' perceptions of testing accommodations: What we know, what we need to know, and why it matters. *Journal of Applied School Psychology*, *29*(1), 72–89. https://doi.org/10.1080/15377903.2013.751477

Manchaiah, V., Beukes, E., & Roeser, R.J. (2022). *Evaluating and conducting research in audiology*. Plural.

Meriam, S.B. (2009). *Qualitative research a guide to design and implementation*. Jossey–Bass.

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

Official Newspaper of Republic of Türkiye. (2018, July 7). *Özel eğitim hizmetleri yönetmeliği [Special education services regulation]* (Number: 30471). https://www.resmigazete.gov.tr/eskiler/2018/07/20180707-8.htm

Ozarkan, H.B., Kucam, E., & Demir, E. (2017). An investigation of differential item functioning according to the visually handicapped situation for the central joint exam math subtest. *Curr Res Educ, 3*(1), 24–34.

Ölçme Seçme ve Yerleştirme Merkezi (Measurement Selection and Placement Center). (2018). *Engeli/sağlık sorunu veya özel durumu olan adaylara yapılan sınav uygulamaları [Testing accommodations for candidates with disability/health problems or special conditions]*. https://www.osym.gov.tr/TR,13930/engelisagliksorunu-veya-ozel-durumu-olan-adaylara-yapilan-sinav-uygulamalari.html

Phillips, S.E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education*, *7*(2), 93-120. https://doi.org/10.1207/s15324818ame0702_1

Prater, M.A. (2018). *Teaching students with high–incidence disabilities: strategies for diverse classrooms*. Sage.

Saka, N., Kleper, D., & Kennet-Cohen, T. (2022). Assessing fairness in selection toward applicants who request accommodations in higher education admissions tests. *Assessment in Education: Principles, Policy & Practice*. https://doi.org/10.1080/0969594X.2022.2122400

Schulte, A.A.G., Elliott, S.N., & Kratochwill, T.R. (2001). Effects of testing accommodations on standardized mathematics test scores: An experimental analysis of the performances of students with and without disabilities, *School Psychology Review*, *30*(4), 527–547. https://doi.org/10.1080/02796015.2001.12086133

Sireci, S.G. (2008). Validity issues in accommodating reading tests. *Jurnal Pendidik dan Pendidikan [Journal of Educators and Education 23*, 81-110. https://files.eric.ed.gov/fulltext/ED500434.pdf

Sireci, S.G., Li, S., & Scarpati, S. (2003). *The effects of test accommodation on test performance: A review of the literature*. Center for Educational Assessment Research Report no. 485. School of Education, University of Massachusetts Amherst. https://nceo.umn.edu/docs/OnlinePubs/TestAccommLitReview.pdf

Sireci, S.G., Scarpati, S.E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research*, *75*(4), 457–490. https://doi.org/10.3102/00346543075004

Şenel, S. (2015). Görme engelli öğrencilerin üniversite giriş sınavı deneyimleri [Experiences of visually impaired students in university entrance exam]. *Hacettepe University Graduate School of Educational Sciences the Journal of Educational Research, 1*(1), 1–9. https://dergipark.org.tr/tr/download/article-file/450547

Şenel, S. (2017). *Bilgisayar ortamında bireye uyarlanmış testlerin görme engelli öğrencilere uygunluğunun incelenmesi [Investigation of the compatibility of computerized adaptive testing on students with visually impaired]* [Doctoral dissertation, Ankara University]. National Thesis Center of Higher Education Board. https://tez.yok.gov.tr/UlusalTezMerkezi/

Şenel, S. (2021). Assessing measurement invariance of Turkish "Central examination for secondary education ınstitutions" for visually impaired students. *Educational Assessment, Evaluation and Accountability*, *33*(4), 621-648. https://doi.org/10.1007/s11092-020-09345-5

Şişman, Y. (2014). Engelliler açısından eşitlik, ayrımcılık ve eğitim hakkı [Equality, discrimination and right to education with regards to disabled people]. *Journal of Social Policy Studies, 32*, 57–85. https://doi.org/10.21560/spcd.37719

Tavşancıl, E., Uluman, M., & Furat, E. (2012, 19–21 September). *Görme engelli öğrencilerin üniversite giriş sınavında karşılaştığı sorunlar ve çözüm önerileri [Problems encountered by visually impaired students in the university entrance exam and suggestions for solutions]* [Conference presentation abstract]. III. National Congress of Assessment and Evaluation in Education and Psychology, Abant İzzet Baysal University, Bolu, Türkiye. https://www.epodder.org/wp-content/uploads/2020/07/EPOD-2012.pdf

Thurlow, M.L., Ysseldyke, J.E., & Silverstein, B. (1993). *Testing accommodations for students with disabilities: A review of the literature*. (Synthesis Report 4). Minneapolis: University of Minnesota, National Center on Educational Outcomes. https://files.eric.ed.gov/fulltext/ED358656.pdf

Weis, R., Dean, E.L., & Osborne, K.J. (2014). Accommodation decision making for postsecondary students with learning disabilities: individually tailored or one size fits all? *Journal of Learning Disabilities, 49*(5), 484-498. https://doi.org/10.1177/0022219414559648

Wightman, L.F. (1993). *Test takers with disabilities: A summary of data from special administrations of the LSAT* (LSAC-Research Report 93–03). Law School Admission Council.

Yıldırım, A., & Şimşek, H. (2016). *Sosyal bilimlerde nitel araştırma yöntemleri [Qualitative research methods in the social sciences]* (10th ed.). Seçkin.

Yılmaz, G. (2019). *Seçme sınavlarının engel durumlarına göre madde yanlılığının incelenmesi [An investigation of item bias for selection exams according to disability situations]* [Master thesis, Hacettepe University]. National Thesis Center of Higher Education Board. https://tez.yok.gov.tr/UlusalTezMerkezi/

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 4, 848–866
https://doi.org/10.21449/ijate.1127664
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Perceived-teacher presenteeism scale: A scale development study

Alper Uslukaya[1], **Zulfu Demirtas**[2], **Muslim Alanoglu**[3,*]

[1]Fırat University, Faculty of Education, Department of Educational Sciences, Elazığ, Türkiye
[2]Fırat University, Faculty of Education, Department of Educational Sciences, Elazığ, Türkiye
[3]National Police Department, Muş, Türkiye

ARTICLE HISTORY
Received: June 10, 2022
Revised: Dec. 8, 2022
Accepted: Dec. 17, 2022

Keywords:
Factorization,
Measurement invariance,
Presenteeism,
Scale development,
Teacher presenteeism.

Abstract: This study aims to develop and test the reliability and validity of a multi-item teachers' perceived presenteeism behavior scale. For this, first of all, a semi-structured interview form was applied to 57 teachers, an item pool was formed for the presenteeism scale with the data obtained, and the draft form of the scale was prepared in line with the expert opinions. Then, the draft scale form was applied to 382 teachers, and exploratory factor analysis was performed with the data obtained. As a result of the analysis, a three-dimensional scale structure consisting of 14 items was obtained. Data were collected from 303 teachers to confirm this structure, and the three-factors scale structure was confirmed based on acceptable fit values with confirmatory factor analysis. It was determined that the validated second-order three-factor model provided convergent and discriminant validity criteria. The measurement invariance of the scale according to gender, marital status, and age groups was tested, and it was observed that the same structure was measured in different groups. Cronbach Alpha internal consistency coefficient and composite reliability values showed that sufficient reliability values were achieved for the scale. Finally, the test-retest performed to test its stability showed that the scale was stable. Thus, it was concluded that the scale is valid and reliable with sufficient conditions to measure the teachers' perceptions of presenteeism.

1. INTRODUCTION

Continuing to work in inappropriate biopsychosocial conditions, referred to in the literature as presenteeism (Vera-Calzaretta & Juarez-Garcia, 2014). Research on presenteeism have shown that this experience has negative psychological effects on employees (Baker-McClearn et al., 2010; Cooper & Lu, 2016); organizational functioning (D'Abate & Eddy, 2007; Ferreira & Martinez, 2012), and affects production relations negatively (Gilbreath & Karimi, 2012). In addition, it was revealed that the negative effects on productivity resulted in a costly loss of approximately 150 billion USD in the USA (Hemp, 2004) and 225 billion Euros in Germany (Abasilim et al., 2015) over one year.

Despite these negative consequences, presenteeism is a new phenomenon for organizational researchers, and a consensus on its definition has yet to be reached (Cooper & Lu, 2016). Certain researchers (e.g., Aronsson et al., 2000; Dew et al., 2005; Kivimäki et al., 2005; Turpin et al., 2004) define presenteeism as the employees being at work while sick, merely by

*CONTACT: Muslim Alanoglu muslimalanoglu@gmail.com National Police Department, Muş, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

associating it with the sickness. However, some other researchers (e.g., D'Abate & Eddy, 2007; Evans, 2004; Johansson & Lundberg, 2004) define presenteeism as the employee's continuing to work despite the circumstances that prevent them from revealing the authentic performance in the workplace. Therefore, according to these researchers, presenteeism is defined as an experience that occurs as a result of many factors (chronic illness, workplace stress, non-work related occupations, special situations related to the employee and negative environmental factors, etc.). However, the tools used to measure the phenomenon in the literature have been developed based on the meaning of the employee continuing to work while sick (e.g., Aronsson et al., 2000; Koopman et al., 2002; Lohaus & Habermann, 2019; Lu et al. 2013; McGregor et al., 2016; Miraglia & Johns, 2016). These tools, which usually consist of one or two-item questions, are designed to measure the frequency of presenteeism (going to work while sick) or the loss of productivity caused by the presenteeism. For this reason, these measurement tools ignore the various dynamics (other than the disease) that the phenomenon may be associated with and its negative consequences other than loss of productivity.

Due to the lack of literature in this area, it is aimed to develop a presenteeism scale in this study according to the perceptions of teachers, who are considered as one of the occupational groups that have experienced presenteeism the most (Bergström et al., 2009; Lohaus & Habermann, 2019), taking into consideration the broadening meaning of presenteeism and its consequences other than loss of productivity.

1.1. What is Presenteeism?

There is inconsistent (Johns, 2011) and complex (Wang et al., 2010) literature on what presenteeism is. Three different research lines related to the concept can be mentioned. The first line of research -especially from European Researchers- defines presenteeism as "continuing to work while sick" (Johns, 2010) examines the phenomenon in a reductionist perspective by distinguishing its premises and consequences. This perspective focuses on factors related to employees, working conditions, and environmental factors associated with presenteeism (Karanika-Murray & Cooper, 2019).

The second line of research, represented by North American researchers (Johns, 2010) defines presenteeism as a loss of productivity due to continuing to work despite health problems (Goetzel et al., 2004; Turpin et al., 2004). This perspective focuses on measuring the loss of productivity caused by presenteeism (Goetzel et al., 2004; Koopman et al., 2002) and necessary medical interventions for emerging physical health problems (Ammendolia et al., 2016).

The first two perspectives formulate presenteeism within the framework of physical health problems, and this situation is called Sickness Presenteeism in the literature. Research that covers the concept more broadly and can be considered as a third line, in addition to physical health problems that will prevent the employee from performing optimally and collecting cognitive energy at work, stress (Gilbreath & Karimi, 2012), depression (Wang et al., 2010), non-work related deals (D'Abate & Eddy, 2007), environmental elements (Hansen & Andersen, 2008), etc. by associating variables with the phenomenon, it defines presenteeism as physically present, but functionally disappeared (Cooper & Lu, 2016). This perspective associates the state of being unwell that will prevent the employee from performing at a high level while at work, with the employee's health problems and organizational, individual and environmental variables. In this study, presenteeism is evaluated within the framework of the third line.

1.2. Presenteeism as a State of Unwell

Presenteeism studies generally focus on the health problems underlying dysfunction in the workplace (Evans, 2004; Johansson & Lundberg, 2004; Turpin et al., 2004). These studies concentrate on physical health problems such as allergies, diabetes, arthritis, asthma, heart

disease, hypertension, migraine/headache, fatigue, respiratory tract infections, neck and back pain (Aronsson et al., 2000; Baker-McClearn et al., 2010; Caverley et al., 2007; Kivimaki et al., 2005). However, the World Health Organization (WHO) defines health as a state of complete physical, mental and social well-being (Witmer & Sweeney, 1992).

Well-being, which is defined as a three-dimensional situation, is effective in increasing the capacity of the employee to use their abilities (Myers & Williard, 2003) and on their performance and productivity in the workplace. The negativity that may arise in any of these dimensions can hinder the energy to perform a task, attention, and motivation (Kiefer, 2008). Therefore, it would be incomplete to consider presenteeism as a process that starts with only physical health problems. Because presenteeism is an experience that begins with the employee's decision to continue working in inappropriate biological, psychological and social conditions (Vera-Calzaretta & Juarez-Garcia, 2014), and physical health problems can be associated with psychological and mental well-being variables that will put a person in a negative well-being state.

1.3. Presenteeism as a Fearful Process

Presenteeism associated with high cost losses was found to be more costly than absenteeism (Cooper & Dewe, 2008). Cross-sectional (Conner & Silvia, 2015; Miraglia & Johns, 2016) and longitudinal studies (Beswick et al., 2018; Chen et al., 2021; Demerouti et al., 2009; Lu et al., 2013) on the subject revealed that presenteeism predicts various negative outcomes. As a matter of fact, productivity in organizational life (Goetzel et al., 2004; Hemp, 2004; Turpin et al., 2004) is negatively associated with work speed, service quality, and organizational creativity (Gilbreath & Karimi, 2012); it is positively associated with work repetition, error rate, and work accidents (D'Abate & Eddy, 2007). It negatively affects employee's mental health, social relationships, physical health (Lu et al., 2013), performance (Berger et al., 2003), work energy (Roe, 2003), teamwork (Borrill et al., 2000), business relations and service quality (Borrill et al., 2000). As a result, presenteeism, which creates a perception of ineffectiveness in the workplace (Ferreira & Martinez, 2012), can be considered a fearful process that must be taken precautionary.

1.4. Teacher Presenteeism

Various variables are considered as basic dynamics that enable teachers to experience presenteeism by hindering them from taking absences based on excuses such as the importance of education-training for the future of students (Aronsson & Gustafsson, 2005) and the high sense of responsibility it creates (Widera et al., 2010), society's expectations from education and training (Grant, 2008), the perception that a teacher cannot be replaced in absenteeism (Caverley et al., 2007). Also, factors such as unsupportive organizational policies (Wrate, 1999), oppressive attitudes of the administration that do not have sufficient information about the effects of presenteeism, and an organizational climate that sees absenteeism as illegitimate (Dew et al., 2005) make teachers potential candidates for the experience of presenteeism. Therefore, teachers are among the employees who experience presenteeism the most (Aronsson et al., 2000; Bergström et al., 2009; Ferreira & Maritnez, 2012).

Intense and widespread experience of presenteeism among teachers may hinder creating and developing a positive and supportive school environment (Jennings & Greesnberg, 2009). It limits a healthy relationship with colleagues and students and a functional participation in the education-training process by predicting a negative mood. It may trigger failure in classroom management (Jennings & Greesnberg, 2009), the loss of the ability to be a correct model for students (Kidger et al., 2016) and weakening of belief in providing healthy guidance (Sisask et al., 2014). Considering these effects, presenteeism, which can cause psychological problems (Perez-Nebra et al., 2020) and learning difficulties (Jennings & Greesnberg, 2009) for students,

can be evaluated as a process that must be taken into consideration in terms of education. No scale has been found in the literature to measure teachers' perceptions of Presenteeism, an experience that cannot be ignored. Therefore, it is important to bring the perceived teacher presenteeism scale, which can be used in presenteeism studies, to the literature. In this context, it is aimed to develop the Perceived-Teacher Presenteeism Scale (P-TPS) in the study.

2. METHOD

2.1. Research Model

This study aimed to develop the perceived teacher presenteeism scale by using a sequential investigative design from mixed methods research. Exploratory sequential design is a sequential process in which the researcher begins qualitative research and continues using a quantitative sequence (Creswell & Plano Clark, 2011). In the study, first of all, a scale item pool regarding presenteeism was created with qualitative data, and the content validity of the item pool was tested by consulting with the field experts. Then, the validity and reliability analyzes of the scale were carried out with the quantitative data that had been collected.

2.2. Study Group

To collect the data to be used in the study, ethical approval was obtained (Ethic no: 20.01.2021-E-97132852-302.14.01-6275), and the research application permission was obtained from Elazıg Governorship Provincial Directorate of National Education (Ethic no: 19.04.2021- E-79137285-605.01). For this study, which was carried out in four stages, data collected from four different study groups were used. During the 2021-2022 academic year, all study data were collected from teachers working in Elazig's city center. The researchers collected the data by personally interviewing the teachers. It was checked whether there was missing data in the data sets, and incompletely filled forms were removed from the data set. In the first stage, which was carried out in the form of qualitative analysis, the opinions of 57 teachers selected by purposive sampling were taken with a semi-structured interview form. In the second stage, the scale was applied to 382 high school teachers, and these data were used in Exploratory Factor Analysis (EFA). In the third stage, data collected from 303 secondary school teachers were used for Confirmatory Factor Analysis (CFA) and measurement invariance analyses. In the fourth stage, data were collected from 109 primary school teachers for test-retest reliability analysis. Information about the research participants is presented in Table 1.

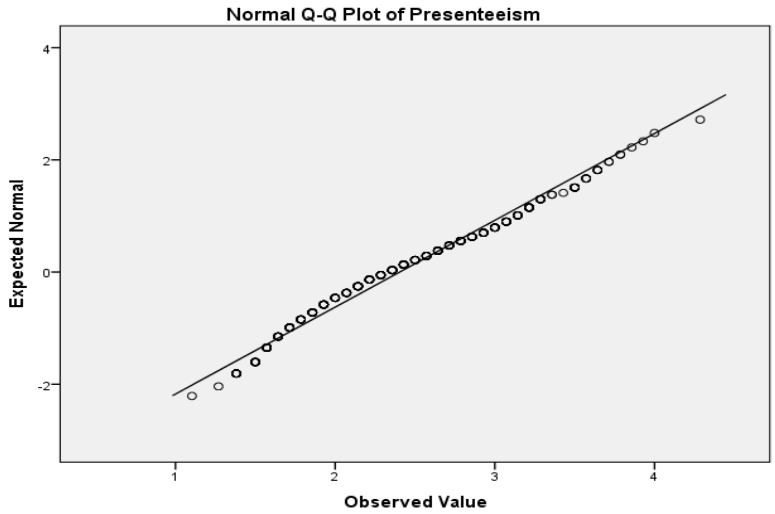
Table 1. *Demographic information of participants.*

		1. Step	2. Step	3. Step	4. Step
		N = 57	N = 382	N = 303	N = 109
Gender	Female	32	200	136	63
	Male	25	182	167	46
Marital Status	Married	36	287	193	87
	Single	21	95	110	22
Age	21-30	8	70	71	-
	31-40	18	131	86	23
	41-50	17	124	88	50
	51-60	11	55	58	35
	61+	3	2	-	1
Instructional Positions	Pre-school	14	-	-	-
	Primary school	17	-	-	109
	Secondary School	12	-	303	-
	High school	14	382	-	-

2.2. Data Analysis

In this study, it was suggested by DeVellis (2003) to be followed in scale development studies; drawing the conceptual and theoretical framework of the variable to be measured, creating the item pool, determining the measurement method, evaluating the item pool by experts, reliability analysis, validity analysis, and finalizing the scale were followed. First, qualitative data were subjected to descriptive and content analysis in the study, and then opinions on content validity were obtained from three education administration field experts and one measurement and evaluation expert. For the semantic validity of the scale, the opinions of two Turkish language experts were consulted. Then, in order to test the comprehensibility of the scale, a focus group interview was held with five teachers, and after it was determined that the scale was comprehensible, factor analysis was started. Kurtosis and skewness values were checked to see if the data sets met the univariate normality, and Mahalanobis distance values and Q-Q graph were checked for the multivariate normality. Since the kurtosis and skewness values are in the range of ± 1 the univariate normality assumption was met (Cokluk et al., 2010). The multivariate normality assumption was provided since Mahalanobis distance values approaching zero were obtained (Seçer, 2015) and as seen in Figure 1, the points were close to the 45-degree reference line on the Q-Q plot.

Figure 1. *Q-Q plot graph.*



For factorization, EFA was performed with the help of the SPSS 22 package program. Kaiser-Meyer-Olkin (KMO) sample adequacy coefficient and Bartlett's Sphericity Test results were examined for the suitability of the data for factor analysis. Since the Bartlett's Test of Sphericity is significant and the KMO value is more than.60, the data are appropriate for factor analysis (Tabachnick & Fidell, 2007). Maximum Likelihood (ML), which is based on the normality assumption and made with continuous indicators, and the direct oblimin rotation technique, which is one of the oblique rotation techniques based on the assumption that the factors are related (Cokluk et al., 2010), were used as factorization techniques. ≥ .50 criterion was determined for item factor loads (Hair et al., 1998). Item evaluation was carried out according to the factor loadings of the items and the common factor variance (h^2) criterion they explained.

In order to verify the scale structure revealed by EFA, CFA was performed with the help of Mplus 7.5 with the ML parameter estimation method, which is used in continuous variables and assumes multivariate normality. It is recommended to use CFI (Comparative Fit Index), TLI (Tucker–Lewis Index), RMSEA (Root Mean Square Error of Approximation), and SRMR (Standardized Root Mean Square Residual) fit criteria to evaluate model fit in CFA (Xu & Tracey, 2017). In addition to these values, Kline (2011) states that the relative chi-square (χ^2/df)

is an important criterion for model fit. In the evaluation of CFA fit indices, CFI and TLI values above .95, RMSEA and SRMR values less than .05, and χ^2/df values less than 2 are perfect fit; CFI and TLI values. .90-.95, RMSEA value of .05-.08, SRMR value between .05-1 and χ^2/df value below 3 indicate acceptable fit (Hu & Bentler, 1999; Kline, 2011). While comparing alternative models that are not nested in CFA, Akaike's information criterion (AIC) was used with the χ^2 difference test. It was accepted that the model with a lower AIC value had a better fit (Barnes & Moon, 2006).

After the scale structure was verified, measurement invariance analysis was conducted to show whether the scale had the same parameter values in different groups. Measurement invariance is a necessary prerequisite for group comparison studies (Vandenberg & Lance, 2000). Therefore, measurement invariance is an important application in the scale development process (Şen, 2020). Failure to ensure measurement invariance may result in erroneous interpretations and results of any group comparisons (Byrne, 2008). In this study, measurement invariance of the scale was tested in terms of categorical variables of gender, marital status, and age.

Each of the measurement invariance, configural, metric, scalar, and strict invariance models are analyzed by comparing them with the previous model and evaluating the change in χ^2. The χ^2 difference test ($\Delta\chi^2$) is used to compare nested models (Brown, 2006; Tabachnick & Fidell, 2007). The non-significant difference for each model is shown as evidence of measurement invariance. However, since the χ^2 test is sensitive to sample size, it is stated that alternative fit values such as ΔCFI and ΔRMSEA can be used for measurement invariance in nested model comparisons. Chen (2007) indicates that the values of ΔCFI≤ -0.010 and ΔRMSEA≤ 0.015 are good cut-off points for the invariance decision for samples greater than 300. In this study, ΔCFI and ΔRMSEA criteria were evaluated together with χ^2 difference tests.

The scale's convergent validity, discriminant validity and reliability were tested using CFA data. Because it calculates the Cronbach Alpha coefficient by equally evaluating the factor load values and error variances of the items, the composite reliability (CR) coefficient gives stronger results than the Cronbach Alpha coefficient (α) in reliability calculations in multidimensional scales (Raykov, 1998). For this reason, the Cronbach Alpha coefficient was tested with the composite reliability coefficient. For convergent validity, it is expected that all CR values for the scale (CR>.70) are greater than AVE (Average Variance Extracted) values, and the AVE value is expected to be greater than .50. For discriminant validity, CR should be >.70, and AVE should be >.50 (CR>AVE), and the square root of the AVE of each construct should be larger than the correlation of the specific construct with any of the other constructs (Fornell & Larcker, 1981). Hair et al. (2014) say that for a scale to be reliable, its Cronbach's alpha internal consistency coefficient and CR value must be above .70. At the last stage, test-retest reliability analysis was performed to test the scale's stability. In the test-retest reliability analysis, the scale's stability depends on the correlation value between the structures measured at different times approaches 1, and the correlation value is significant (Gravesande et al., 2019).

3. RESULT

In this part, findings related to the validity and reliability of the scale have been presented respectively.

3.1. First Stage

In this study, presenteeism, which is examined following the third tradition, is defined as a process that starts with the employee's working and foresees various negative results despite their unwellness. Based on this definition and the literature, a semi-structured interview form consisting of three questions was prepared. The prepared form was submitted to review two experts from the field of educational administration, and necessary corrections were made

according to the feedback. In addition, it was examined by a measurement and evaluation expert to check the form in terms of scientific research logic and a Turkish language expert examined it to determine the points that were not understood. Due to the ambiguous meaning of a question, it was changed and the form was given its final form. It was the way the questions were;

1. Have you continued to work at school in the last month even though you did not feel well (psychologically, physically, or mentally)? (Yes/No).
2. If your answer is yes, what were the factors/reasons that prevented you from feeling well?
3. What were the consequences of continuing to work despite not feeling well?

3.1.1. *Creation of the item pool*

The data collected from the teachers with a semi-structured interview form were analyzed separately by three researchers. It was created by content analysis which can predict the results of the work, despite the factors that make it unwell and unwell at work. The premises and outcomes reached by each researcher have been listed, and the results were compared. Specific premises and outcomes were cocompiled under more general premises and outcomes, and the decisions were tabulated on them (Table 2).

Table 2. *Premises and outcomes of unwellness.*

Premises	*f*	%	Presenteeism Outcomes	*f*	%
Economic uncertainty in the	43	12.1	Distraction/inability to focus	32	16.2
Epidemic diseases	39	11.0	Unproductiveness	28	14.1
Economic problems	35	9.9	Lack of motivation	25	12.6
Authoritarian principal behaviors	27	7.6	Disruption of business	19	9.6
Family problems	27	7.6	Inability to use its capacity	19	9.6
Health problems	24	6.8	Inability to complete tasks	19	9.6
Students' discipline problems	23	6.5	Inability to be energetic	17	8.6
Unfair management style	20	5.6	Unrest	12	6.1
Natural disasters	19	5.4	Business Failure/Fault	8	4
Psychological problems	17	4.8	Inability to give oneself to the lesson	8	4
Stressful environment at school	16	4.5	Forgetfulness	6	3
Time pressure	16	4.5	Inability to pick up what to teach	5	2.5
Incompetent managers	14	3.9			
Excessive workload	13	3.7			
Exclusion	6	1.7			
Hygiene problems	3	0.8			
Political conflicts	2	0.6			
Visual and noise pollution	2	0.6			
Adverse climatic conditions	1	0.3			
Negative changes in legislation	1	0.3			
The absence of teacher career ladders	1	0.3			
Private affairs	1	0.3			
Crowded classes	1	0.3			
Works performed outside of education and training	1	0.3			
Polarizations in school	1	0.3			
Disregard	1	0.3			
Being ignored	1	0.3			

For example, blood pressure, heart disease, rheumatoid arthritis, diabetes, seborrheic dermatitis, etc., health problems; burn-out, depression, obsession, psychological problems; reluctance, demoralization, etc., were grouped under the heading of low motivation. After these results were confirmed by 10 teachers from the same participant group, the item was written. The algorithm prepared in the computer program and the causes and results of being unwell were matched to create the item pool. The algorithm is based on the principle of combining the most emphasized result of the participants stating a certain reason. For example, "focus", which is the most emphasized result of the participants who stated "family problems", was brought together and the scale item "I have no problem focusing on my work at school despite my family problems (reverse item)" was created. In this way, a meaningful 27-item pool was created about presenteeism.

The focus group interview technique was used to analyze the item pool. Focus group interviews are the exchange of views between 4-12 people on the subject of interest under the guidance of a researcher (Marshall, 1999). This technique, built on a discussion strategy on a certain subject, aims to clarify a subject, to clarify, to reveal incomprehensible points, and reach a maximum level of consensus. In this context, focus group interviews were conducted with 5 teachers, each from a different branch, to evaluate the items and to determine whether there were items that were not understood. The items were distributed to these five teachers before the interview, and they were asked to review them. In the interview, which lasted an average of 30 minutes, it was decided to remove seven items that were considered to have the same meaning or had ambiguous expressions from the pool and to make changes in the expressions of four items. The corrected item pool was examined by three educational administrators and an assessment and evaluation expert in terms of content and construct validity and by two Turkish language field experts in terms of semantic validity to get their opinions on content validity and expressions. In the light of expert opinions, 3 more items were removed from the pool, and the expressions (in terms of results) of two items were changed. As a result, a 17-item scale form was created, four of which were reverse items. The scale was graded in a five-point Likert type as "*Always* (5)", "*Mostly* (4)", "*Sometimes* (3)", "*Rarely* (2)" and "*Never* (1)" considering the item statements.

3.2. Second Stage

At this stage, it was aimed to determine the factor structure of the 17-item scale and to make item analyzes.

3.2.1. *Exploratory factor analysis*

In the analysis performed to determine whether the sample is suitable for EFA, the KMO value .87 was found, and the sample adequacy condition was laid down. Bartlett's test of sphericity ($\chi^2 = 3474.11$; $df = 136$; $p = 0.00$) was found to be statistically significant, so it was determined that the data set was suitable for factor analysis. EFA results have been presented in Table 3.

When Table 3 is examined, it was seen that the scale, which was subjected to factor analysis, consisted of three factors with an eigenvalue higher than 1. It is stated that the item can be removed from the scale in cases where the difference between the loads under the two factors is less than .10 (Hair et al., 1998); therefore, two items were removed. The low common factor variance also indicates that the item should be removed from the scale (Kalaycı, 2010), so one item was removed for this reason. The eigenvalues and variances of the factors obtained were 4.56 (32.56), 2.18 (15.59), and 1.56 (11.16), respectively. All three factors together explain 59.31% of the total variance. It is seen that factor loads vary between .64 and .88.

Table 3. *Exploratory factor analysis results for the scale.*

	Factor Loads			
	Factor 1	Factor 2	Factor 3	h^2
I8- The school principal's authoritarian attitudes cause me to make mistakes in school.	.85			.66
I10- I can complete my tasks despite the excessive workload. *	.75			.61
I6- The stressful environment at school affects my professional performance negatively.	.71			.50
I7-The incompetence of school administrators affects my motivation negatively.	.71			.55
I12- The pressure of time I am exposed to by the managers consumes my energy.	.69			.51
I13- Despite the discipline problems of our school students, I can use my full capacity in my lessons. *	.68			.51
I9- Unfair attitudes and behaviors exhibited by school administrators make me feel restless at school.	.64			.51
I1- Despite my economic problems, I concentrate on my work at school. *		.88		.89
I2-Despite my family problems, I have no problem focusing on my work at school. *		.87		.72
I4-My emotional problems hinder my works at school.		.84		.70
I3- Because of my health problems, I cannot show the performance I want while teaching.		.74		.60
I15- I cannot concentrate on my works at school due to the economic uncertainties in the country.			.72	.53
I17- Due to epidemics (COVID-19, Flu, etc.), I cannot be as productive as I would like at school.			.72	.52
I16- Natural disasters and climate change affect my motivation at school negatively.			.70	.50
Eigenvalue	4.56	2.18	1.56	
Total Variance Explained %	32.56	15.59	11.16	

* Reversely coded items

The first factor consists of seven items with loads ranging from .64 to .85; the second factor consists of four items with load values between .74 and .88; the third factor consists of three items that take load values between .70 and .72. The common variance values (h2) being .50 and above have seen as important evidence of the homogeneity of the scale (Çokluk et al., 2021; Thompson, 2004).

Organization-Related Presenteeism (OP), Individual-Related Presenteeism (IP), and Environment-Related Presenteeism (EP) were determined based on the contents of the items collected under the factors and the available literature.

3.3. Third Stage

At this stage, it was aimed to verify the scale structure obtained as a result of EFA and to test the measurement invariance. For this, the data of 303 secondary school teachers were used. Although there are different opinions, Comrey and Lee (1992) expressed that a participant group consisting of 300 people was good in their CFA analysis. Therefore, it can be said that the determined participant group (Table 1) is good according to the specified criteria.

3.3.1. *Confirmatory factor analysis*

CFA was conducted to validate the scale structure that emerged in EFA and to test alternative models. Three-factor, second-order, and one-factor CFA model fit values for the scale are shown in Table 4.

Table 4. *Fit values of models related to factor structure of the PTP scale.*

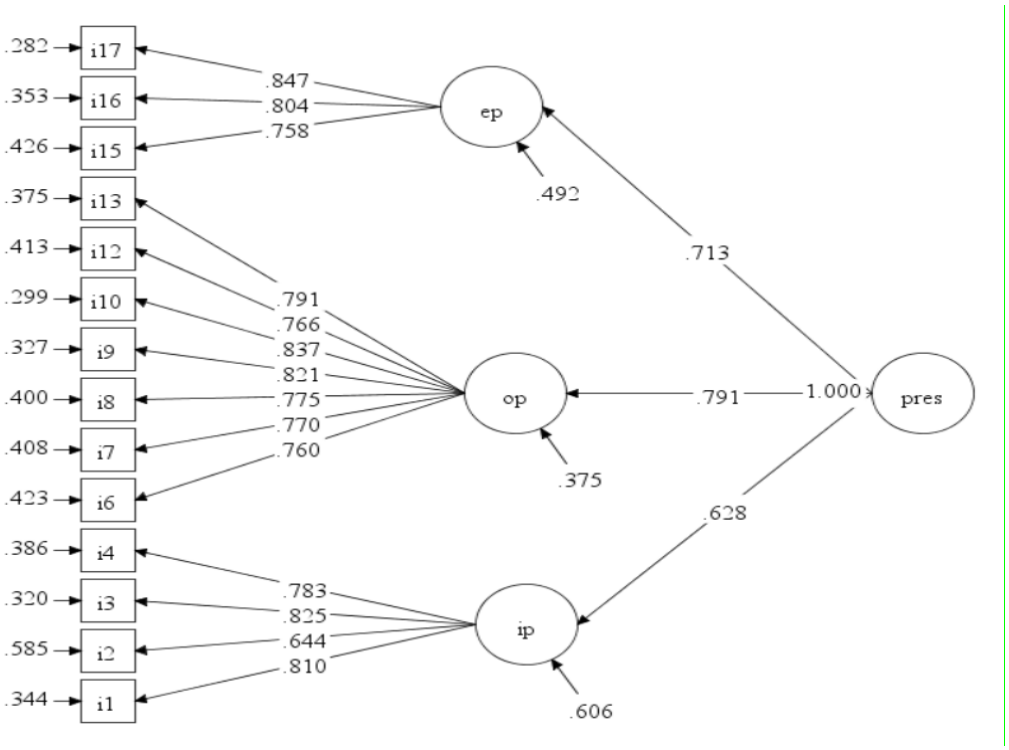
Modeller	χ^2	*df*	χ^2/df	$\Delta\chi^2$	RMSEA	CFI	TLI	SRMR	AIC
The-three factor model	150.69	74	2.03		.058	.968	.961	.032	9318.942
Second order- three factor model	150.69	74	2.03		.058	.968	.961	.032	9318.942
The one-factor model	775.16	77	10.06	624.47*	.173	.713	.660	.114	9937.409
Used model fit indices			≤3		<.05	>.95	>.95	<.05	

CFI: comparative fit index, TLI: Tucker–Lewis index, RMSEA: root mean square error of approximation, SRMR: standardized root mean square Residual

* p<.001

Comparing the three-factor model, second order-three-factor model, and the one-factor model in Table 4, it has been seen that the three-factor model and second-order three-factor models have good fit values. The fit values of the models were detected χ^2 = 150,693 (df = 74; p = .000), RMSEA = .058 (90% CI = .045-.072), CFI = .968, TLI = .961, and SRMR = .032. χ^2/df was also below 3. We also tested the single-factor model, as the indicators showed high correlation values with each other. However, the fit values of the single factor model deteriorated compared to the other models ($\Delta\chi^2$ = 624.47, p <.001, ΔAIC = 555.467), and the model fit values were outside the acceptable limits [(χ^2 = 775.16, df = 77; p = .000), χ^2/df = 10.06, RMSEA = .173 (90% CI = .162-.184), CFI = .713, TLI = .660, and SRMR = .114]. Therefore, it can be said that the three-factor and second-order three-factor models have good fit values. Therefore, it can be said that the construct validity of the model created in EFA and conceptualized theoretically is ensured. The second-order three-factor model obtained from the CFA result of the scale is shown in Figure 2.

Figure 2. *The second order-three factor confirmatory factor analysis model of the scale.*



The second order-three-factor scale structure in Figure 2 was seen to vary the item loads of the "Organization-Related Presenteeism" factor between .760 and .837, the item loads of the "Individual-Related Presenteeism" factor between .644 and .825, and the item loads of the "Environment-Related Presenteeism" factor between .758 and .847.

The CR, AVE, the square root of AVE, and correlations between factors were calculated for the scale's convergent and discriminant validity, whose construct validity was proven by CFA. The results are presented in Table 5.

Table 5. *Fit values of models related to factor structure of the PTP scale.*

Factor	α	AVE	CR	1	2	3
1. IP	.84	.59	.85	**.76***		
2. OP	.91	.62	.92	.50	**.78***	
3. EP	.84	.64	.84	.51	.62	**.80***

α = Cronbach Alpha; AVE = Avarage Variance Extracted; CR = Composite Reliability

* The square root of AVE

Table 5 shows that CR values for all factors are higher than .70, AVE values are higher than .50, and AVE values are lower than CR values. Thus, it can be said that the scale has convergent validity. CR values from AVE values and .70; likewise, the square roots of the AVE values were higher than the correlation values between the factors. These estimations show that the scale has discriminant validity. Therefore, although the scale measures conceptually similar concepts, it has been seen that the measurements are sufficiently different from each other. Both Cronbach Alpha and CR values show that all factors have high reliability.

3.3.2. *Measurement invariance*

The data obtained as a result of the measurement invariance analysis between the categorical variables of the scale's gender (female-male), marital status (married-single), and age (early adulthood-middle adulthood) are presented in Table 6.

Table 6. *Measurement invariance fit indexes (N = 303).*

Models	χ^2	*df*	SRMR	TLI	CFI	RMSEA	$\Delta\chi^2$	Δdf	*p*	ΔCFI	ΔRMSEA
Female (N = 136) Male (N = 167)											
Configural model	238.395	148	.042	.955	.963	.063					
Metric model	248.455	159	.048	.958	.964	.061	10.060	11	.525	.001	-.002
Scalar model	266.275	170	.051	.958	.961	.061	17.820	11	.085	-.003	.000
Strict model	278.210	184	.053	.962	.962	.058	11.935	14	.611	-.001	-.003
Married (N = 193) Single (N = 110)											
Configural model	234.046	148	.040	.957	.965	.062					
Metric model	251.158	159	.050	.957	.962	.062	17.112	11	.104	-.003	.000
Scalar model	259.174	170	.053	.961	.964	.059	8.017	11	.711	.002	-.003
Strict model	268.597	184	.053	.966	.965	.055	9.423	14	.803	.001	-.004
Early adult (N = 157) Mid adult (N = 146)*											
Configural model	238.484	148	.041	.955	.963	.064					
Metric model	254.893	159	.052	.955	.961	.063	16.409	11	.126	-.002	-.001
Scalar model	271.579	170	.055	.956	.958	.063	16.685	11	.117	-.003	.000
Strict model	285.707	184	.062	.959	.958	.060	14.128	14	.440	.000	-.003

* This study was based on Levinson's (1986) age curve. The author includes the age range of 20-40 years in early adulthood; middle adulthood 40-60 years old; defines the age of 60 and above as late adulthood. Participants who are 40 years old are in the early adulthood group; The ones who are over the age of 41 were evaluated in the middle adult group.

When Table 6 is examined; configural model fit indexes according to gender (female-male) variable $\chi^2(148) = 238.395$; RMSEA = .063; CFI = .963; TLI = .955; and SRMR = .042. Metric model fit values $\chi^2(159) = 248{,}455$; RMSEA = .061; CFI = .964; TLI = .958; and SRMR = .048, and ΔCFI and ΔRMSEA values show that the metric invariance conditions were satisfied.

Scalar model fit values $\chi^2(170) = 266.275$; RMSEA = .061; CFI = .961; TLI = .958; and SRMR = .051 and ΔCFI and ΔRMSEA values show that scalar invariance conditions were satisfied. Strict model fit values $\chi^2(184) = 278{,}210$; RMSEA = .058; CFI = .962; TLI = .962; and SRMR = .053, and ΔCFI and ΔRMSEA values show that strict invariance conditions were satisfied. Therefore, the insignificance of the χ^2 difference test and the changes in CFI and RMSEA show that configural, metric, scalar and strict invariances for gender are fully satisfied.

Configural invariance model fit indexes in terms of marital status (married-single) variable $\chi^2(148) = 234.046$; RMSEA = .062; CFI = .965; TLI = .957; and SRMR = .040. Metric model fit values $\chi^2(159) = 251.158$; RMSEA = .062; CFI = .962; TLI = .957; and SRMR = .050 and ΔCFI and ΔRMSEA values show that the metric invariance conditions were satisfied. Scalar model fit values $\chi^2(170) = 259{,}174$; RMSEA = .059; CFI = .964; TLI = .961; and SRMR = .053, and ΔCFI and ΔRMSEA values show that scalar invariance conditions were satisfied. Strict model fit values $\chi^2(184) = 268{,}597$; RMSEA = .055; CFI = .965; TLI = .966; and SRMR = .053, and ΔCFI and ΔRMSEA values show that strict invariance conditions were satisfied. Thus, the insignificance of the χ^2 difference test and the changes in CFI and RMSEA show that the configural, metric, scalar and strict invariances for marital status are fully satisfied.

Configural invariance model fit indexes for age (early adult-mid adulthood) variable $\chi^2(148) = 238{,}484$; RMSEA = .064; CFI = .963; TLI = .955; and SRMR = .041. Metric model fit values $\chi^2(159) = 254.893$; RMSEA = .063; CFI = .961; TLI = .955; and SRMR = .052, and ΔCFI and ΔRMSEA values show that the metric invariance conditions were satisfied. Scalar model fit values $\chi^2(170) = 271.579$; RMSEA = .063; CFI = .958; TLI = .956; and SRMR = .055, and ΔCFI and ΔRMSEA values show that scalar invariance conditions were satisfied. Strict model fit values $\chi^2(184) = 285.707$; RMSEA = .060; CFI = .958; TLI = .959; and SRMR = .062, and ΔCFI and ΔRMSEA values show that strict invariance conditions were satisfied. Therefore, the insignificance of the χ^2 difference test and the changes in CFI and RMSEA show that the age variable's configural, metric, scalar and strict invariances are fully satisfied.

3.4. Fourth Stage

3.4.1. *Test-retest*

Test-retest technique was used to determine the stability of the scale. The scale was applied to the determined participants (Table 7) with an interval of 3 weeks, and the stability of the scale was tried to be estimated by calculating the Pearson correlation (*r*) values over the data set reached. The results of the correlation analysis are shown in Table 7.

Table 7. *Test-retest correlation values of the scale.*

	N	1. P-TPS	IP (1)	OP (1)	EP (1)	2. P-TPS	IP (2)	OP (2)	EP (2)	α
1. P-TPS	109	1								
IP (1)	109	.86*	1							.83
OP (1)	109	.91*	.64*	1						.87
EP (1)	109	.78*	.66*	.55*	1					.82
2. P-TPS	109	**.86***	.70*	.82*	.64*	1				
IP (2)	109	.48*	**.52***	.39*	.36*	.67*	1			.81
OP (2)	109	.75*	.52*	**.81***	.47*	.81*	.33*	1		.82
EP (2)	109	.60*	.53*	.43*	**.72***	.66*	.26*	.40*	1	.80

*$p < .01$; α: Cronbach Alpha

Table 7 shows the results of the correlation analysis. As can be seen in Table 7, the relationship between 1st P-TPS and 2nd P-TPS is .86; the relationship between IP (1) and IP (2) is .52; the correlation between OP (1) and OP (2) was calculated as .81 and between EP (1) and EP (2) as

.72. Test-retest results were found to be significant at the $p<.01$ level in terms of the overall scale and its factors, and it was determined that the stability of the scale was at a sufficient level.

4. DISCUSSION and CONCLUSION

Presenteeism is associated with significant cost losses and has negative effects on both organizations and employees (Abasilim et al., 2015; Baker-McClearn et al., 2010; Bakker & Demerouti, 2007; Cooper & Lu, 2016; D'Abate & Eddy, 2007; Ferreira & Martinez, 2012; Gilbreath & Karimi, 2012; Li et al., 2019) is an important organizational reality. However, contrary to this importance, the absence of a measurement tool in the literature to measure this experience for teachers who experience presenteeism more intensely compared to other occupational groups has been seen as an important deficiency. Based on this deficiency, it is aimed to develop a useful scale with high validity and reliability to measure teachers' perceptions of presenteeism.

Although there are different understandings about presenteeism (Johns, 2010), in this study presenteeism means that the employee works despite being unwell (Gilbreath & Karimi, 2012; Wang et al., 2010), which will prevent him/her from being functional at work (D'Abate & Eddy, 2007) and thus results with negative consequences. In this context, other variables (stress, economic problem, non-work related, bad management, etc.) other than health problems that may be associated with the employee's well-being have been associated with presenteeism.

In the first stage of the development of the P-TPS, the factors associated with the employee's unwellness and the results they can predict were determined with open-ended questions, and scale items were created with the algorithm that had been developed over the obtained data. The scale items applied in the second stage were subjected to exploratory factor analysis. It was determined that the P-TPS had a three-factor structure (individual-related presenteeism, organization-related presenteeism, and environment-related presenteeism) according to the content of the items associated with unwellness. The dimension of “individual-related presenteeism” consists of four items; the “Organization-related presenteeism” factor consists of seven items; The factor of “environment-related presenteeism” consists of three items. The scale explains 59.31% of the total variance. There is no exact value for the minimum variance that a scale should explain, but it is stated that the variance explained by scales with two or more factors should not be less than 50%, especially in social sciences (Liau et al., 2011). There are four reverse items with positive statements, two in the IRP (individual-related presenteeism,) factor and two in the OP factor. These items should be reverse coded when coding the responses on the scale. The highest score that can be obtained on the scale is 70, and the lowest is 14. High scores on the scale and its factors indicate a high perception of presenteeism. In the third stage, as a result of the confirmatory factor analysis of the three-dimensional and 14-item P-TPS, good fit values were estimated, and thus construct validity was ensured. Although there is no consensus in the literature about the fit indices to be considered in determining the model fit in CFA, in addition to the χ^2/df value (Kline, 2011), RMSEA (Steiger, 1990), CFI (Bentler, 1990), TLI (Bentler & Bonett, 1980; Tucker & Lewis, 1973) and SRMR (Byrne, 2008; Hu & Bentler, 1999) fit indices are frequently recommended. For this reason, these fit indices were used in model evaluation in the research. As a result of the analysis of the square root of AVE, CR, and AVE and the correlation coefficients between the factors reached by CFA, it was seen that the scale met the conditions of convergent and discriminant validity. In the reliability analysis of the scale, Cronbach's alpha coefficient and CR values were examined, and if these values are above .80, it shows that sufficient conditions for reliability are met. In addition, measurement invariance analysis (Millsap, 2011), which is used to indicate whether the scale measures the same structure among the groups, was tested over the P-TPS. Considering the insignificance of the χ^2 difference tests and the changes in CFI and RMSEA, it shows that the P-TPS meets the configural, metric, scalar, and strict invariance conditions

regarding gender, age, and marital status variables. Therefore, it has been revealed that the PTP scale can measure the same structure among groups that differ in terms of these variables. In this sense, it can be said that the scale can be used to compare the perceptions of presenteeism among different groups.

Finally, the stability of the scale, which provided internal consistency with Cronbach's alpha and combined reliability conditions with CR values, was tested with the test-retest method, correlation values were examined in terms of the overall scale and the factors, and significant values were estimated. Therefore, it is possible to say that the scale has a stable structure, and consistent results can be achieved when applied at different times. When the validity and reliability proofs of the scale are evaluated together, it can be stated that the scale can be used safely to determine teachers' perceptions of presenteeism.

4.1. Limitations and Recommendations

This study analyzed the validity, reliability and measurement invariance of the three-dimensional and 14-item P-TPS. Therefore, it can be used as an effective measurement tool for in-depth analysis in future empirical, relational, and descriptive research on presenteeism. However, there are some limitations to this study. First, the scale was developed with presenteeism, approaching from a specific perspective (3rd tradition). In this sense, the scale may need to be adjusted according to other perspectives. Secondly, the sample was selected from among the teachers working in Turkey in 2021, and an item pool was created according to the answers given by these teachers to open-ended questions. Therefore, the content of the scale reflects the realities of the time the answers were collected (COVID- 19, economic problems, etc.) and the professional and organizational characteristics of the teachers. Some revisions may be necessary for it to be used in another period and other professional fields. Thirdly, the semi-structured interviews for the creation of the item pool during the COVID-19 pandemic process and the focus interviews applied to evaluate the item pool have been minimized as much as possible. Therefore, overlooked, some important facts can be found. Finally, the second-order three-factor CFA results confirmed the structure of the scale. The scale scores show the teacher's presenteeism. For this reason, future researchers who will use the scale should be careful to use the second-order three-dimensional structure of this scale.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number:** E-11611387-100-38896

Authorship Contribution Statement

Alper Uslukaya: Collected data, Investigation, Resources, Visualization, Software, Formal Analysis and Writing-original Draft. **Zulfu Demirtas**: Methodology, Supervision, and Validation. **Muslim Alanoglu:** Software, Formal Analysis, Writing-original Draft, Supervision, Validation.

Orcid

Alper Uslukaya https://orcid.org/0000-0003-1455-8438
Zulfu Demirtas https://orcid.org/0000-0002-1072-5772
Muslim Alanoglu https://orcid.org/0000-0003-1828-4593

REFERENCES

Abasilim, U.D., Salau, O.P., & Falola, O.H. (2015). Toward an understanding of presenteeism and its effects in the workplace. *Journal of Public Administration & Management, 1*(1), 74-85. http://eprints.covenantuniversity.edu.ng/7380/

Ammendolia, C., Côté, P., Cancelliere, C., Cassidy, D., Hartvigsen, J, Boyle, E., Soklaridis, S., Stern, P., & Amick, B. (2016). Healthy and productive workers: Using intervention mapping to design a workplace health promotion and wellness program to improve presenteeism. *BMC Public Health, 16*(1), 1190. https://doi.org/10.1186/s12889-016-3843-x

Aronsson, G., & Gustafsson, K. (2005). Sickness presenteeism: prevalence, attendance-pressure factors, and an outline of a model for research. *Journal of Occupational and Environmental Medicine*, *47*(1), 958-966. https://doi.org/10.1097/01.jom.0000177219.75677.17

Aronsson, G., Gustafsson, K., & Dallner, M. (2000). Sick but yet at work: An empirical study of sickness presenteeism. *Journal of Epidemiology Community Health*, *54*(1), 502-509. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1731716/

Baker-McClearn, D., Greasley, K., Dale, J., & Griffith, F. (20l0). Absent management and presenteeism: The pressures on employees to attend work and impact of attendance on performance. *Human Resource Management Journal, 20*(3), 311-328. https://doi.org/10.1111/j.1748-8583.2009.00118.x

Bakker, A., & Demerouti, E. (2007). The job demands-resources model: State of the art. *Journal of Managerial Psychology, 22*(3), 309-328. https://doi.org/10.1108/02683940710733115

Barnes, K.L., & Moon, S.M. (2006). Factor structure of the psychotherapy supervisor development scale. *Measurement and Evaluation in Counseling and Development*, *39*(3), 130-140. https://doi.org/10.1080/07481756.2006.11909794

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(1), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bentler, P.M., & Bonett, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(1), 588-606. https://doi.org/10.1037/0033-2909.88.3.588

Berger, M.L., Howell, R., Nicholson S., & Sharda, C. (2003). Investing in healty human capital. *Journal of Occupational and Environmental Medicine*, *45*(12), 1213-1225. https://doi.org/10.1097/01.jom.000102503.33729.88

Bergström, G., Bodin, L., Hagberg, J., Lindh, G., Aronsson, G., & Josephson, M. (2009). Sickness presenteeism today, sickness absenteeism tomorrow? A prospective study on sickness presenteeism and future sickness absenteeism. *Journal of Occupational and Environmental Medicine, 51*(6), 629-638. https://doi.org/10.1097/JOM.0b013e3181a8281b

Beswick, D.M., Mace, J.C., Rudmik, L., Soler, Z.M., DeConde, A.S., & Smith, T.L. (2018). Productivity changes following medical and surgical treatment of chronic rhinosinusitis by symptom domain. *International Forum of Allergy & Rhinology*, *8*(12), 1395-1405. https://doi.org/10.1002/alr.22191

Borrill, C., West, M.A., Shapiro, D., & Rees, A. (2000). Team working and effectiveness in health care. *British Journal of Health Care Management, 6*(8), 364-371. https://doi.org/10.12968/bjhc.2000.6.8.19300

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

Burton, W.N., Pransky, G., Conti, D.J., Chen, C.Y., & Edington, D.W. (2004). The association of medical conditions and presenteeism. *Journal of Occupational and Environmental Medicine*, 38-45. https://doi.org/10.1097/01.jom.0000126687.49652.44

Byrne, B.M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema*, *20*(4), 872-882. https://pubmed.ncbi.nlm.nih.gov/18940097/

Caverley, N., Cunningham, J.B., & MacGregor, J.N. (2007). Sickness presenteeism, sickness absenteeism, and health following restructuring in a public service organization. *Journal of management studies, 44*(2), 304-319. https://doi.org/10.1111/j.1467-6486.2007.00690.x

Chen, F.F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*(1), 464–504. https://doi.org/10.1080/ 10705510701301834

Chen, J.W., Lu, L., & Cooper, C.L. (2021). The compensatory protective effects of social support at work in presenteeism during the coronavirus disease pandemic. *Frontiers in Psychology*, *12*(1), 689. https://doi.org/10.3389/fpsyg.2021.643437

Comrey, A.L., & Lee, H.L. (1992). *A first course in factor analysis.* Erlbaum.

Conner, T.S., & Silvia, P.J. (2015). Creative days: A daily diary study of emotion, personality, and everyday creativity. *Psychology Aesthetics Creat Arts, 9*(1), 463-470. https://doi.org /10.1037/aca0000022

Cooper, C.L., & Lu, L. (2016). Presenteeism as a global phenomenon: Unraveling the psychosocial mechanisms from the perspective of social cognitive theory. *Cross Cultural and Strategic Management, 23*(2), 216-231. https://doi.org/10.1108/ccsm-09-2015-0106

Cooper, C., & Dewe, P. (2008). Well-being – absenteeism, presenteeism, costs and challenges. *Occupational Medicine (London), 58*(8), 522–4. https://doi.org/10.1093/occmed/kqn124

Creswell, J.W., & Plano Clark, V.L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Sage Publications.

Çokluk, O., Şekercioğlu, G., & Büyüköztürk, Ş. (2021). *Sosyal bilimler için çok değişkenli istatistik, SPSS ve LISREL uygulamları (6nd ed.) [Multivariate statistics SPSS and LISREL applications for social sciences]*. PegemA.

D'Abate, C.P., & Eddy, E.R. (2007). Engaging in personal business on the job: Extending the presenteeism construct. *Human Resource Development Quarterly,18*(3), 361-384. https://doi.org/10.1002/hrdq

Demerouti, E., Le Blanc, P.M., Bakker, A.B., Schaufeli, W.B., & Hox, J. (2009). Present but sick: A three-wave study on job demands, presenteeism and burnout. *Career Development International, 14*(1), 50–68. https://doi.org/10.1108/13620430910933574

DeVellis, R.F. (2003). *Scale development: Theory and applications* (2nd ed.). Sage Publications.

Dew, K., Keefe, V., & Small, K. (2005). Choosing' to work when sick: Workplace presenteeism. *Social Science & Medicine*, *60*(10), 2273-2282. https://doi.org/10.1016/j.socscimed.2004.10.022

Evans, C.J. (2004). Health and work productivity assessment: State of the art or state of flux? *Journal of Occupational and Environmental Medicine*, *46*(1), 3-11. https://doi.org/10.1097/01.jom.0000126682.37083.fa

Ferreira, A.I., & Maritnez, L.F. (2012). Presenteeism and burnout among teachers in public and private Portuguese elementary schools. *The International Journal of Human Resource Management, 23*(20), 4380-4390. https://doi.org/10.1080/09585192.2012.667435

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://www.jstor.org/stable/3151312?seq=1

Gilbreath, B., & Karimi, L. (2012). Supervisor behavior and employee presenteeism. *International Journal of Leadership Studies*, *7*(1),114-131. https://www.researchgate.net/publication/236946893_Supervisor_Behavior_and_Employee_Presenteeism

Goetzel, R.Z., Long, S.R., Ozminkowski, R.J., Hawkins, K., Wang, S., & Lynch, W. (2004). Health, absence, disability and presenteeism cost estimates of certain physical and mental

health conditions affecting US employers. *Journal of Occupational and Environmental Medicine, 46*(4), 398-412. https://doi.org/10.1097/01.jom.0000121151.40413.bd

Goldstein, I., Goren, A., Li, V.W., Maculaitis, M.C., Tang, W.Y., & Hassan, T.A. (2019). The association of erectile dysfunction with productivity and absenteeism in eight countries globally. *International Journal of Clinical Practice*, *73*(11), e13384. https://doi.org/10.1111/ijcp.13384

Grant, A.M. (2008). The significance of task significance: job performance effects, relational mechanisms, and boundary conditions. *Journal of Applied Psychology,* 93(1), 108-124.

Gravesande, J., Richardson, J., Griffith, L., & Scott, F. (2019). Test-retest reliability, internal consistency, construct validity and factor structure of a falls risk perception questionnaire in older adults with type 2 diabetes mellitus: A prospective cohort study. *Archives of Physiotherapy, 9*(1), 1-11. https://doi.org/10.1186/s40945-019-0065-4

Hair, J.D., Anderson, R.E., Tatham, R.L., & Black, W.C. (1998). *Multivariate data analysis* (5th ed.). Prentice Hall.

Hair, J.F., Black, W.C., Babin, B.J., & Anderson, R.E. (2014). *Multivariate data analysis* (7nd ed.). Pearson Education.

Hansen, C.D., & Andersen, J.H. (2008). Going ill to work–What personal circumstances, attitudes and work-related factors are associated with sickness presenteeism?. *Social Science & Medicine*, *67*(6), 956-964. https://doi.org/10.1016/j.socscimed.2008.05.022

Hemp, P. (2004). Presenteeism: At work-but out of it. *Harvard Business Review*, *82*, 49–58.

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

Jennings, P.A., & Greenberg, M.T. (2009). The prosocial classroom: Teacher social and emotional competence in relation to student and classroom outcomes. *Review of Educational Research*, *79*(1), 491-525. https://doi.org/10.3102/0034654308325693

Johansson, G., & Lundberg, I. (2004). Adjustment latitude and attendance requirements as determinants of sickness absence or attendance. Empirical tests of the illness flexibility model. *Social Science and Medicine*, *58*(10), 1857–1868. https://doi.org/10.1016/S0277-9536(03)00407-6

Johns, G. (2010). Presenteeism in the workplace: A review and research agenda. *Journal of Organizational Behavior, 31*(1)*,* 519–542. https://doi.org/10.1002/job.630

Johns, G. (2011). Attendance dynamics at work: The antecedents and correlates of presenteeism, absenteeism, and productivity loss. *Journal of Occupational Health Psychology, 16*(4), 483-500. https://doi.org/10.1037/a0025153

Kalaycı, Ş. (2010). Factor analysis. In Ş. Kalaycı, (Ed.), *SPSS applied multivariate statistical techniques* (pp. 234-255). Asil Publication.

Karanika-Murray, M., & Biron, C. (2019). The health-performance framework of presenteeism: Towards understanding an adaptive behaviour. *Human Relations,* 1-20. https://doi.org/10.1177/0018726719827081

Kidger, J., Brockman, R., Tilling, K., Campbell, R., Ford, T., Araya, R., King, M., & Gunnell, D. (2016). Teachers' wellbeing and depressive symptoms, and associated risk factors: A large cross sectional study in English secondary schools. *Journal of Affective Disorders*, *192*(1), 76-82. https://doi.org/10.1016/j.jad.2015.11.054

Kiefer, R.A. (2008). An integrative review of the concept of well-being. *Holistic Nursing Practice*, *22*(5), 244-252. https://doi.org/10.1097/01.HNP.0000334915.16186.b2

Kivimäki, M., Head, J., Ferrie, J.E., Hemingway, H., Shipley, M.J., Vahtera, J., & Marmot, M.G. (2005). Working while ill as a risk factor for serious coronary events: The Whitehall II study. *American Journal of Public Health, 95*(1), 98-102. https://doi.org/10.2105/AJPH.2003.035873.

Kline, R.B. (2011). *Principles and practice of structural equation modeling*. The Guilford Press.

Koopman, C., Kenneth R.P., James, F.M., Claire, E.S., Marc L.B., Robin, S.T., Paul, H., Pamela, G., Danielle, M.H., & Talor, B. (2002). Stanford presenteeism scale: Health status and employee productivity. *Journal of Occupational and Environmental Medicine*, *44*(1), 1- 12. https://doi.org/10.1097/00043764-200201000-00004

Levinson, D.J. (1986). A conception of adult development. *American Psychologist*, *41*(1), 3-13. https://doi.org/10.1037/0003-066X.41.1.3.

Li, Y.X., Zhang, J.H., Wang, S.N., & Guo, S.J. (2019). The effect of presenteeism on productivity loss in nurses: the mediation of health and the moderation of general self-efficacy. *Frontiers in Psychology, 10*(1), 1745. https://doi.org/10.3389/fpsyg.2019.01745

Liau, A., Tan, T.K., & Khoo, A. (2011). Scale measurement: Comparing factor analysis and variable clustering. *SAS Global Forum 2011*, Nevada, Las Vegas. https://www.researchgate.net/publication/264496804_Scale_Measurement_Comparing_Factor_Analysis_and_Variable_Clustering

Lohaus, D., & Habermann, W. (2019). Presenteeism: a review and research directions. *Human Resource Management Review, 29*(1), 43–58.https://doi.org/10.1016/j.hrmr.2018.02.010

Lu, L., Lin, H.Y., & Cooper, C.L. (2013). Unhealthy and present: motives and consequences of the act of presenteeism among Taiwanese employees. *Journal of Occupational Health Psychology*, *18*(4), 406-416. https://doi.org/10.1037/a0034331

Marshall, G. (1999). *Sociology dictionary.* Science and Art Publications

McGregor, A., Magee, C.A., Caputi, P., & Iverson, D. (2016). A job demands-resources approach to presenteeism. *Career Development International, 21(*4), 402-418. https://doi.org/10.1108/CDI-01-2016-0002

Millsap, R.E. (2011). *Statistical approaches to measurement invariance*. Routledge.

Miraglia, M., & Johns, G. (2016). Going to Work III: A meta-analysis of the correlates of presenteeism and a dual-path model. *Journal of Occupational Health Psychology. 21*(3), 261. https://doi.org/10.1037/ocp0000015

Myers, J.E., & Williard, K. (2003). Integrating spirituality into counselor preparation: A developmental approach. *Counseling and Values*, *47*(1), 142-155. https://doi.org/10.1002/j.2161-007x.2003.tb00231.x

Perez-Nebra, A.R.P., Queiroga, F., & Oliveira, T.A. (2020). Presenteeism of class teachers: Well-being as a critical psychological state in the mediation of job characteristics. *Revista de Administração Mackenzie, 21*(1), 1-26. https://doi.org/10.1590/1678 6971/eRAMD200123

Raykov, T. (1998). A method for obtaining standard errors and confidence intervals of composite reliability for congeneric items. *Applied Psychological Measurement*, *22*(4), 369-374. https://doi.org/10.1177/014662169802200406

Roe, R. (2003). Health and performance. In W. B. Schaufeli, A.B. Bakker, & J. De Jonge (Eds.), *Psychology of Work and Health* (In German) (pp. 375-388). Bohn Stafleu Van Loghum. https://link.springer.com/chapter/10.1007/978-90-313-6556-2_19

Seçer, İ. (2015). *SPSS ve LISREL ile pratik veri analizi* (2nd ed.) *[Practical data analysis with SPSS and LISREL]*. Anı Publication.

Sisask, M., Värnik, P., Värnik, A., Apter, A., Balazs, J., Balint, M., Bobes, J., Brunner, R., Corcoran, P., Cosman, D., Feldman, D., Haring, C., Kahn, J-P, Poštuvan, V., Tubiana, A., Sarchiapone, M., Wasserman, C., Carli, V., Hoven, C.W., & Wasserman, D. (2014). Teacher satisfaction with school and psychological well-being affects their readiness to help children with mental health problems. *Health Education Journal*, *73*(4), 382–393. https://doi.org/10.1177/0017896913485742

Steiger, J.H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*(1), 173-180.

Şen, S. (2020). *Mplus ile yapısal eşitlik modellemesi uygulamaları [Structural equation modeling applications with Mplus]*. Nobel Publishing

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

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. American Psychological Association

Tucker, L.R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*(1), 1-10. https://link.springer.com/article/10.1007/BF02291170

Turpin, R.S., Ozminkowski, R.J., Sharda, C.E., Claire, E., Collins, J.J., Berger, M.L., Billotti, G.M., Baase, C.M., Olson, M.J., & Nicholson, S. (2004). Reliability and validity of the Stanford Presenteeism Scale. *Journal of Occupational and Environmental Medicine, 46*(1),1123–1133. https://doi.org/10.1097/01.jom.0000144999.35675.a0

Vandenberg, R.J., & Lance, C.E. (2000). A review and synthesis of the measuremant invarience literature: Suggestion, practices and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-70. https://doi.org/10.1177/109442810031002

Vera-Calzaretta A., & Juarez-Garcia A. (2014). Presenteeism. In A.C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research (pp.* 5040-5046*)*. Springer. https://doi.org/10.1007/978-94-007-0753-5_2254

Wang, J., Schmitz, N., Smailes, E., Sareen, J., & Patten, S.M.D. (2010). Workplace characteristics, depression, and health-related presenteeism in a general population sample. *Journal of Occupational and Environmental Medicine*, *52*(8), 836-842. https://doi.org/10.1097/JOM.0b013e3181ed3d80

Widera, E., Chang, A., & Chen, H.L. (2010). Presenteeism: A public health hazard. *Journal of General Internal Medicine, 25*(11), 1244-1247. https://link.springer.com/article/10.1007/s11606-010-1422-x

Witmer, J.M., & Sweeney, T.J. (1992). A holistic model for wellness and prevention over life span. *Journal of Counseling and Development*, *71*(2), 140-148. https://doi.org/10.1002/j.1556-6676.1992.tb02189.x

Wrate, R.M. (1999). Increase in staff numbers may reduce doctors' presenteeism. *British Medical Journal*, 319(1), 1502. https://doi.org/10.1136/bmj.319.7223.1502a

Xu, H., & Tracey, T.J. (2017). Use of multi-group confirmatory factor analysis in examining measurement invariance in counseling psychology research. *The European Journal of Counselling Psychology, 6*(1), 75-82. https://doi.org/10.5964/ejcop.v6i1.120

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 4, 867–882
https://doi.org/10.21449/ijate.1208809
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

An investigation of data mining classification methods in classifying students according to 2018 PISA reading scores

Emrah Buyukatak[1,*], **Duygu Anil**[2]

[1]Independent Researcher
[2]Hacettepe University, Faculty of Education, Department of Educational Sciences, Ankara, Turkey

ARTICLE HISTORY

Received: June 10, 2022
Revised: Sep. 13, 2022
Accepted: Nov. 22, 2022

Keywords:
Data Mining,
Artificial Neural Networks,
Decision Trees,
Cluster Analysis,
Classification.

Abstract: The purpose of this research was to determine classification accuracy of the factors affecting the success of students' reading skills based on PISA 2018 data by using Artificial Neural Networks, Decision Trees, K-Nearest Neighbor, and Naive Bayes data mining classification methods and to examine the general characteristics of success groups. In the research, 6890 student surveys of PISA 2018 were used. Firstly, missing data were examined and completed. Secondly, 24 index variables thought to affect the success of students' reading skills were determined by examining the related literature, PISA 2018 Technical Report, and PISA 2018 data. Thirdly, considering the sub-classification problem, the students were scaled in two categories as "Successful" and "Unsuccessful" according to the scores of PISA 2018 reading skills achievement test. Statistical analysis was conducted with SPSS MODELER program. At the end of the research, it was determined that Decision Trees C5.0 algorithm had the highest classification rate with 89.6%, the QUEST algorithm had the lowest classification rate with 75%, and four clusters were obtained proportionally close to each other in Two-Step Clustering analysis method to examine the general characteristics according to the success scores. It can be said that the data sets are suitable for clustering since the Silhouette Coefficient, which is calculated as 0.1 in clustering analyses, is greater than 0. It can be concluded that according to achievement scores, all data mining methods can be used to classify students since these models make accurate classification beyond chance.

1. INTRODUCTION

One of the most important criteria for the success of educational policies of countries is to be able to train qualified and successful individuals in accordance with the information and data era. Success is determined by evaluating the performances at the national and international levels whether the planned targets in the education systems have been achieved in the recent period. Today for this purpose, education systems are evaluated by using large-scale exams which are applied to large groups covering the specified knowledge and skills for more than one course to monitor what students learn in the school environment. In addition, the learning skills of students of a certain age and school group in different countries are regularly monitored and compared.

*CONTACT: Emrah Buyukatak ✉ ebuyukatak@hotmail.com

e-ISSN: 2148-7456 /© **IJATE** 2022

In large-scale exams, it has become important to use open-ended questions or open-ended and multiple-choice questions together, which allows measuring high-level cognitive skills and allows students to give their own answers since open-ended questions give students the opportunity to think and create their own answers.

PISA measures students' high-level cognitive skills by investigating not only whether the basic knowledge learned at school is re-used, but also whether students can guess about what they do not know using knowledge that they have learned and whether they can apply what they know inside and outside of school. In PISA, not only knowledge and skills in Turkish, mathematics and science, but also attitudes towards Turkish, mathematics, and science are discussed, and also whether they are aware of what opportunities the scientific competencies they gain at school will create for them is evaluated (Anıl, 2008). Large-scale achievement tests such as PISA are achievement tests that mostly consist of multiple subtests or dimensions in different grade levels and courses. PISA is applied to large student groups and a huge amount of data are obtained from this exam.

PISA is carried out regularly and information on many variables is collected. Since there is a large amount of information about students in such a large-scale exam, data in this application are also defined as big data. This information in different formats, which emerges from both test scores and questionnaires and is also obtained from more than half a million students, constitutes a large pile of data. The important thing here is to determine which is meaningful and which is meaningless from such a large amount of data in PISA in the decision process. As a result, decisions can be made as to whether this data can be used in data mining since large amount of information obtained from students in recent years is big data (Nisbet, Elder, & Miner, 2009). With these methods, behavioral patterns of individuals are analyzed and predictions are made for future behaviors.

The amount of information produced and stored at the global level is unimaginably large and on the increase every day. However, data in these areas should be stored and managed securely in a magnetic environment using database systems. As a result of such needs, powerful systems and tools are needed to systematically reveal efficient information from large amounts of data and to transform them into organized data and then knowledge. Data mining emerged in the 1980s when computers began to be used to solve data analysis problems. Data mining is called an interdisciplinary field of study that combines various techniques such as machine learning, pattern recognition, statistics, databases, and visualization to solve the problem of obtaining information from large data sets (Cabena et al., 1998). Data mining is also expressed as the process of applying one or more computer learning techniques to automatically extract and analyze information from the data in the database (Roiger, 2017). In addition, this process is the use of multiple data analysis tools to reveal patterns in the data and the relationship between the data in order to make valid predictions. In this direction, data mining techniques make it possible to reveal the relationship between the parameters of large amounts of data in largescale exams such as PISA and TIMMS.

Data such as the most important element of the education process, students' personal information, grade status, absenteeism, and successful and unsuccessful courses are obtained by Educational Data Mining (EDM), which examines data mining in terms of education. By applying different models to these data, it is possible to determine the reasons for success, to increase their success, to prevent their absenteeism, to choose the courses they will take, and to make recommendations regarding their career goals (Rizvi, Rienties, & Khoja, 2019). In this way, the discovery of patterns based on these data and the use of discovered patterns in the improvement of the learning process and in instructional design have emerged as important issues today. By this means, data mining techniques are used in education in forming groups according to students' personal characteristics and individual learning similarities, predicting

undesirable student behaviors such as low motivation, absenteeism, dropping out of school, and not following school rules, and taking necessary precautions (Aksoy, 2014).

Educational Data Mining (EDM) is the creation, research, and application of analysis methods in digital environments to detect patterns in multi-volume educational data, which is very difficult to analyze due to large data (Romero & Ventura, 2013). Data in EDM are not limited to interactions of students, and data from students, administrative data, demographic data, and emotional characteristics of trainees together constitute the EDM data (Witten & Frank, 2000). To make determinations about student success, to make inferences about failure in the education environment and its causes, and to create educational environments that meet the needs, educational data mining, which uses many different disciplines such as psychometry, learning analytics, and statistics can be benefitted (Özbay, 2015).

Nowadays, it has been thought that data mining will be useful especially in the selection and classification made taking into account the measurement results in the field of education. In this way, it will be possible to understand the learning level and behavior of students better by determining which variable may be effective in which cluster or class. As a result, the number of prediction studies conducted to determine the factors affecting student success and the shaping of this success has increased significantly (Anıl, 2008; Gelbal, 2008; Erdil, 2010; Özer & Anıl, 2011). In addition, it is very difficult to make prediction and classifications in groups that are similar to each other, which makes it necessary to carefully select the methods used in research and ensure the classification with the most accurate prediction.

When the related studies are evaluated as a whole, data mining methods can be seen to have been used intensively on a sectoral basis, especially in industry and banking. Although such methods offer a wide field of study in the field of education and the number of studies in education related to the concept of data mining has increased nationally and internationally, it is observed that very few studies have been carried out and specifically domestic studies and resources have been scarce. However, using the data collected in education is of central importance for achieving success and increasing student achievement in this field. As a result of collecting more data in the field of education along with technological developments, this research is important in terms of examining data mining methods in educational fields other than the usual sectoral basis.

Different methods and algorithms have been used in the literature as to recent prediction and classification studies in data mining, and the models used in these applications have a unique algorithm. Evaluating the algorithms by comparison or revealing which algorithm is successful in situations is important in terms of increasing classification performances. Research in data mining has generally been limited to Artificial Neural Networks and Decision Trees. However, this study is important in terms of using and comparing Artificial Neural Networks, Decision Trees, K-Nearest Neighbor, and Naive Bayes methods for classification models that will allow predicting the future success of students. In this study, apart from the most used classification methods, other data mining classification methods that are considered to make significant contributions to the literature are examined. In addition, using the Two-Step Clustering analysis, different groups gathered in the same cluster according to the similar characteristics of the students' data in the large-volume PISA 2018 data set were determined and the importance of the variables on these groups was examined.

The purpose of this research was to determine classification accuracy of the factors affecting the success of students' reading skills and the success scores of reading skills, based on PISA 2018 data by using data mining classification methods such as Artificial Neural Networks, Decision Trees, K-Nearest Neighborhood, and Naive Bayes and to examine the general characteristics of success groups. For this purpose, the following sub-problems were examined in this study.

1.Considering the factors affecting the students' success in 2018 PISA reading skills and their success scores, at what accuracy rate do Artificial Neural Networks, Decision Trees, K-Nearest Neighborhood, and Naive Bayes analyses classify students according to their success?

2.What are the general characteristics of the achievement groups according to the factors affecting the 2018 PISA reading skills success of the students and their success scores in reading skills?

3.What are the results regarding the comparison of the general classification rates of the students of Artificial Neural Networks, Decision Trees, K-Nearest Neighborhood, and Naive Bayes methods according to their success?

2. METHOD

This study was conducted to examine different classification models. In this respect, the research is a descriptive study.

2.1. Study Group

186 schools and 6890 students represented Turkey in the PISA 2018 application. Since the items that were mostly not answered or not entered any responses in the study were excluded from the data set, the sample of the study consisted of 6431 students.

2.2. Data Collection Tools

Each PISA application focuses on one of the fields of mathematics, reading, and science. PISA 2018 focused on predominantly reading skills and also mathematical literacy and science literacy.

PISA 2018 included cognitive tests aiming to measure the academic performance of students and questionnaires of student and school were prepared to evaluate the student as a whole. Students were expected to answer the questionnaire, which consisted of questions about oneself, family and home, language learning at school, the Turkish / Turkish Language and Literature Lesson learned at school, thoughts about life, school, school program, and learning periods. The main student questionnaire, computer-based, consisted of 79 questions and lasted 35 minutes. The data used in the study consist of student questionnaire and cognitive test results and were downloaded from the OECD website.

2.3. Data Analysis

In the literature there are not any assumptions that need to be tested before these techniques can be applied. However, missing data analysis was done by considering the mechanisms of missing data patterns and amounts. As a result of the examination carried out before the analysis in the study, 459 data were removed from the data set due to responses either mostly not answered or not entered at all, and the analysis was carried out with 6431 data. In addition, after the missing data analysis, it was determined that 1678 data were missing. Since the exclusion of the data of 1678 students from the analysis would not give correct results, missing data were completed with the EM logarithm.

In the study, importance was given to the selection of variables that affect the success of students' reading skills in the selection of variables based on PISA 2018 data. Within the scope of variable selection, literature, PISA 2018 Technical Report, and PISA 2018 data were examined. As a result, 24 indices that were considered to affect the success of students' reading skills were determined in this study. The variables used in the study were "Index of Economic, Social and Cultural Status (ESCS)", "Family Wealth (WEALTH)", "Understanding and Remembering (UNDREM)", "Summarizing (METASUM)", "Reading and Using Strategies (METASPAM)", "Joy/Like Reading (JOYREAD)", "Disciplinary Climate (DISCLIMA)", "Home Educational Resources (HEDRES)", "Home Possessions (HOMEPOS)", "Information

and Communication Technologies Resources (ICTRES)", "Cultural Possessions at Home (CULTPOSS)", "Teacher Support (TEACHSUP)", "Teacher's Stimulation of Reading Engagement Perceived by Student (STIMREAD)", "Self-Concept of Reading: Perception of Competence (SCREADCOMP)", "Self-Concept of Reading: Perception of Difficulty (SCREADDIFF)", "Perception of Difficulty of the PISA Test (PISADIFF)", "Parents' Emotional Support Perceived by Student (EMOSUPS)", "Perceived Feedback (PERFEED)", "Subjective Well-Being: Positive Affect (SWBP)", "Perception of Cooperation at School (PERCOOP)", "Subjective Well-Being: Sense of Belonging to School (BELONG)", "Use of ICT in Leisure Activities out of School (ENTUSE)", "Use of ICT for School Work Outside of School (HOMESCH)" and "Use of ICT at School (USESCH)".

Students were scaled in two categories as "Successful-Unsuccessful" according to the scores in PISA 2018 reading skills achievement test. First, "average reading achievement score" variable was formed by taking the average mean of the 10 reading achievement scores (Plausible Value: PV1READ, PV2READ … PV10READ) of every student. Then the mean of this variable was calculated as 470. If any students' "average reading success score" is below 469.9, it is called "unsuccessful-0", and if it is above 469.9, it is called "successful-1". The "success status" variable was created in such a way. In the light of these regulations, the number of "successful1" students was 3212 with 49.9%. The number of "unsuccessful-0" students was 3219 with 50.1% in the PISA 2018 Turkey application, in which 6431 students participated.

During the model evaluation process, both Cross Validation and Bootstrap methods were used to ensure that many models were created and tested. In order to increase the accuracy of the methods and algorithms, the analysis was run with Boosting, and the 10-fold Cross Validation technique was used in the development of the models. Before the analysis, the data set was divided into 70% training and 30% test data. In the literature, some studies split the data into three parts as training, test, and validation, while some research splits the data into training and test sets. In this study, data was split into two parts; namely, data as training and test set because in the study Cross Validation method was used to ensure that many models were created and tested. When using a method such as cross validation, two partitions may be sufficient and effective, thereby averaged after repeated rounds of model training and testing to help reduce bias and variability (Xu & Goodacre, 2018). The seed value of analysis to reproducibility is 2695748.

In the study, "success status" variable is a dependent variable and 24 index variables are independent ones. Artificial Neural Networks, Decision Tree algorithms, K-Nearest Neighborhood, and Naive Bayes analyzes were made using SPSS Modeler 18.0 program. As a result of the analyses, the correct classification rates of each model and algorithm's training and test data were calculated. The overall correct classification rate of all data set was calculated using the following equation:

$$\text{The Overall Correct Classification Rate} = \frac{\text{Number Of Correctly Classified Samples}}{\text{Total Number Of Samples}}$$

To test the accuracy of the classification of a model, the relative and maximum chance criteria need to be calculated and compared. According to the success status of the sample, the maximum chance criterion of "successful" and "unsuccessful" students is 0.51 The relative chance criterion is 0.49 In this study, the percentages of classification accuracy determined were evaluated by comparing them with the maximum and relative chance criteria.

In this specific research, clustering analysis was performed in order to group the ungrouped data according to their similarities. Two-Step Clustering algorithm is preferred for large and high-dimensional data consisting of both categorical and continuous data. In the study,

Two-Step Clustering Analysis was carried out through the SPSS program in order to determine the different groups by collecting the data of the students in the same cluster according to their similar characteristics (variables) and to examine the importance of the variables on these groups. In this analysis "success status" variable is a dependent variable and 24 index variables are the independent variables.

3. RESULT

This section presents the research findings obtained from the analyses carried out in parallel with the research questions and makes brief interpretations of these findings as well. To predict the success of students' reading skills, artificial neural networks "Multilayer Perceptron (MLP)" model was used. One dependent and 24 independent index variables were included in this model. Using the total data set, 70.8% (n=4556) of the data were allocated to the training set and 29.2% (n=1875) to the test set. While predicting the success of reading skills, 50 trials were made to find the architecture of the network that gives the best performance. Artificial neural network is three layers, and there are 24 artificial nerve cells (neurons) in the first layer (input layer) and seven artificial nerve cells in the hidden layer, which is the second layer. In the last layer (output), there are two nerve cells representing each level of the dependent variable. “Hyperbolic Tangent Function” is applied as activation function in hidden layer and “Softmax Function” is applied in output layer. The results of the analysis regarding the success of reading skills with the artificial neural network are shown in Table 1. The seed value of analysis to reproducibility is 2695748.

Table 1. *Analysis Results on Artificial Neural Networks Reading Skills Achievement.*

Sample	Observed	Estimated		
		Unsuccessful	Successful	Classification Rate
Training	Unsuccessful	1776	507	77.8%
	Successful	553	1691	75.4%
	Total	51.4%	48.6%	76.6%
Test	Unsuccessful	688	248	73.5%
	Successful	276	692	71.5%
	Total	50.6%	49.4%	72.5%

When Table 1 is examined, it is seen that the artificial neural network correctly predicted the reading skills success of the students in the training sample with a performance of 76.6% and the reading skills success of the students in the test sample with a performance of 72.5%. In addition, it correctly classified 75.4% of the successful students in the training dataset and 71.5% of the successful students in the test dataset. While 77.8% of the unsuccessful students in the training dataset were classified correctly, 73.5% of the unsuccessful students in the test dataset were classified correctly. The overall correct classification rate of the training and test data sets was calculated as 75.4%. The maximum chance criterion of the sample is 0.51 and the relative chance criterion is 0.49 The value of 75.4% is above the maximum and relative chance criterion. This result shows that artificial neural networks can be used successfully in classification in this model.

When the degree of importance of the independent variables used in the analysis on the success of reading skills is examined, the most important input variables related to the success of reading skills can be seen as "Home Possessions (100%)" and "Family Wealth (82%)" as the most important determinants of success regarding reading skills. The independent variables that have the least effect on the success of reading skills include “Teacher's Stimulation of Reading Engagement Perceived by Student (14.8%)” and “Subjective Well-Being: Positive Affect (9.9%)”.

ROC analyzes are performed in the analysis of artificial neural network models. The area under the ROC curve is called the "AUROC or AUC" (Area under the ROC curve) and is a measurement that helps determine the reliability of the model. The AUROC value takes values between 0.5 and 1.0. The closer the probabilities of the AUROC index are to one, the more successful the result will be. In the relevant literature, it is stated that the discrimination ability of the prediction model can be classified as follows:

'AUROC' =0.5 No prediction probability, so no discrimination.

0.7≤ 'AUROC'≤ 0.8 statistically acceptable discrimination.

0.8≤ 'AUROC' ≤0.9 statistically perfect discrimination.

'AUROC'>0.9 is statistically outstanding.

As can be seen in Table 2, a statistically perfect discrimination ability with 0.837 value was presented by the model. With this analysis, the performance of the model was also tested.

Table 2. *Areas under the Curve as a Result of ROC Analysis.*

		Areas Under the Curve
Success Status	Unsuccessful	0.837
	Successful	0.837

To predict students' reading skills success with decision tree, the results of analysis of four decision tree algorithms were examined. One dependent and 24 independent index variables were included in the analysis of the decision trees algorithm, and using the total data set, 69.6% (n=4476) of the data were determined for training and 30.4% (n=1955) for testing.

In the study, the C5.0 algorithm was run with "Boosting" to increase the accuracy rate. In this model, a 10-fold cross-validation test was used as a validation test. The standard deviation of the model determined by the cross-validation method was 0.7% and the depth of the decision tree was 21. The analysis results regarding the success of reading skills with the C5.0 algorithm are shown in Table 3.

Table 3. *Analysis Results of C5.0 Algorithm.*

Sample	Observed	Estimated		
		Unsuccessful	Successful	Classification Rate
Training	Unsuccessful	2014	219	91%
	Successful	259	1984	88.5%
	Total	50.7%	49.2%	89. 3%
Test	Unsuccessful	895	91	90.8%
	Successful	96	873	90%
	Total	50.7%	49.3%	90.4%

When Table 3 is examined, it is seen that the C5.0 algorithm correctly predicted the success in reading skills of the students in the training sample with a performance of 89.3%, and the success in reading skills of the students in the test sample with a performance of 90.4%. In addition, it correctly classified 88.5% of the successful students in the training dataset and 90% of the successful students in the test dataset. While 91% of the unsuccessful students in the training dataset were classified correctly, 90.8% of the unsuccessful students in the test dataset were classified correctly. According to this result, the overall correct classification rate of the C5.0 algorithm was calculated as 89.6%. The maximum chance criterion of the sample is 0.51 and the relative chance criterion is 0.49 Since overall correct classification rate is above these values, it can be concluded that the C5.0 algorithm can be used successfully in classification.

When the degree of importance of the independent variables used in the analysis on the success of reading skills is examined, it is seen that the "Self-Concept of Reading: Perception of Difficulty (0.047)" variable is the most important determinant of the success of reading skills. It is seen that the variability in the " Self-Concept of Reading: Perception of Difficulty " greatly affects the success of reading skills. The variable "Index of Economic, Social and Cultural Status (0.037)", that is, the socio-economic status of students, is the most ineffective independent variable on the success of reading skills. This situation shows that the socio-economic development and wealth of the student are not important on the success of reading and do not contribute to their success.

In the analysis of the CHAID algorithm, the largest tree depth was 10, the chi-square calculation method is Pearson, the stopping criteria were calculated as 2% for the root node, 1% for the child node, and the largest iteration was 100. The analysis results regarding the success of reading skills with the CHAID algorithm are shown in Table 4.

Table 4. *Analysis Results of CHAID Algorithm.*

Sample	Observed	Estimated		
		Unsuccessful	Successful	Classification Rate
Training	Unsuccessful	1830	403	82%
	Successful	387	1856	82.7%
	Total	49.5%	50.4%	82.3%
Test	Unsuccessful	669	317	68%
	Successful	291	678	70%
	Total	49.1%	50.9%	69%

When Table 4 is examined, it is seen that the CHAID algorithm correctly predicted the reading skills success of the students in the training sample with a performance of 82.3%, and the reading skills success of the students in the test sample with a performance of 69%. In addition, it correctly classified 82.7% of the successful students in the training dataset and 70% of the successful students in the test dataset. While 82% of unsuccessful students in the training dataset were classified correctly, 68% of unsuccessful students in the test dataset were classified correctly. Based on this result, it is possible to say that the CHAID algorithm gives good results in predicting successful students. The overall correct classification rate of the CHAID algorithm was calculated as 78.2%. It can be concluded that the CHAID algorithm can be used successfully in classification because the classification rate of the CHAID algorithm, which is 78.2%, is above the maximum and relative chance criterion values.

When the degree of importance of the independent variables used in the analysis on the success of reading skills is examined, it is seen that the most important input variables are "Reading and Using Strategies (0.18)" and "Summarizing (0.11)". It is seen that "Teacher Support (0.004)" is the most ineffective one on the success of reading skills. This situation shows that teachers' help to students in learning and their support in understanding a subject are not much important on the success of reading, and teacher support on the success of reading does not contribute to their success in learning and comprehension.

In the C&RT algorithm analysis, the largest tree depth is five, the largest number of proxies is zero (indicating that there is no missing value in the data set), impurity measurement is Gini for the categorical target area, stopping criteria is 2% for the root node and 1% for the child node. The analysis results obtained for the C&RT algorithm are shown in Table 5.

Table 5. *Analysis Results of C&RT Algorithm.*

Sample	Observed	Estimated		
		Unsuccessful	Successful	Classification Rate
Training	Unsuccessful	1734	799	68.4%
	Successful	488	1755	78.2%
	Total	46.5%	53.4%	77.9%
Test	Unsuccessful	736	250	74.6%
	Successful	264	705	72.7%
	Total	51.1%	48.8%	73.7%

Table 5 shows that the C&RT algorithm correctly predicted the reading skills success of the students in the training sample with a performance of 77.9%, and the reading skills success of the students in the test sample with a performance of 73.7%. In addition, it correctly classified 78.2% of the successful students in the training dataset and 72.7% of the successful students in the test dataset. While 68.4% of the unsuccessful students in the training dataset were classified correctly, 74.6% of the unsuccessful students in the test dataset were classified correctly. According to these results, it is possible to say that the C&RT algorithm gives good results in predicting especially successful students in the same way as the CHAID algorithm does. The overall correct classification rate of the C&RT algorithm was calculated as 76.6%. It can be concluded that the C&RT algorithm can be used successfully in classification because the value of 76.6% is above the maximum and relative chance criteria of the sample.

When the degree of importance of the independent variables used in the analysis on the success of reading skills is examined, it is seen that as in the CHAID algorithm analysis the most important input variables are "Reading and Using Strategies (0.18)" and "Summarizing (0.10)", while “Teacher Support (0.005)” is the most ineffective one on the success of reading skills. However, values of the degree of importance are different from those in the CHAID algorithm analysis. This situation shows that the teachers' help to students in learning and their support in understanding a subject are not much important on the success of reading, and teacher support on the success of reading does not contribute to their success in learning and comprehension.

Quadratic separation analysis was used in the QUEST algorithm, and each node was divided into two subgroups. Analysis parameters are maximum tree depth 10, maximum number of proxies 0, Alpha (for splitting) 0.05, stopping criteria 2% for root node, and 1% for child node. Analysis results of the QUEST algorithm are presented in Table 6.

Table 6. *Analysis Results of QUEST Algorithm.*

Sample	Observed	Estimated		
		Unsuccessful	Successful	Classification Rate
Training	Unsuccessful	1715	518	76.8%
	Successful	555	1688	75.2%
	Total	50.7%	49.2%	76.3%
Test	Unsuccessful	732	254	74.2%
	Successful	268	701	72.3%
	Total	51.1%	48.8%	73.3%

When Table 6 is examined, it can be seen that the QUEST algorithm correctly predicted the reading skills success of the students in the training sample with a performance of 76.3% and the reading skills success of the students in the test sample with a performance of 73.3%. In addition, it correctly classified 75.2% of the successful students in the training dataset and 72.3% of the successful students in the test dataset. While 76.8% of the unsuccessful students

in the training dataset were classified correctly, 74.2% of the unsuccessful students in the test dataset were classified correctly. According to these results, it is possible to say that the QUEST algorithm gives good results in predicting unsuccessful students. The overall correct classification rate of the QUEST algorithm was calculated as 75%. The maximum and relative chance criteria of the sample are 0.51 and 0.49, respectively. The results of the QUEST algorithm analysis show the classification rate of 75% above these values. This result shows that the QUEST algorithm can be used successfully in classification in this model.

When the degree of importance of the independent variables used in the analysis on the success of reading skills is examined, it is seen that the most important input variables are "Reading and Using Strategies (0.196)" and "Summarizing (0.115)". The independent variable that has the least effect on the success of reading skills is Subjective Well-Being: Positive Affect (0.0001)". "Subjective Well-Being: Positive Affect " variable, that is, different emotions that students may have when they evaluate themselves (joyful, cheerful, and happy), has little effect on the success of their reading skills. It shows that the positive effects and emotions of the students are not important on their success of reading, and the happiness of the students does not contribute to their success in reading.

In the K-Nearest Neighbor analysis, Manhattan Distance Measure was chosen as the distance measure since it increases the accuracy rate. For the validity test, the k value, which gives the lowest error rate as a result of the 10-fold cross-validation test, was calculated as five. K-Nearest Neighbor method analysis results are presented in Table 7.

When Table 7 is examined, it can be seen that the K-Nearest Neighbor method correctly predicted the reading skills success of the students in the training sample with a performance of 81.2%, and the reading skills success of the students in the test sample with a performance of 82.1%. In addition, it correctly classified 81.9% of the successful students in the training dataset and 83% of the successful students in the test dataset. While 80.6% of the unsuccessful students in the training dataset were classified correctly, 81.2% of the unsuccessful students in the test dataset were classified correctly. According to this result, it is possible to say that the K-Nearest Neighbor method gives good results, especially in predicting successful students. The overall correct classification rate of the K-Nearest Neighbor was calculated as 81.5%. This result of analysis shows that K-Nearest Neighbor method can be used successfully in classification in this model, since the classification rate is above the maximum and relative chance criterion values.

Table 7. *Analysis Results of K-Nearest Neighbor.*

Sample	Observed	Estimated		
		Unsuccessful	Successful	Classification Rate
Training	Unsuccessful	1801	432	80.6%
	Successful	406	1837	81.9%
	Total	49.3%	50.6%	81.2%
Test	Unsuccessful	801	185	81.2%
	Successful	164	805	83%
	Total	49.3%	50.6%	82.1%

When the degree of importance of the independent variables on the success of reading skills is examined, it is seen that the most important input variables for the K-Nearest Neighbor method are "Reading and Using Strategies (0.0438)" and "Summarizing (0.0433)". The effects of the variables are very close to each other, but the variability in "Reading and Using Strategies" greatly affects the success of reading skills. In addition, the variable "Perception of Cooperation at School (0.0409)", that is, cooperation among students in learning, is the most ineffective

independent variable on the success of reading skills. This situation that cooperation between students, or giving importance to cooperation, is not important on the success of reading and does not contribute to their reading success.

The findings regarding the success of reading skills with Naive Bayes analysis are shown in Table 8.

Table 8. *Analysis Results of Naive Bayes.*

Sample	Observed	Estimated		
		Unsuccessful	Successful	Classification Rate
Training	Unsuccessful	1716	517	76.8%
	Successful	517	1726	76.9%
	Total	49.8%	50.1%	76.9%
Test	Unsuccessful	757	229	76.7%
	Successful	225	744	76.7%
	Total	50.2%	49.7%	76.78%

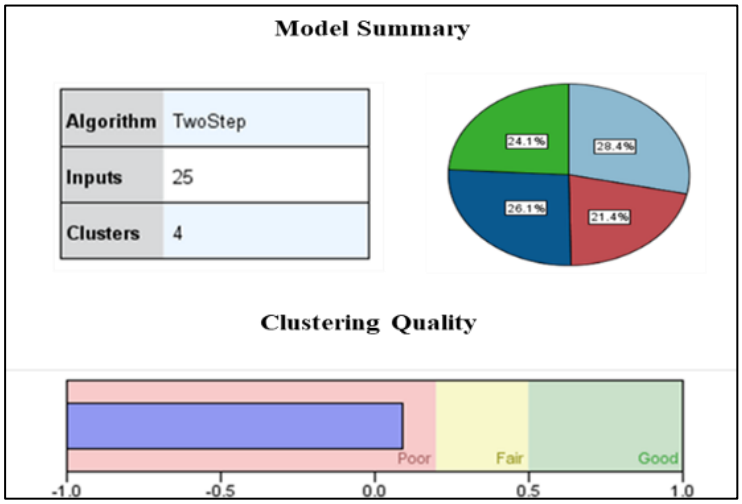
When Table 8 is examined, it can be seen that the Naive Bayes method correctly predicted the reading skills success of the students in the training sample with a performance of 76.9% and the reading skills success of the students in the test sample with a performance of 76.78%. In addition, it correctly classified 76.9% of the successful students in the training dataset and 76.7% of the successful students in the test dataset. While 76.8% of the unsuccessful students in the training dataset were classified correctly, 76.7% of the unsuccessful students in the test dataset were classified correctly. The overall correct classification rate of the Naive Bayes method was calculated as 76.8%. The overall classification rate as a result of analysis is above the maximum and relative chance criteria of the sample. According to this result, it can be concluded that the Naive Bayes method can be successfully used in classification in this model.

When the degree of importance of the independent variables on the success of reading skills is examined, it is seen that the most important input variables are “Disciplinary Climate (0.667)” and “Perception of Difficulty of the PISA Test (0.623)”. The independent variable that has the least effect on the success of reading skills is “Use of ICT at School (0.367)”. This situation that students' use of information and communication technologies at school is not important on the success of reading and does not contribute to their reading success.

In the Two-Step Clustering Analysis using one dependent and 24 independent index input variables, the Silhouette Coefficient was calculated as 0.1 and the clustering quality indexed to the Silhouette coefficient is shown in Figure 1. In the literature, a precise threshold value is not defined in the evaluations regarding the Silhouette coefficient. However, it is stated that a coefficient value greater than 0 is sufficient for clusters, and the larger the coefficient, the better the quality of the cluster. In this context, it can be concluded that although the Silhouette coefficient value (0.1) in the Two-Step Clustering Analysis is small, it is sufficient for clustering.

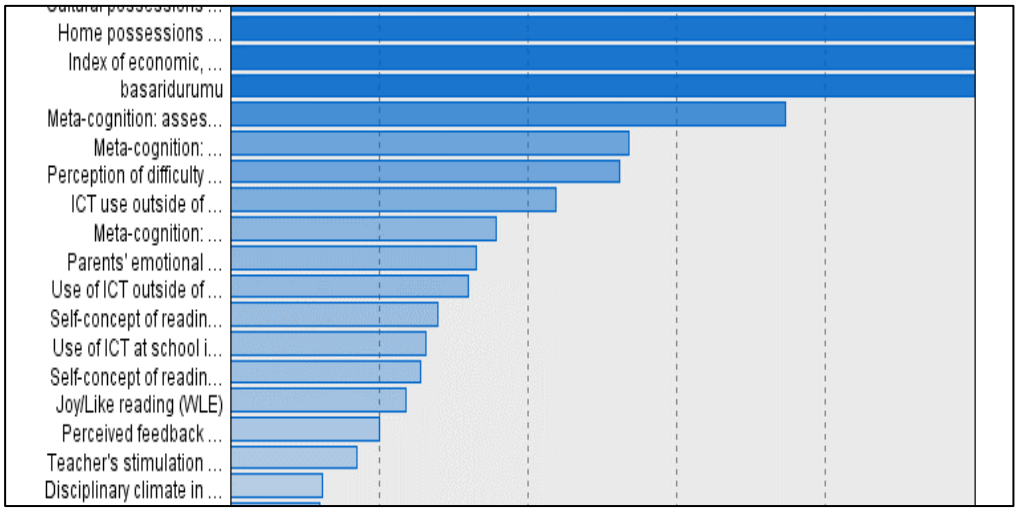
As a result of the clustering analysis, Silhouette coefficient four clusters were obtained, and it was determined that the distributions of these clusters were proportionally close to each other. The ratio from the largest to the smallest cluster was found to be 1.33. This ratio should be less than 2. In this context, it is seen that the size of the clusters and the ratio from the largest to the smallest cluster are appropriate. Variables according to their importance in cluster analysis are shown in Figure 2.

Figure 1. *Clustering Quality Indexed to Silhouette Coefficient.*



According to the findings, successful students gathered in the First and Second Clusters. It is seen that "Use of ICT for School Work Outside of School ", "Perceived Feedback", " Perception of Cooperation at School", " Perception of Difficulty of The PISA Test", and " Self-Concept of Reading: Perception of Difficulty" did not have a significant effect on successful students, while "Success Status", "Reading and Using Strategies", "Summarizing" "Family Wealth", "Information and Communication Technologies (ICT) Resources", and "Home Educational Resources " variables had a significant effect by performing well. On the other hand, unsuccessful students were gathered in the Third and Fourth Clusters. In terms of variables, it is revealed that "Reading and Using Strategies", "Summarizing", "Meta-Cognition: Understanding and Remembering" and "Joy/Like Reading" did not have a significant effect on unsuccessful students, while "Use of ICT For School Work Outside of School" of “Using” and “Use of ICT at School” hag a significant effect by performing well.

Figure 2. *The degree of importance of cluster analysis independent variables.*



The results of the comparison of correct classification rates according to the analysis of Artificial Neural Networks, Decision Trees, K-Nearest Neighbor, and Naive Bayes methods are shown in Table 9.

Table 9. *Analysis Results of Naive Bayes.*

Method		Classification Rate (%)
Artificial Neural Networks		75.4
Decision Trees	C5.0	89.6
	CHAID	78.2
	C&RT	76.6
	QUEST	75
K-Nearest Neighbor		81.5
Naive Bayes		76.8

As is seen in Table 9, Decision Trees C5.0 algorithm has the highest classification rate with 89.6%. The second highest rate is the K-Nearest Neighbor method with 81.5%. QUEST algorithm has the lowest classification rate with 75%. However, the classification rates of other methods and algorithms are close to each other. The results of the analysis made according to the success status of the students participated in the PISA 2018 Turkey application are above the maximum chance criterion and the relative chance criterion of the samples. According to the results, Artificial neural networks, Decision Tree algorithms, K-Nearest Neighborhood, and Naive Bayes methods can be used successfully in classifying students according to their success since these models make accurate classification beyond chance.

These results are in parallel with the study of Calis, Kayapınar, and Çetinyokuş (2014), who used decision trees for classification in data mining and tested the accuracy of classification according to demographic structures of individuals in four decision tree algorithms and revealed that C5.0 had a higher correct classification rate than that of other algorithms. Similarly, credit scores were calculated by comparing neural networks, M5, logistic regression, and K-Nearest Neighborhood (KNN) algorithms in the study by Liu and Schumann (2005), and the highest classification accuracy was obtained with the K-Nearest Neighbor (KNN) method. In the study, where the models obtained by Artificial Neural Networks and Decision Trees methods to compare the insurance risk estimation performances, the prediction success of the Decision Trees method was found to be higher, although both methods are at an acceptable level (Şahin, 2018). These results also show that there is a parallelism between the studies.

4. DISCUSSION and CONCLUSION

Statistical results and inferences are revealed with the analysis of data types that occur for the solution of research problems in the field of education with Educational Data Mining. In this study, based on PISA 2018 data, those factors affecting the success of students' reading skills and the success scores of their reading skills were examined with data mining classification methods and classification accuracies.

When the findings are evaluated, it is seen that Artificial Neural Networks classify with 75.4%, Decision Tree algorithms C5.0 89.6%, CHAID 78.2%, CART 76.6%, QUEST 75%, K-Nearest Neighborhood 81.5%, and Naive Bayes method 76.8%. In the study Decision Trees C5.0 algorithm has the highest classification rate with 89.6%, and the QUEST algorithm has the lowest classification rate with 75%. It is seen that the classification rates of other methods and algorithms are close to each other. The K-Nearest Neighbor method has the second highest rate of classification by having a higher classification rate than that of other methods. These findings coincide with the study in which the C5.0 decision tree algorithm makes the best prediction based on the analysis of Logistic Regression, Artificial Neural Networks, Decision Tree algorithms using student credentials, previous success status, and electronic learning data (Aydın, 2007). In addition, it is possible to say that there is a parallelism with the study in which the success rate of the K-Nearest Neighbor (KNN) analysis was found to be much higher than

that of other data mining algorithms, as a result of examining the success rate of the model developed for estimating the cellular location of proteins in the field of biotechnology (Cai & Chou, 2003).

In order to examine the general characteristics of the success groups according to the 2018 PISA reading achievement scores of the students, four clusters were obtained as a result of the Two-Step Cluster Analysis method, and it was determined that the distributions of these clusters were proportionally close to each other. In the Two-Step Cluster Analysis, the Silhouette Coefficient was calculated as 0.1. Since this coefficient is greater than 0.1, it can be said that the data set is suitable for clustering. The ratio of largest cluster to the smallest cluster, which should be less than 2, is 1.33. According to these findings, it is revealed that clustering is appropriate. When the variables are examined according to their importance, it is seen that the degree of importance of "Information and Communication Technologies (ICT) Resources", "Family Wealth", "Home Educational Resources", "Cultural Possessions at Home", "Home Possessions", "Index of Economic, Social and Cultural Status", and "Success Status" is 1 and they are effective in clustering and the most distinguishing variables by making a significant difference. It was found that the variables of "Disciplinary Climate", "Subjective Well-Being: Positive Affect", "Subjective Well-Being: Sense of Belonging to School", "Perception of Cooperation at School", and "Teacher Support" are not effective in distinguishing those with the lowest discrimination and do not make a significant difference.

The maximum chance criterion calculated within the scope of the ratios of the "successful" and "unsuccessful" students in the sample is 0.51, and the relative chance criterion is 0.49.9. It is evaluated that these Artificial Neural Networks, Decision Trees, K-Nearest Neighbor (KNN), and Naive Bayes methods can be used to classify students according to their success status and the produced models can correctly classify beyond chance, since the classification rates are above the sample's maximum and relative chance criterion values. It has been revealed that reading achievement scores are effective in separating students according to their success status and make a significant difference.

There are different methods and algorithms for prediction and classification, which has been studied extensively in data mining. However, when the studies are examined, it is revealed that Artificial Neural Networks and Decision Trees are the most studied methods. In other studies, on the same or similar sample, success can be estimated or predicted by means of such other classification methods as Regression Support Vector Machines, K-Means, and Time Series Analysis.

In the first stage of the study, loss and missing data were completed, and then the analyses were made. However, some analyses of classification methods in data mining can also be performed with missing data. In this context, how the analyses of the same or similar data sets and other classification methods perform in missing data can be examined.

SPSS Modeler program was used for the analysis. There are many data mining analysis Üprograms. In order to compare the programs, data mining methods and analysis programs can be compared using the same data sets. However, similar studies can be conducted on exams with different data such as TIMMS and PIRLS.

Studies in the field of education are carried out with different sample sizes, including different variables and different methods to divide students as successful and unsuccessful such as upper-lower 27% groups. Therefore, it is necessary to use a large number of methods and algorithms for classification and comparison purposes in order to determine which method or algorithm performs better in the sample used.

Acknowledgments

This study is based on a summary of the doctoral dissertation entitle "Examination of Reading Literacy Levels in PISA 2018 Turkey Sample with Different Data Mining Classification Methods".

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

Emrah Büyükatak: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing-original draft. **Duygu Anıl:** Methodology, Supervision, and Validation. Authors may edit this part based on their case.

Orcid

Emrah Büyükatak https://orcid.org/0000-0002-5341-5053
Duygu Anıl https://orcid.org/0000-0002-1745-4071

REFERENCES

Aksoy, E. (2014). *Determination of the mathematically gifted and talented students using data mining in terms of some variables* [Master Thesis] Dokuz Eylül University Department of Educational Sciences.

Anıl, D. (2008). The analysis of factors affecting the mathematical success of Turkish students in the PISA 2006 evaluation program with structural equation modeling. *American-Eurasian Journal of Scientific Research*, *3*(2), 222-227.

Aydın, S. (2015). *Data mining and an application on Anadolu University distance education system* [Doctoral dissertation]. Anadolu University.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovering data mining: from concept to implementation*. Prentice-Hall, Inc.

Cai, Y.D., & Chou, K.C. (2003). Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. *Biochemical and Biophysical Research Communications*, *305*(2), 407-411. https://doi.org/10.1016/S0006-291X(03)00775-7

Çalış, A., Kayapınar, S., & Çetinyokuş, T. (2014). An application on computer and internet security with decision tree algorithms in data mining. *Journal of Industrial Engineering*, *25*(3), 2-19. https://dergipark.org.tr/en/pub/endustrimuhendisligi/issue/46771/586362

Erdil, Z. (2010). Relationship of academic achievement and early intervention programs for children who are at socio-economical risk. Journal of *Hacettepe University Faculty of Nursing*, *17*(1), 72-78. https://dergipark.org.tr/en/pub/hunhemsire/issue/7840/103271

Gelbal, S. (2010). The effect of socio-economic status of eighth grade students on their achievement in Turkish. *Education and Science*, *33*(150). http://eb.ted.org.tr/index.php/EB/ article/view/626

Liu, Y., & Schumann, M. (2005). Data mining feature selection for credit scoring models. *Journal of the Operational Research Society*, *56*(9), 1099-1108. https://doi.org/10.1057/palgrave.jors.2601976

Nisbet, R., Elder, J., & Miner, G. (2009). *Handbook of statistical analysis and data mining applications.* Academic Press.

Özbay, Ö. (2015). The current status of distance education in the world and Turkey. *The Journal of International Educational Sciences*, *2*(5), 376-394.

Özer, Y., & Anıl, D. (2011). Examining the factors affecting students' science and mathematics achievement with the structural equation modeling. *Hacettepe University - Journal of Education,* 41, 313-324. https://app.trdizin.gov.tr/makale/TVRJMU1qa3lNZz09

Rizvi, S., Rienties, B., & Khoja, S.A. (2019). The role of demographics in online learning; A decision tree based approach. *Computers & Education*, 137, 32-47. https://doi.org/10.1016/j.compedu.2019.04.001

Roiger, R.J. (2017). *Data mining: a tutorial-based primer*. Chapman and Hall/CRC.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135-146. https://doi.org/10.1016/j.eswa.2006.04.005

Şahin, M. (2018). *Risk assessment in car insurance using decision trees and artificial neural networks* [Doctoral dissertation]. Yıldız Technical University Department of Statistics.

Witten, I.H. & Frank, E. (2000). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann Publishers.

Xu, Y., & Goodacre, R. (2018). On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing, 2*(3), 249-262.

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 4, 883–905
https://doi.org/10.21449/ijate.980870
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Assessment tools and strategies used by Jamaican secondary school teachers

Clavia T. Williams-McBean[1,*]

[1]University of the West Indies, Mona Campus, Faculty of Humanities & Education, School of Education

ARTICLE HISTORY
Received: Aug. 9, 2021
Revised: Jul. 7, 2022
Accepted: Sep. 18, 2022

Keywords:
Classroom assessment,
Jamaica,
Traditional assessment,
Alternative assessment,
Assessment tools.

Abstract: There is an increasing understanding that assessment is an integral part of teaching and learning and that teachers are largely not adequately prepared for their assessment responsibilities. Consequently, there is a need for research on what teachers need to improve their assessment practices. To determine what Jamaican secondary school teachers need, this mixed methods study was conducted to describe the assessment tools and strategies used by secondary school teachers of various subjects and in different types of schools as the basis for future interventions. Data was collected from a survey of 1088 secondary school teachers of varying subjects and school types and further explored through interviews and observations of 32 teachers of English. Analysis of the data using descriptive statistics and ANOVA in the quantitative phase revealed that secondary school teachers primarily used traditional assessment tools and strategies, particularly tests, despite school type. Pattern coding and pattern matching in the qualitative phase confirmed these results. The findings also revealed statistically significant differences in the frequency of use of traditional and alternative assessment tools and strategies based on the subject the teachers taught. Qualitative explorations revealed that school policies that require a quota of grades and state or express positive attitudes towards tests influenced teachers despite school type to use traditional methods. The findings imply that school administrators need to implement supportive school-level policies and display positive attitudes toward alternative assessments to maximize the use of assessment to improve learning.

1. INTRODUCTION

Assessment has been given international attention in recent times as the need for educational accountability increased (Kubiszyn & Borich, 2013), and as the use of assessment to improve student learning (i.e., formative assessment) has been promoted, investigated, and reported (Monteiro et al., 2021). Because of the promise of assessment, particularly formative assessment, in improving student learning, there have been worldwide efforts to improve teachers' assessment knowledge and skills. However, repeated reports have confirmed that teachers' knowledge and skill in engaging in effective assessment practices that can make the promise of assessment a reality need improvement (Acar-Erdol & Yıldızlı, 2018; Sewagegn, 2019; Organisation for Economic Cooperation and Development [OECD], 2019). Research has also indicated the need for empirical studies on what teachers need to improve their assessment

*CONTACT: Clavia T. Williams-McBean claviawilliams@gmail.com University of the West Indies, Mona, Faculty of Humanities & Education, School of Education

e-ISSN: 2148-7456 /© **IJATE** 2022

practice (Jiang, 2020). To determine what teachers need, it is crucial to determine, understand and describe current assessment practices. Understanding where the teachers are makes efforts to determine and take them where they are supposed to be more effective.

In the Jamaican context, The National Education Inspectorate (NEI) has repeatedly reported that teachers' use of assessment needs improvement (National Education Inspectorate [NEI], 2013; 2014; 2015; 2016; 2017). Additionally, there is a dearth of empirical research on assessment in Jamaica, in general, and the formative use of assessment in the Jamaican context (Williams-McBean, 2021). I found only one paper on the assessment strategies Jamaican teachers used, and it focused on only one of the seven education regions in Jamaica (Onyefulu, 2018). Additionally, Onyefulu (2018) reported that she could find no published article on Jamaican teachers' classroom assessment practices. Therefore, I embarked on the Spotlight on Assessment in Jamaica Project (SAAJP). The project aims to study existing assessment principles and practices in Jamaican schools in all seven regions, use the information to determine what teachers need to improve their assessment practices, then design and implement interventions to improve assessment policies and practices in Jamaica. The premise is that understanding what exists will increase the effectiveness of later interventions to improve teaching and learning through assessment. This paper is the first in a series that shares the results of the first phase of the project that describes the existing nature of assessment in secondary schools across Jamaica. In describing the existing nature of assessment, I focused on what assessment tools and strategies secondary school teachers used, how they used them, and, the factors that influenced their choice of assessment. However, because of the extensiveness of the data, this paper only focused on the assessment tools and strategies used by secondary teachers. Subsequent papers will report the findings on the other two areas of focus. In seeking to describe the assessment tools and strategies used by Jamaican secondary school teachers, I sought answers to the following research questions:

1. What assessment tools and strategies do Jamaican secondary school teachers use most frequently?
2. What is the difference in teachers' reported frequency of use of the different types of assessment tools and strategies based on subject?
3. What is the difference in teachers' reported frequency of use of the different types of assessment tools and strategies based on school type?

1.1. Assessment Tools and Strategies Used by Teachers

The 2014 Standards for Educational and Psychological Measurement distinguishes 'test' from 'assessment' by outlining that 'tests' refer to "scales, inventories, pen-and-paper tests, orals, free-format responses, and authentic assessments" (American Educational Research Association [AERA] et al., p. 2) and defining 'assessment' as "a process that integrates test information with information from other sources (e.g., information from other tests, inventories, and interviews; or the individual's social, educational, employment, health, or psychological history)" (AERA et al., p. 2). These definitions indicate that the 'test' is the instrument (traditional or alternative) that is used to measure learning. I agree with this definition despite the lack of consensus on the definition and specific differences between the two terms among researchers in the field (see, for example, Kubiszyn & Borich, 2013; Miller et al., 2013; Popham, 2018). However, since most people (including the respondents in this research) associate tests with the traditional pen-and-paper, one-shot examinations, I used it in that way for shared understanding. The term assessment tools and strategies refer to all the testing tools and techniques used to provide the measurement and qualitative data used in the assessment process. Assessment refers to the process by which the measurement and/or qualitative data on the nature and extent of students' learning are used by teachers, students, or administrators for formative, summative, and evaluative purposes.

Assessment tools and strategies are differentiated by their format (traditional or alternative, testing or performance), purpose (diagnostic, formative, summative, evaluative), location (internal or external, classroom or standardised), their relative weight or importance (high stakes or low stakes) and the interpretation of the results (norm-referenced and criterion-referenced) (Acar-Erdol & Yıldızlı, 2018; Kubiszyn & Borich, 2013). In this research, the focus is on the format of the assessment used in the classroom by teachers of varying subjects across Jamaica. In terms of format, assessment tools and strategies are classified as traditional or alternative. Traditional classroom assessment models the format and administration of standardised, public examinations and refers to pen-and-paper examinations that usually utilise items such as multiple-choice, true/false, matching, short answers, and essays (Dikli, 2003; Gronlund, 2006, Koh, 2017; Miller et al., 2013) although some writers classify essays as performance assessment (Wren & Gareis, 2019). In contrast, alternative assessment methods include authentic and open-ended performance assessment that requires students to use or apply their knowledge and skills while performing a task in a realistic setting. It also requires direct observation of the performance by the assessor, who uses a rubric to evaluate the quality of the performance (Brookhart, 2009). Examples of alternative assessment strategies include performances, concept maps, open-ended questions, interviews, exhibits, presentations, oral and practical demonstrations, hands-on execution of experiments, simulations (with or without the use of computers), observations, student journals, peer-assessment, self-assessment, projects, and portfolios (Adeyemi, 2015; Berry, 2008; Bland & Gareis, 2018; Dandis, 2013).

In education, over 30 years of research have reported that traditional assessment tools and strategies have dominated (Brookhart, 2013; Esomonu & Eleje, 2020; OECD, 2019; Stiggins & Conklin, 1992). However, there has been increasing advocacy for the increased use of alternative assessment methods. This advocacy is based on research results that alternative assessments impact more positively on students' intrinsic motivation and engagement than traditional assessments (Hess et al., 2020; Koh, 2017); promote and measure affective learning (Koh, 2017); more effectively allow for formative assessment (Black & Wiliam, 1998; Koh, 2017); have greater authenticity (Wren & Gareis, 2019;); and, that they are focused on deeper learning and higher-order thinking skills (Koh, 2017; Wren & Gareis, 2019;). This shift has accompanied the shift from behaviourism to constructivism and from a focus on summative assessment to formative assessment (Buhagiar, 2007; Dogan, 2011; Koh, 2017).

At the same time, some writers have taken a "middle of the road" stance. They argue that both are useful and should be used in conjunction to get the most accurate picture of student achievement (Popham, 2005; Wren & Gareis, 2019). In explaining his support for what he calls "balanced assessment," Burke (2009) posits that it should include three types of assessment: traditional (focusing on knowledge, curriculum, and skills), portfolio (process, product, and growth), and performance (standards, application, and transfer). In this way, a more comprehensive range of student skills is measured, and a more valid assessment of student achievement can be made.

The more positive impact of alternative assessment indicates that improved educational outcomes can result from its use. However, most of the studies reviewed found that despite the pedagogical shifts and the curricula rewrites, teachers' assessment practices at the secondary level have remained predominantly traditional, with tests being the most frequently used type of assessment (Acar-Erdol & Yıldızlı, 2018; Bramwell-Lalor, 2019; Brookhart, 2013; Dandis, 2013; Esomonu1 & Eleje, 2020; Guskey & Link, 2019; Saefurrohman, 2017; Vlachou, 2018). Berry (2010) reported that even when teachers used strategies or tools labelled as alternative assessments, for example, projects, their objective was to measure lower-order thinking skills and to assess knowledge acquisition and retention. Other researchers also found that elementary teachers used more varied assessment methods, including informal evidence and observations,

while secondary teachers used paper-pencil objective tests, whether commercially prepared, teacher-made, or derived from textbooks (Brookhart, 2009, 2013; Guskey & Link, 2019; Ong, n.d.; Vlachou, 2018; Zhang & Burry-Stock, 2003).

I found only one study investigating teachers' assessment practices in the Jamaican context: Onyefulu (2018). Onyefulu (2018) surveyed 157 primary and secondary school teachers in Region 1 in Jamaica and confirmed testing dominance with 51% of the primary school teachers and 85% of the secondary school teachers surveyed reporting that they most frequently used closed-book tests to assess their students. However, there were only eight assessment methods included on the research instrument. Five of the eight were a type of test (closed book test, open-book test, collaborative or negotiated test, cooperative testing, and take-home test). The other three methods were portfolio assessment, peer-assessment, and self-assessment. This research includes the reported frequency of use of 22 assessment tools and strategies from secondary school teachers from all seven educational regions in Jamaica. Additionally, since all except one of the studies reviewed were conducted outside of Jamaica and none included teachers from across the country, it was prudent to investigate if the same obtained in Jamaica. Nevertheless, these studies helped identify various assessment tools and strategies and classify them as traditional or alternative.

1.2. The Difference in Assessment Tools and Strategies Used Based on Subject

Most of the studies reviewed focused on the assessment tools and strategies used in a single subject. Therefore, they did not allow for comparisons across subjects. This inclusion is another way in which this study contributes to the existing body of literature. Additionally, among the studies reviewed that included different subjects, the findings are contradictory Some researchers reported that teachers of Mathematics indicated that they used alternative assessment methods with greater frequency than all other subject areas (Bol et al., 1998) or more than teachers of language arts, science, and social studies (Zhang & Burry-Stock, 2003). Bol et al. (1998) explained that the greater use of alternative assessment tools and strategies resulted from the Mathematics teachers' greater focus on process than product. On the other hand, researchers have found that teachers of Mathematics use predominantly traditional assessment tools and strategies (Dandis 2013; Senk et al., 1997; Watt, 2005).

The contradiction is evident for other subjects as well. For example, Zhang and Burry-Stock (2003) corroborated Marso and Pigge's (1988) study and reported that "language-arts teachers used paper-pencil tests more often than did teachers in nonacademic subjects" (p. 333). McMillan (2001) also reported that English teachers reported more frequent use of constructed-response assessment strategies than both mathematics and science teachers. Constructed response items include essays, which may be classified as traditional assessment. The term, however, also includes alternative assessment tools and strategies. Therefore, it is unclear what type of assessment (traditional or alternative the teachers in this study were using. Furthermore, Brookhart (2009) reported that teachers of Social Studies used traditional assessment (constructed-response items) more frequently than all other subjects. Berry (2010) corroborated the difference in assessment tools and strategies based on subject. However, she did not assess which subject area had a greater propensity toward what type of assessment tool or strategy. She did, however, establish that subject content played a role in the assessment tool and strategy selection of the participants in her study. According to Berry (2010), teachers reported using alternative assessment strategies if the content was "activity-based" (p. 104). To add to the contradiction, Duncan and Noonan (2007) and Ong (n.d.) reported no difference based on the subjects taught. Therefore, the results on the difference in assessment tools and strategies used by secondary school teachers based on subject are conflicting and worthy of further investigation, especially in the Jamaican context where this area is mainly unaddressed. Consequently, it was an area of focus in this research.

1.3. Difference in Assessment Tools and Strategies Used based on School Type

It is essential to consider school-type differences in Jamaica because there are grave disparities in student academic achievement among the different types of secondary schools: traditional high schools for boys, traditional high schools for girls, coeducational traditional high school, upgraded high schools, and technical high schools (Clarke, 2011; Williams-McBean, 2021). Top performers in the primary-level exit examinations are usually placed in traditional high schools. As students' academic achievement (as measured by the exit examinations) decreases, they are placed in upgraded high schools and technical high schools (Clarke, 2011; Williams-McBean, 2021). However, individual upgraded and technical high schools outperform some traditional high schools, and technical high schools outperform some upgraded high schools. In addition, research purport that the use of alternative assessment can increase student achievement (e.g., Guha et al., 2018). Since school type and academic achievement are so interconnected in the Jamaican context, investigating the types of assessment used in the different types of schools would be useful. Therefore, this research disaggregated schools based on the five types of secondary schools in Jamaica.

2. METHOD

Data was collected using a multiphase mixed methods design, which began with a quantitative phase, followed by a qualitative phase, followed by an intervention phase. However, the data presented in this paper are from the first two phases.

2.1. The Quantitative Phase

In this phase, the researcher surveyed 1,088 secondary school teachers on the types of assessment tools and strategies used, the frequency of use, the factors that influenced their choice of assessment tools and strategies, and the types of feedback they give to students.

2.1.1. *The sample*

The quantitative sample consisted of 1,088 secondary school teachers from 45 secondary schools across Jamaica. The schools were ranked (above average, average and below average) based on a three-year average of students' performance in Caribbean Secondary Examinations Certificate English A examinations – the exit examination for English at the secondary level. Therefore, the schools were stratified according to school type and rank and a sample was selected using proportionate, stratified random sampling. Of the 1,088 teachers, 587 or 54% teachers were from upgraded high schools, 213 or 19.6% from coeducational traditional high schools, 60 or 5.5% from traditional high schools for boys, 100 or 9.2%, from traditional high schools for girls, and 128 or 15.5% from technical high schools. The quantitative sample consisted of male and female teachers with varying years of experience who reported teaching various subjects categorized into nine different groups: English (English Language, English Literature, Communication Studies), Mathematics, Social Sciences (e.g., Social Studies, Religious Education, History), Sciences (e.g., Biology, Chemistry, Physics, Integrated Sciences), Business (e.g., Principles of Business, Principles of Accounts, Office Administration, Information Technology), Practical Arts (e.g., Physical Education, Woodwork, Electrical and Electronic Technology, Food and Nutrition), Performing Arts (e.g., Dance, Drama, Art), Modern Languages (Spanish and French) and Mixed (a combination of any of the categories) (see Table 1). The disproportionality within each sample variable represents the disparity that exists in the teacher population of Jamaica. There was a 95% overall response rate.

Table 1. *The quantitative sample.*

Sample Characteristic		N	%
Gender	Male	325	31
	Female	726	69
Age	Young adult	149	18
	Middle-aged	913	82
Years of Experience	0 – 5 years	275	216
	6 – 10 years	328	32
	11 – 15 years	163	16
	16 – 20 years	112	11
	≥ 20 years	154	15
School Type & Rank	**Traditional High school (Coed)**	**213**	**20**
	Above Average	47	4
	Average	85	9
	Below Average	81	7
	Traditional High school (Boys)	**60**	**6**
	Above Average	20	2
	Average	20	2
	Below Average	20	2
	Traditional High school (Girls)	**100**	9
	Above Average	33	3
	Average	33	3
	Below Average	34	3
	Upgraded High School	**587**	**54**
	Above Average	195	18
	Average	196	18
	Below Average	196	18
	Technical High School	**128**	**11**
	Above Average	37	3
	Average	52	5
	Below Average	39	3
Subject	English	191	18
	Mathematics	132	13
	Social Sciences	177	17
	Sciences	115	11
	Business	119	11
	Practical Arts	175	17
	Performing Arts	34	3
	Modern Languages	43	4
	Mixed	60	6

2.1.2. *Quantitative data collection method*

A self-developed Teacher Assessment Practices Questionnaire was used to collect data in this phase. The questionnaire was developed by relying heavily on the literature (e.g., Alkharusi, 2011; Berry, 2010; Dandis, 2013). The questionnaire contained 41 questions that were divided into four sections. The first section presented four items to capture demographic details that researchers identified as influencing teachers' choice and frequency of use of different assessment tools and strategies: gender (Alsarimi, 2000); age; years of service (Alkharusi, 2011), subject(s) taught (Alkharusi, 2011; Berry, 2010; Dandis, 2013). The second section

consists of assessment strategies and techniques scale: 22 items on a 6-point Likert scale ranging from *Don't know* (to be selected if the respondent does not know the assessment strategy) to *Frequently use*d. Each assessment tool or strategy was identified in previous studies. The tools and strategies were also classified as traditional or alternative based on Gronlund's (2006) specification. To increase the clarity of the items, the method 'Test' was used to refer to the traditional pen-and-paper test because that is how it is understood by the respondents, and (multiple-choice, true/false, matching, short answers) were included in brackets to clarify further. These item formats were identified in previous literature as items commonly used on traditional tests (see, for example, Koh, 2017; Miller et al., 2013). Additionally, putting clarifying terms in brackets after a concept is recommended by Cobern and Adams (2020) as part of the basic steps to instrument validation. Though essays are also frequently used on pen-and-paper tests, it was separated because some writers classify essays as performance assessment (e.g., Wren & Gareis, 2019). The separation allowed for more specific identification and examination of teachers' frequency of use of traditional as differentiated from alternative assessment tools and strategies. Section three consisted of 19 items on a 4-point Likert scale ranging from *Least influential* to *Extremely influential*) that listed factors, also identified from the literature, that influenced teachers' choice of assessment tools and strategies. Section four consisted of one item with five types of feedback: Grades (e.g., 70%, 9/10, B+), Ticks and Xs, Oral feedback, Written feedback on students' strengths and weaknesses and Grades accompanied by written feedback. These types of feedback were identified from the literature, and the respondents were required to select the type of feedback they most frequently gave their students.

The validity and reliability of the instrument were assured using data from two pilot studies, member checking, expert checking and a literature-validated theoretical model. According to Cobern and Adams (2020), theoretical models where the researcher uses the literature to create a model on which the survey instrument is developed "provide the first line of validity evidence for the survey" (p. 408). The selected demographic details, the assessment tools and strategies and their classification as traditional or alternative, the factors and types of feedback were all derived from the related literature. In addition, experts in quantitative research and educational assessment checked the questionnaire for content validity since expert checks are the best way to ensure content validity (Zohrabi, 2013). The educational assessment experts, who had at least a master's degree in educational measurement and taught in the area, confirmed the grouping of the tools and strategies as traditional and alternative and suggested no change to the instrument. The experts also affirmed the logical groupings of the individual factors into three categories: Student Factors (students' grade level, students' academic abilities, students' behaviour, students' motivational levels, number of students in the class, number of students in the school and expectations of the students' parents), Teacher Factors (formal teacher training, teachers' experiences as teachers, teacher's experiences as learners, teachers' knowledge of current research, teacher content knowledge, and Assessment Factors (the format of standardized tests (e.g. CSEC), availability of past papers, workload of the assessment strategy, national assessment practices, the school's assessment policy, time constraints and the demands of the national curriculum). No additions were suggested by the 10 secondary school teachers, including five heads of department, who were interviewed about the clarity and completeness of the tools and strategies, factors and types of feedback They suggested adding 'please turn over' on the first page of the instrument and increasing the spacing. Both suggestions were implemented before the questionnaire was administered to the main sample. Expert and respondent feedback was also used to ensure face validity (Oluwatayo, 2012).

The reliability of the instrument was assessed using Cronbach's alpha in SPSS. The two sub-scales in Section 2 had acceptable reliability of .65 for frequency of use of traditional assessment tools and strategies and .83 for frequency of use of alternative assessment tools and

strategies. The subscales in Section three also had acceptable reliability of .73, .60 and .71 for Student Factors, Teacher Factors, Assessment Factors, respectively. Though alpha of or greater than .70 is usually considered acceptable, researchers have also purported that alphas of .60 are acceptable (Churchill Jr. & Peter, 1984; Taber, 2018) especially for newly developed measures (Nunnally, 1978, 1988).

2.1.3. *Quantitative data analysis and presentation*

To answer the first research question: (*What assessment tools and strategies do Jamaican secondary school teachers use most frequently?*), I calculated the mean score of the non-missing values for each assessment strategy (Bryman & Cramer, 2011). Then, using Gronlund's (2006) specifications, I categorised the 22 individual assessment tools and strategies as "Traditional Assessment Strategies" and "Alternative Assessment Strategies" in SPSS. The traditional assessment tools and strategies were tests, questioning, oral quizzes, teacher observation, and essays. Concept maps, checklists, flow charts, peer evaluations, portfolios, speech/debate/drama, case studies, research reports, rubrics, self-evaluations, practical tests, role plays, student journals, contracts, conferences, anecdotal records, and interviews were categorized as alternative assessment tools and strategies. Descriptive statistics were then used to answer the question in the quantitative phase. To respond to questions 2: (*What is the difference in teachers' reported frequency of use of the different types of assessment tools and strategies based on subject?*), a one-way between-groups ANOVA with a post-hoc test was done as all the subjects were collapsed into nine categories: English, Mathematics, Social Sciences, Science, Business, Practical Arts, Performing Arts, Modern Languages and Mixed. Finally, a two-way between-group analysis of variance was used to assess the difference in teachers' reported frequency of use of the different types of assessment tools and strategies based on school type. This technique was suitable because, in this study, school type referred to the type of school (traditional, technical, upgraded) as well as the rank of the school (above average, average, below average). The results of the quantitative phase were used to select the sample for the subsequent qualitative phase.

2.2. The Qualitative Phase

2.2.1. *Research design*

Qualitative data was also collected to answer the research questions and to add depth to the research. I observed teachers in their natural settings (to determine if they used the same assessment tools and strategies, and with the same frequency, as they had reported — to add credibility to the quantitative findings and the overall conclusions from this study. A multiple-case instrumental case study design (Creswell, 2014; Yin, 2014) was used in this research phase. The cases (teachers) were embedded within the context of the schools, and they were deliberately selected to unearth different perspectives about the issue of teachers' assessment practices. Hence, they were "instrumental cases" (Creswell, 2014, p. 493).

2.2.2. *The participants*

I selected the participants in the qualitative phase through stratified purposive sampling (Patton, 1990). The six schools I selected were stratified from the quantitative phase by school type and rank. The quantitative findings showed no difference in the frequency with which teachers used traditional or alternative assessment tools and strategies based on school type. Therefore, I reduced the number of school types represented from five to three. However, I maintained the three major types – traditional, upgraded, and technical – to explore possible school type differences qualitatively. I randomly selected two schools from the three different school types retained. Of the two schools, one was from the above average rank and the other from the below average. Five or six language teachers from each selected school were observed and interviewed to explore further the methods of assessment teachers used and explain the quantitative findings.

I selected all the teachers from each school who had participated in the initial survey and were willing to continue into the qualitative phase. I interviewed all the teachers of English in each school even after saturation was achieved. After these schools had been selected, one teacher from an average, traditional high school for boys requested to continue participating in the study. Therefore, 32 teachers of English, two males, and 30 females from four types of schools, with varying years of experience, were interviewed and observed. I selected the English department for further investigation because it is the area in which I am most knowledgeable and skilled, having been a teacher and researcher of issues in English Language and Literature education at the secondary level for approximately nine years. It was also the area in which the formative assessment intervention was to be subsequently implemented. The English group also represented the largest subject group from the quantitative sample: 191 or 18%. Consequently, while the qualitative findings provide useful insights into why teachers from different school types predominantly used traditional assessment tools and strategies, the specific findings are reflective of the teachers of English within these schools.

2.2.3. *Qualitative data collection methods and procedures*

I collected data through interviews, observations, and document analysis. I interviewed the participants using in-depth, semi-structured interviews guided by an interview schedule (DiCicco-Bloom & Crabtree, 2006). The interview questions were informed by the findings of the quantitative phase, the literature reviewed, and the research questions. Participants were asked about the types of assessment used, the factors that influenced their choices in general and specific factors in their schools that would have influenced them to select any assessment tool or strategy most frequently. The interviews lasted 20–90 minutes, with a mode of 45 minutes. At the end of the day or week of each interview, I transcribed and emailed the transcripts to each participant for their verification. I intertwined data collection and analysis to allow the analysis results to guide subsequent interviews and observations. After I interviewed the participants, I observed each of them three times while they taught three different classes, with class periods lasting 45 minutes (single session) to 90 minutes (double session). However, the first observation for each teacher was not recorded to reduce reactivity (Fraenkel & Wallen, 2003; Johnson & Turner, 2003). In the other two observations, I observed classroom practices without participating in the activities. I tape-recorded each classroom observation and supplemented the recordings with my field notes. I also observed other school functions, such as prize-giving ceremonies, devotions, student activity during recess, and school paraphernalia (notice boards and paintings on the walls) — to understand the context better. I extended my field notes immediately after the observations or at the end of the day when the information was fresh in my mind. No more than three observations were conducted for a day and the observations were transcribed at the end of the day. It took five months to complete all the observations. These observations provided a direct picture of the teachers' assessment practices. It also allowed me to corroborate, refute or extend the assessment practices the teachers reported on the questionnaires and in the interviews (Charmaz, 2006). The teachers' lesson plans were also analysed to increase the accuracy of the findings through triangulation.

2.2.4. *Qualitative data analysis and presentation*

Marshall and Rossman (2016) purported that typical procedures for analyzing qualitative data involve "immersion in the data, generating categories and themes, coding the data, offering interpretations through analytical memos, searching for alternative understanding and writing the report or other format for presenting the study" (p. 209). These were the methods I employed in this study. I read through the transcripts for each case to get an overall impression of the teachers' assessment practices. Then, beginning with my research questions, I listed possible theory-generated codes and categories (Marshall & Rossman, 2016). For example, for the research question, *(What assessment tools and strategies do Jamaican secondary school*

teachers use most frequently?), the individual assessment tools and strategies: *questioning, tests, oral quizzes, teacher observation, and extended writing (essays, written speeches, and stories)* were listed as theory-generated codes under the category of *traditional assessment tools and strategies*. I also classified the other tools and strategies from the questionnaire used in the quantitative phase as alternative assessment tools and strategies. Since I was using the qualitative phase to corroborate the findings of the quantitative phase, I used all the assessment strategies on the questionnaire as codes. However, I was keen to identify tools and strategies that were not on the questionnaire. Therefore, I coded the data deductively and inductively (Saldaña, 2016). My literature review, the quantitative results, and my initial exploration of the qualitative data generated many of the deductive codes.

Using QDAMiner, I first coded sentences and chunks and employed independent coding (Thomas, 2006) by a lecturer and veteran qualitative researcher to validate my codes and coding. Then, I categorised the codes using pattern coding before seeking answers to the research questions through pattern matching (Yin, 2014). Pattern matching is where the researcher "compare[s] an empirically based pattern — that is, one based on the findings from your case study — with a predicted one made before you collected your data (or with several alternative predictions)" (Yin, 2014, p. 143). I made predictions for each research question. For example, for the first research question (*What assessment tools and strategies do Jamaican secondary school teachers use most frequently?*), I predicted that teachers used predominantly traditional assessment methods, with pen-and-paper tests being the most frequently used assessment method. This prediction was based on the review of the extant literature, the findings of the quantitative phase of this research, and the findings of the qualitative pilot study. I ran a code frequency on the category, "*Assessment tools and strategies used," to match the empirical data from the qualitative data.* This output combined the assessment tools and strategies reported by all the teachers and those I observed them using. I then separated the frequency of use across the different types of data (interview and observations) to ascertain the difference between teachers' reported and observed frequency of use. After that, I classified the tools and strategies in this list as traditional or alternative, based on Gronlund's (2006) specifications, as was done in the quantitative phase. With information on the types of assessment and the frequency of use, I assessed whether and to what extent the empirical data matched my initial prediction. I also used pattern matching to identify possible answers to the other research questions.

After the individual case analyses, I conducted cross-case analyses within the context (type of school) and across cases and contexts. These analyses were done by using the same set of categories and profiles of each case, arranging them in a matrix, and then checking for replications (similarities) and contrasts (differences) across cases (Yin, 2014). In doing the cross-case analyses, I utilised explanation building (Yin, 2014) because I wanted to explain the findings from the quantitative phase, particularly why there was no difference in teachers' reported frequency of use of assessment tools and strategies based on school type. Explanation building allowed me to provide these explanations and explore rival explanations while strengthening the credibility of the findings by showing how "these rival explanations cannot be supported given the actual set of case study findings" (Yin, 2014, p. 150). It also allowed me to "build a general explanation that fits each case, even though the cases will vary in their details" (Yin, 2014, p. 148).

To interpret the data, I looked at patterns in the data (causes and effects, sequence, hierarchy, frequencies) and extrapolated possible explanations for these relationships. I also used the "most useful data segments to support the emerging story, to illuminate the questions being explored" (Marshall & Rossman, 2016, p. 219). I looked for alternative explanations throughout, as supported by the data collected. According to Yin (2014), the findings of

multiple case studies may be reported as an overall cross-case analysis with separate sections devoted to different topics. I used this reporting format in this study. I also interspersed exemplars from the individual cases throughout the different sections.

3. FINDINGS

3.1. Types and Frequency of Use of Assessment Tools and Strategies

Based on the data analysis in the quantitative and qualitative phases, the prediction that teachers predominantly used traditional assessment tools and strategies, especially pen-and-paper tests, was corroborated. In the quantitative phase, the teachers reported using traditional forms of assessment most frequently, with tests (98.9%, n = 1072), questioning (98.4%, n = 1077), teacher observations (95.2%, n = 1063), practical tests (92.8%, n = 1053), and oral quizzes (94.4%, n = 1081) being the most frequently used tools and strategies (see Table 2). Though a higher overall percentage of the sample reported that they used oral quizzes over practical tests, practical tests were ranked higher because more teachers reported that they *always* used them. This level of frequency resulted in a higher mean score for practical tests (M = 3.68, SD 1.23) than for oral quizzes (M = 3.63, SD = 1.14). The percentage of teachers who indicated that they always used tests and questioning, 51.8 (n = 1072) and 51.6 (n = 1077), respectively, further underscored the high frequency of reported use of traditional assessment tools and strategies.

Table 2. *Assessment tools and strategies used by classroom teachers (Quantitative phase).*

Tools & Strategies	*n*		*M*	*SD*	DK	NU	SU	U	FU	AU	%
	Valid	Missing									
Tests	1072	16	4.32	0.856	0.4	0.7	1.9	11.8	33.4	51.8	98.9
Questioning	1077	11	4.27	0.946	0.7	0.7	3.4	12.7	30.7	51.6	98.4
Teacher observations	1063	25	3.85	1.194	1.5	3.3	6.4	25.9	23.4	39.5	95.2
Practical tests	1053	35	3.68	1.234	0.1	6.2	8.5	24.3	28	32	92.8
Oral quizzes	1081	7	3.63	1.137	0.7	4.9	8	29.5	31.1	25.8	94.4
Self-evaluations	1063	25	3.49	1.199	0.9	4.7	13.5	32.3	22.6	26	94.4
Essays	1064	24	3.23	1.404	1.3	14.8	14.1	22.7	23.8	23.4	84
Peer evaluations	1053	35	3.1	1.156	1.5	7.8	18.1	35.8	25.2	11.6	90.7
Roleplays	1071	17	3.1	1.262	0.7	12.4	17.7	31.5	21.4	16.3	86.9
Rubrics	1053	35	3.01	1.421	5.4	10.7	16.8	30.3	17.9	18.8	83.8
Checklist	1060	28	2.8	1.181	2.5	12.3	22.1	36.7	18.7	7.8	85.3
Speech/Debate/Drama	1053	35	2.76	1.245	1.1	18.5	20.4	32.5	17.9	9.5	80.3
Research reports	1043	45	2.74	1.222	2.2	14.9	23.9	34.4	15.2	9.4	82.9
Portfolios	1050	38	2.67	1.168	1.1	16.8	25.5	34.6	15.2	9.4	84.7
Concept maps	1058	30	2.62	1.205	2.7	16.8	26.2	30.1	18.4	5.8	80.5
Flow charts	1052	36	2.5	1.205	2.9	19.8	27.3	30.1	14.1	5.9	77.4
Student journals	1054	34	2.41	1.201	1.6	25.1	26.9	29.6	10.1	6.6	73.2
Interviews	1070	18	2.34	1.245	2.2	29.1	24.3	27.7	9.9	6.8	68.7
Case studies	1057	31	2.21	1.24	3	32.3	26.6	22	10.6	5.5	64.7
Anecdotal records	1031	57	1.93	1.402	15.5	29.7	20.6	20.7	7.6	6	54.9
Conferences	1046	42	1.85	1.207	8.2	40.6	21.9	19.5	6.6	3.2	51.2
Contracts	1040	48	1.53	1.144	13.7	47.5	19.7	13	3.6	2.6	38.9

Note. DK = Don't Know, NU = Never Used, SU = Sometimes Used, U = Used, FU = Frequently Used and AU = Always Used

On the other hand, the five least reportedly used assessment strategies were interviews (68.7%, *n* = 1070), case studies (64.7%, *n* = 1057), anecdotal records (54.9%, *n* = 1031), conferences (51.2%, *n* = 1046) and contracts (38.9%, *n* = 1040). It is also noteworthy that anecdotal records and contracts are the two strategies that were most frequently left unanswered — with 57

(0.05%) and 49 (0.05%) missing responses, respectively. This omission could indicate that more teachers did not know about these strategies but were unwilling to indicate their lack of knowledge.

The qualitative results confirmed that traditional assessment tools and strategies were reported and observed being used more frequently by the participants. In the qualitative phase, all the traditional assessment tools and strategies, except Oral Quiz, were in the top five, with 'Test' (selected response and short answer items only), 'Teacher Observation,' 'Questioning' and 'Extended Writing (essays, written speeches, and stories) ranked 1–4, respectively, as the most frequently used assessment tools and strategies (see Table 3). When I asked the participants which assessment tools or strategies they used most frequently (participants were allowed to select more than one assessment tool or strategy), 20 of the 32 participants responded 'Test'.

However, 'Teacher Observation' was the most frequently observed strategy (88 times by 21 participants), although only three teachers reported using it most frequently. Therefore, although overall traditional assessment tools and strategies were reported and observed to be the more frequently used, the specific traditional assessment tool and strategy used differed. Tests were the most frequently reported (20 counts in 20 cases), and Teacher observation (88 counts in 23 cases) was the most frequently observed. Conversely, the only alternative assessment strategy in the top five was peer assessment, listed at number five among the most frequently observed assessment tools and strategies but reported by none of the participants as the most frequently used strategy.

Table 3. *Comparison of teachers' reported and observed frequency of use of assessment tools and strategies (Qualitative phase).*

Top 10 Assessment Tools and Strategies Reported and Observed	Reported Use		Observed Use		Total	
	Counts	Cases	Counts	Cases	Counts	Cases
Test (MCQS, T/F, Short answer)	20	20	62	32	82	32
Teacher Observation	3	3	88	23	91	23
Questioning	4	4	50	27	54	27
Extended Writing	8	8	20	17	28	19
Peer-assessment	-	-	25	15	25	15
Presentation	4	4	23	13	23	13
Oral Quiz	-	-	7	4	7	4
Research Report	-	-	6	6	6	6
Dramatization	1	1	4	4	5	4
Game/Puzzle	1	1	1	1	1	1

Note. - = none was reported

3.2. Different Subject, Different Assessment Tools and Strategies

The second research question assessed differences in teachers' reported frequency of use of the different types of assessment tools and strategies based on the subject the teachers taught. This was analysed quantitatively using a one-way between-groups ANOVA with a post-hoc test as all the subjects were collapsed into nine categories: English, Mathematics, Sciences, Social Sciences, Business, Performing Arts, Practical Arts, Modern Languages, and Mixed and traditional assessment and alternative assessment as dependent variables in separate analyses.

3.2.1. *Subject differences for traditional assessment tools and strategies*

The assumptions of normality and homogeneity of variance for ANOVA had been violated for the frequency of use of traditional assessment tools and strategies. However, both the Welsh and Brown-Forsythe tests revealed a significant difference between teachers' reported frequency of use of traditional assessment tools and strategies based on subject ($p < 0.001$ for

both tests). The non-parametric, Kruskal-Wallis test which both Field (2013) and Pallant (2013) recommend instead of a one-way between-group ANOVA when the distribution is not normally distributed, also showed a significant difference (see Table 4). Therefore, it was concluded that there was a significant difference in teachers' reported frequency of use of traditional assessment tools and strategies based on subject.

Table 4. *Kruskal-Wallis test for subject*traditional assessment tools & strategies.*

Null Hypothesis	Test	Sig.	Decision
The distribution of traditional assessment tools and strategies is the same across categories of subject.	Independent-Sample Kruskal-Wallis Test	< 0.001	Reject the null hypothesis

Note. Asymptotic significances are displayed. The significance level is .05.

The post hoc test results revealed that the differences between teachers of English and teachers of Mathematics and Practical Arts were significant at a confidence interval of .05. The teachers of English reported using traditional assessment tools and strategies more frequently than teachers of Mathematics and Practical Arts (M = 4.04, SD .64 for teachers of English and M = 3.54 and 3.75, SD = .56 and .69 for teachers of Mathematics and Practical Arts, respectively). There were significant differences among other subject areas as well. The teachers of Mathematics reportedly used traditional assessment tools and strategies significantly less frequently than the teachers of Social Studies, Science, and Business. There was also a significant difference between the Social Sciences teachers (M = 4.07, SD .64) and the teachers of the Practical Arts (M = 3.79, SD = .69). The effect size was moderate at .06. Consequently, I concluded that there were practical differences.

Overall, the teachers of Mathematics reported using traditional assessment tools and strategies with the least frequency (M = 3.54, SD = .56). On the other hand, the Social Sciences and English teachers reported the highest use of traditional assessment tools and strategies. This result is arguably because 'essay', which is frequently used as an assessment tool by English teachers, was classified as a traditional assessment. This probability was supported by another ANOVA with subjects as the independent variable and essays as the continuous, dependent variable. It revealed that the mean score for English was the highest, (M = 4.07, SD = .9) followed by Social Sciences (M = 3.94, SD = 1.06). Additionally, the teachers of English had significant differences in the reported frequency of use from all the other subject areas except Social Sciences and Modern Languages. Predictably, the teachers of Mathematics reported using essays with the least frequency, which was significantly different from all the other subject groups. Since only the teachers of English participated in the qualitative phase of the research, subject differences were not explored in this phase.

3.2.2. *Subject differences for alternative assessment tools and strategies*

The difference in teachers' reported frequency of use of the alternative assessment tools and strategies based on subject was also analysed quantitatively using a one-way between-groups ANOVA with a post-hoc test. The assumption of normality had been violated (p = .002 on the K-S test). However, as Elliott and Woodward (2007) and Pallant (2013) stipulated, when the sample size is greater than 30 or 40, parametric tests can be used even if there is a violation of the assumptions of normality. Therefore, I proceeded with the ANOVA since the sample was 1088.

The results of the ANOVA revealed that there was a significant difference (p < 0.001). A subsequent examination of the post hoc test results revealed significant differences between teachers of English and teachers of Mathematics and Science at a confidence interval of 0.05. The teachers of English reported using alternative assessment tools and strategies more frequently than teachers of Mathematics and Science (M = 2.75, SD .56) for teachers of English

and *M* = 2.26 and 2.49, *SD* = .63 and .60 for teachers of Mathematics and Science, respectively. Significant differences were also found between teachers of Mathematics and all the other subject groups except Science and Modern Languages. The teachers of Mathematics reportedly used alternative assessment tools and strategies less frequently than all the other subject groups, including Science and Modern Languages. This meant that the teachers of Mathematics reported that they used alternative assessment tools and strategies with the least frequency. A mean score of 2.27 out of 5 meant that, on average, the teachers of Mathematics reported that they sometimes used the alternative assessment tools and strategies on the instrument. Apart from the significant difference between teachers of Social Science and Mathematics discussed earlier, there were also significant differences between the teachers of Social Sciences (*M* = 2.84, *SD* .63) and the teachers of the Science (*M* = 2.49, *SD* .60), Practical Arts (*M* = 2.58, *SD* .57) and Modern Languages (*M* = 2.48, *SD* = .58). The teachers of Social Sciences reported using alternative assessment tools and strategies more frequently than these other subject areas as well. The teachers of the sciences and the teachers of subjects categorized as the performing arts differed significantly as well, with the performing arts teachers reporting a higher frequency of use. This significant difference is in addition to the significant differences found between English and Social Sciences teachers reported earlier. Business differed significantly from Mathematics and Performing Arts. While the reported frequency of use by business teachers was higher than that of teachers of Mathematics, it was lower than that reported by the performing arts teachers. The performing arts teachers reported the highest frequency of use of alternative assessment tools and strategies (*M* = 3.03, *SD* = .69). However, while this is the highest, it is much lower than the highest mean score for the reported frequency of use of traditional assessment (*M* = 4.07, *SD* = .64 for Social Sciences). It is also lower than the lowest mean score for reported frequency of use of traditional assessment tools and strategies (*M* = 3.54, *SD* = .56) for Mathematics.

In continuing, the results also showed that performing arts teachers reported using alternative assessment tools and strategies significantly more frequently than practical arts (*M* = 2.58, *SD* = .58), modern languages (*M* = 2.48, *SD* = .58) and mixed teacher (*M* = 2.61, *SD* = .65). This is in addition to all the other subjects discussed earlier (Mathematics, Business, Science and Practical Arts). Finally, the teachers who taught more than one category of subjects (Mixed) differed from the teachers of Mathematics and the performing arts, as was discussed earlier. They reported using alternative assessment tools and strategies more frequently than teachers of Mathematics but less frequently than the performing arts teachers. The effect size was moderate at .09, which indicated that the differences were not by chance.

3.3. Different School Type, Same Assessment Tools and Strategies, Same Assessment Policy

A two-way between-group analysis of variance was used to determine if there were differences in teachers' reported frequency of use of the different types of assessment tools and strategies based on school type. This technique was suitable because, in this study, school type referred to the type of school (traditional, technical, and upgraded) as well as the rank of the school (above average, average, below average). All the assumptions except normality and homogeneity of variance for Frequency of Use of Traditional Assessment Tools and Strategies (FUTATS) were met. However, since Elliott and Woodward (2007) and Pallant (2013) purport that with a larger sample, the assumption of normality is frequently violated, and ANOVA is robust to violations of the assumption of normality and "reasonably robust" to violations of the assumption of homogeneity of variance (Pallant, 2013, p. 204), I continued with the ANOVA.

The results of the ANOVA showed that the interaction effect was not significant (p = .74). There was also no significant difference in FUTATS based on school type or school rank (p = .20 and .27, respectively). There was also no significant difference in FUAATS based on school

type or rank (p = .64 for SchoolType*SchoolRank, .72 for School Rank, and .29 for School Type. Therefore, the quantitative analyses revealed no significant difference in teachers' frequency of use of either traditional or alternative assessment methods based on school type.

3.3.1. *Qualitative explanations of the absence of significant difference based on school type*

Based on the quantitative findings, I used the subsequent qualitative phase to explain why there was no difference in the frequency of use of traditional and alternative assessment tools and strategies based on school type. This quantitative finding was surprising given the grave disparities in student academic ability, infrastructural development and support, parental and alumni support, and teacher qualification among the different types of schools: traditional, upgraded, and technical high schools. When I analysed contextual data in the qualitative phase, I observed that traditional high schools benefited from better infrastructural development, alumni, and parental support and had teachers with higher qualifications. They also had more well-behaved students with higher overall academic achievement and achievement in English. For example, each classroom in the top-performing traditional high school was outfitted with projectors and HDMI connections for technology integration, while there were insufficient classrooms and, desks and chairs in the low-performing upgraded and technical high schools. Additionally, while all the teachers in the traditional high schools had a degree in English Language Education and some a master's degree, some of the teachers in the technical and upgraded high schools only had teaching diplomas. Some of the teachers in the below-average upgraded high school were trained to teach at the primary level and not to teach English. (For an extended discussion, see Williams-McBean 2021). Therefore, I wanted to find out why there was no difference in teachers' frequency of use of traditional and alternative methods despite the contextual differences. The data revealed similarities in the schools' assessment policies that led teachers to select traditional assessment tools and strategies more frequently than alternative assessment tools and strategies. These similarities include mandatory, standardised testing and a quota of grades.

3.3.1.1. Mandatory, Standardised Testing Led to Greater Use of Traditional Assessment Tools and Strategies. In all the participating schools, the schools' assessment policies propelled teachers into using traditional assessment tools and strategies by stipulating mandatory tests and essays. All the teachers reported that their schools' assessment policy required that teachers administer monthly or six weekly tests in addition to end-of-term and end-of-year examinations, which are usually standardised pen-and-paper tests. Even when not specified, the administrators' negative attitude to other assessment tools and strategies propelled teachers to use written assessments (tests and essays). This negative attitude is typified in the explanation provided by Mrs. Moody, from the below-average traditional high school, as to why she used written tests most frequently. She explained,

> I used to like doing a lot of drama first time …. But it's difficult now because the push is about the homework, the classwork, the test. It is more now of an academic institution right throughout, instead of making the students whole. I think the culture of the school is dying and where we can be creative that is basically taken away because when a drama presentation with students was suggested as the graded test for grade nine, it was shunned by the Head of Department and administrators. (Interview with Mrs. Moody)

Another example was seen when Mrs. Black from the above-average traditional high school shared that her school's assessment policy stipulated that teacher's term assessment classwork or homework "must include at least one essay and one comprehension task." The school administrators stated or expressed a preference for written tasks influenced teachers to select traditional assessment tools and strategies more frequently. Since this preference was evident in all the schools in the study, it partially explained why there was no statistically significant

difference in teachers' frequency of use of traditional or alternative assessment tools and strategies based on school type.

3.3.1.2. Higher Quota of Grades Led to Greater Use of Traditional Assessment Tools and Strategies. Schools that require a quota of grades from teachers also influenced teachers to use predominantly traditional assessment tools and strategies, despite school type. For accountability purposes, in each school, each teacher was required to input a set number of grades into the school's grading system per month, six-week period, or term (see Table 5).

Table 5. *The quota of grades required for each school.*

School Name (pseudonyms)	School Type & Rank	Number of Grades Required per Term		Type of Grades
		Language	Literature	
Sunnydale High School	Traditional Above Average	14		Three tests, two classwork, and two homework per subject
		7	7	
James Stewart High School	Traditional Average	12		One homework or classwork and one test per subject every six weeks.
		6	6	
Harrison High School	Traditional Below Average	12		One homework or classwork and one test per subject every six weeks.
		6	6	
Roaring River High School	Upgraded Above Average	36		Two homework, two classwork, one test and one affective every six weeks.
		18	18	
Willow High School	Upgraded Below Average	6		One homework, one classwork, one test.
		3	3	
Hill Top High School	Technical Above Average	24		One homework, one classwork, one test per month
		12	12	
Northside High School	Technical Below Average	4		Midterm and end of term exams.
		2	2	

These grades usually come from classwork, homework, and tests and were sent to parents on report cards. While the number of grades varied in each school, ranging from four to 36 per term for three terms (Christmas, Easter, and Summer terms), the impact of the quota requirement was similar in most of the schools. The more grades required, the higher the likelihood of teachers assessing students using traditional assessment tools and strategies.

When I asked the participants how the school's assessment policy impacted their choice of assessment, most of them explained that the required number of grades led them to use traditional assessment tools and strategies. These traditional assessments were primarily selected-response items with one correct answer because they were easier to mark. In that way, they could meet their grade quota more easily. This impact was most evident in Roaring River High School, which had the highest required number of grades per term (36). The explanation was typified in the response from Mrs. Turner. She explained:

> It has a lot of influence on it [her classroom assessment practices] because I teach so many classes, and I have so many grades to give in for the month. What I do is I plan some assessments that are not time-consuming to mark, especially for literature. So, what happens is that it is not as meaningful as I would like it to be. Because when I would give them like an essay or something, or have them do some extended writing, with the number of grades … If I have to give in five pieces of grades for literature, I have to give the students some questions based on the chapter and

> I give them like one to ten and so on, and they use a couple minutes and answer those questions, short answer questions. Or I give them something that is multiple choice if I have like a paper that I set before – a past paper that is multiple choice. I give them like from one to a certain number and have them answer the questions and in quick time I finish marking it and I give them a grade.

This teacher gave the students easy-to-mark assignments just to get a grade, and because traditional assessments are easier to mark and less time-consuming, they would be used more frequently.

Another teacher, Ms. Hall, also from Roaring River, explained how she changed from using activities that focused on the students' ability to speak in English to written pieces that were easy to mark to meet the quota of grades. I watched Miss Hall give her students a test comprised of 10 short answer questions on a chapter from the novel the class was reading in Literature class. I asked her why she decided to use a written test. She responded:

> Let me just say something. What I'm accustomed to in the classroom … my focus was mainly on learning. Well, that's what I believe, teaching and learning are the focus, right? So, before I came here, we spent more time teaching the concepts and evaluating the students on actually understanding the concepts. And evaluation didn't mean like four pieces or five pieces for the month. It would probably be like three pieces over a six-week period or something like that, so it was not that frequent. So, the whole speaking aspect of it came into play because then I had them speak more. They had the chance to take part more and not be afraid that I was going to mark every piece of work they did. That's what I'm used to. That's the kind of environment that I'm used to. So, this (*She points to the test paper*.) is a shock to me, and so I'm gradually getting accustomed to it. That is all I can tell you.

From the excerpt, it is evident that the other school in which Miss Hall taught (that was not included in this qualitative phase of this study) also required a quota of grades from the teachers. It is also evident that using traditional assessment tools and strategies becomes more likely when the quota is higher. The higher the quota of grades required, the higher the marking load and the less time the teachers have to focus on alternative assessments that take more time to administer and score.

While the impact of the quota of grades was most evident in Roaring River High School, it was evident in the other schools. In most schools, the teachers found the school's assessment policy "challenging" because of the amount of marking required or because of the frequency of the assessment coupled with the large class sizes. As Mrs. Peart from Sunnydale High School explained:

> Sometimes it is challenging to ensure that you have the number of pieces because you must have two classwork pieces as well, and I think two homework pieces for both language and literature, so it takes a lot. It's a lot of marking. (Interview with Mrs. Peart)

Ms. Khan from Hill Top High School gave a similar explanation: "I think some [classes] probably have like forty-six or so. I think the lowest number is forty-five. Yeah, so you can just imagine having all those books to mark, and all those assignments".

Ms. Hunter from Harrison High School also explained:

> So alright, the term starts in September. Six weeks take us to mid-October, and I teach, and I test at that time. It's going to take me to — and I have to mark all of those pieces. While marking, I must still be teaching, and still, I have to be setting another set of six-week work again. The testing time is too much! (Interview with Ms. Hunter)

In essence, the stipulated grades caused the teachers to view the policies as challenging because to ensure they met their grade quota, the teacher had to be marking students' work much more frequently while teaching and engaging in other school activities. The challenge was also associated with large class sizes, as seen in the excerpt taken from Ms. Khan's interview. To overcome this challenge, many of the teachers used selected-response items.

The challenge of the grade quota system was evident in all the schools except Willow High School (the below-average upgraded high school), where there was no formal assessment policy and three grades were required per subject per term. At Willow High School, most teachers of English were required to submit three grades per term for English Language because English Literature was only taught to the top-streamed class in each grade. In this school, the teachers' preference for written assessment was primarily influenced by the format of the internal exams (End of Term and End of Year) and national assessments (CSEC, City & Guilds). The teachers taught to prepare students for these assessments. Therefore, they tested using similar formats (primarily tests and essays) but introduced projects after a project-based school-based assessment was added to the CSEC English examinations. This explanation is exemplified in the excerpt taken from the interview with Mrs. Downs. She explained:

> Sometimes you give them homework and projects. We try to give them at least one project per term so that they can get used to it, especially for the SBAs. Because we're having a problem with them at grades ten and eleven when they are to do the SBAs, we are trying from grade seven to say, okay you must do projects, and we're going to teach you skills for doing projects, so we trying to do that.

In sum, the assessment policies in the schools that participated in the qualitative phase of the research were largely similar in requiring or expressing a preference for traditional assessment tools and strategies and specifying a quota of grades that the teachers had to supply per month, six weeks, or term. These policy requirements influenced teachers of English to use traditional assessment tools and strategies more frequently because they were more manageable and less time-consuming to administer and score. The higher the grade quota, the more likely teachers would use selected-response and short-answer questions to assess students. The focus of assessment became to provide grades rather than to assess students' learning meaningfully. Tests (consisting of selected-response items only) were easier to mark, save teachers time, and ensured they met their grade quota. This largely accounted for the absence of differences across school types. The absence of difference in teachers' frequency of use of traditional assessment tools and strategies was also due to the format of internal and external summative examinations. Since those were primarily traditional, the teachers used traditional assessment formats as well. However, efforts were made to introduce projects since it was introduced as a part of the secondary exit English examinations offered by the Caribbean Examinations Council.

4. DISCUSSION of THE FINDINGS, IMPLICATIONS, and RECOMMENDATIONS

The findings of both the quantitative and qualitative phases of this mixed methods study confirmed the findings of previous studies that secondary school teachers primarily used traditional assessment tools and strategies. Among the traditional assessment methods, pen-and-paper tests which primarily included selected-response items, were most frequently used. Previous international researchers also reported the dominance of testing (see, for example, Acar-Erdol & Yıldızlı, 2018; Berry, 2010; Brookhart, 2013; Dandis, 2013; Esomonu1 & Eleje, 2020; Guskey & Link, 2019; OECD, 2019; Saefurrohman, 2017; Vlachou, 2018). The same was reported in the lone local study conducted by Onyefulu (2018). This dominance has persisted despite pedagogical shifts, curricular rewrites, and increased advocacy for the greater use of alternative assessment tools and strategies. Since classroom assessment is primarily supposed to be used to improve teaching and learning (Acar-Erdol & Yıldızlı, 2018) and that improvement can be increased by using alternative assessment tools and strategies (Berry 2010; Black & Wiliam, 1998; Koh, 2017; McMillan, 2014), there is need for research on why teachers continue to use traditional assessment tools and strategies with far greater frequency.

The explanations provided by the teachers of English who participated in the qualitative phase of this research provided some useful insights. The teachers primarily used tests to assess their students despite variation in students' academic ability, infrastructure which allowed for

innovations in assessment, teacher qualification, and parental support because the school's assessment policies required or expressed more positive attitudes towards traditional tests. They also used selected-response and short answer tests to meet the school administration's quota of grades per month, six weeks, or term. The higher the grade quota, the more frequently these tests were used, even if the administrators allowed teachers to choose the assessment format. Finally, the teachers used traditional tools and strategies more frequently because they modelled internal and external, standardised, summative examinations, which primarily used written examinations. However, as the format of these examinations changed, for example, to include school-based assessments, teachers included alternative assessments (i.e., projects). Other studies have also reported that the format of external, standardised assessment has influenced teachers to select and create and use traditional tests in the classroom (Berry, 2010; McMillan, 2003; Ong, n.d.).

Consequently, changes in assessment must be accompanied by policy changes at the school level to allow teachers time to administer and score alternative assessment tools and strategies. School administrators must also demonstrate more positive attitudes towards alternative assessment tools and strategies in practice and reduce the required number of grades. The focus on grades should be replaced with a focus on learning, from the summative use of assessment to the formative use of assessment. However, in the absence of supportive school-level attitudes and practices, researchers must focus on how traditional assessment tools and strategies can be created and used to improve learning (i.e., formatively) and not just for grading (i.e., summatively). Empirical studies on best practices related to the formative use of traditional and summative tests and their impact on students' learning are also needed to improve educational outcomes. Additional research should also be done to find out if the explanations provided by the teachers of English hold true for teachers of other subjects.

This study also showed that teachers' frequency of use of traditional and alternative assessment tools and strategies differed significantly based on subject: English, Mathematics, Sciences, Social Sciences, Business, Practical Arts, Performing Arts, Modern Languages, and Mixed. There were many differences among the groups that were discussed in this paper. Most notable were that teachers of Mathematics reported using both traditional and alternative assessment tools less frequently than teachers of all other subject groups, and the teachers of Social Sciences and English reported the highest use of traditional assessment tools and strategies. The result for the teachers of Social Sciences and English is arguably because 'essays' which are frequently used as an assessment tool by English teachers, were classified as traditional assessments. Some writers classify essays as traditional assessments (Dikli, 2003; Gronlund, 2006, Koh, 2017) while others do not (Frey & Schmitt, 2010; Wren & Gareis, 2019). In this study, essays were classified as traditional assessment primarily because it is popular on the external, standardised examinations offered by the CXC for secondary schools in the Caribbean. In this study, teachers of English and the Social Sciences reported that they used essays with a significantly higher frequency than all other subjects, which largely accounted for their significantly greater use of traditional assessment tools and strategies. In contrast, the teachers of the Performing Arts used alternative assessment tools and strategies with the highest frequency, and the teachers of English also reported using alternative assessment tools and strategies more frequently than teachers of Mathematics and Science.

There have been conflicting reports from previous studies on differences based on subject, with some researchers reporting significant differences (Alkharusi, 2011; Berry, 2010; Bol et al., 1998; Dandis, 2013; Duncan & Noonan, 2007; Zhang & Burry-Stock, 2003) and others reporting finding no significant difference (Duncan & Noonan, 2007; Ong, n.d.). There are contradictions among those who previously reported significant differences as well. Some researchers reported that teachers of Mathematics indicated that they used alternative

assessment methods with greater frequency than all other subject areas (Bol et al., 1998) or more than teachers of language arts, science and social studies (Zhang & Burry-Stock, 2003), while others reported that teachers of Mathematics use predominantly traditional assessment tools and strategies (Dandis 2013; Watt, 2005). The findings of the latter group of researchers were confirmed in this study. The findings of previous studies that reported that teachers of English and Social Studies used paper-pencil tests and constructed-response items including essays (Berry, 2010, Brookhart, 2009; Zhang & Burry-Stock, 2003) was also confirmed in this study.

The explanations provided by the teachers of English in the qualitative phase of this study, which were previously discussed, provide some insights as to why teachers of English used tests so frequently. However, since this qualitative exploration was not done with the teachers of other subjects in this study, future studies could provide said qualitative explanations. Even the qualitative explanations provided by the teachers of English in this study should be explored in other contexts as what obtains in one region, country, school, or classroom may differ from another. Diverse contextual issues not identified in these schools and Jamaica may become evident in future studies. It is through identifying and responding to these issues can we hope to improve teachers' assessment practices and improve teaching and learning through the formative use of assessment.

Acknowledgments

I wish to thank all the schools and teachers who participated in this project and allowed me to shed some light on assessment in Jamaican secondary schools.

Declaration of Conflicting Interests and Ethics

The author 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.

Orcid

Clavia T. Wlliams-McBean https://orcid.org/0000-0003-3434-8913

REFERENCES

Acar-Erdol, T., & Yıldızlı, H. (2018). Classroom Assessment Practices of Teachers in Türkiye. *International Journal of Instruction*, *11*(3), 587-602. https://doi.org/10.12973/iji.2018.11340a

Adeyemi, B. (2015). The efficacy of authentic assessment and portfolio assessment in the learning of social studies in junior secondary schools in Osun state, Nigeria. *IFE Psychologia: An International Journal*, *23*(2), 125-132.

Alkharusi, H. (2011). Teachers' classroom assessment skills: Influence of gender, subject area, grade level, teaching experience and in-service assessment training. *Journal of Turkish Science Education*, *8*(2), 39-48).

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

Berry, R. (2008). *Assessment for learning*. Hong Kong University Press.

Berry, R. (2010). Teachers' orientations towards selecting assessment strategies. *New Horizons in Education*, *58*(1), 96-107.

Bland, L.M., & Gareis, C.R. (2018). Performance assessments: A review of definitions, quality characteristics, and outcomes associated with their use in K-12 schools. *Teacher Educators' Journal*, *11*, 52-69.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Bol, L., Stephenson, P.L., O'Connell, A.A., & Nunnery, J.A. (1998). Influence of experience, grade level, and subject area on teachers' assessment practices. *The Journal of Educational Research*, *91*, 323–330. https://doi.org/10.1080/00220679809597562

Bramwell-Lalor, S. (2019) Assessment for learning on sustainable development. In: Leal Filho W. (eds) *Encyclopedia of sustainability in higher education*. Springer, Cham. https://doi.org/10.1007/978-3-319-63951-2_1-1

Brookhart, S.M. (2009). Assessment and examinations. In L.J. Sasha & A.G. Dworkin (Eds.), *International handbook of research on teachers and teaching*. Springer Science & Business Media.

Brookhart, S.M. (2013). Comprehensive assessment systems in service of learning: Getting the balance right. In R. W. Lissitz (Ed.), *Informing the practice of teaching using formative and interim assessment: A systems approach* (pp. 165–184). Information Age Publishing.

Buhagiar, M. (2007). Classroom assessment within the alternative assessment paradigm: Revisiting the territory. *Curriculum Journal*, *18*(1), 39-56. https://doi.org/10.1080/09585170701292174

Burke, K. (2009). *How to assess authentic learning*. Corwin Press.

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative data analysis*. Sage.

Cobern, W.W., & Adams, B.A. (2020). Establishing survey validity: A practical guide. *International Journal of Assessment Tools in Education*, 7(3), 404-419. https://doi.org/doi.org/10.21449/ijate.781366

Clarke, M.G. (2011). Rescue upgraded high schools – Gov't must address inequities in education sector. Retrieved from http://jamaicagleaner.com/gleaner/20110821/cleisure/cleisure2.html

Creswell, J. (2014). *Educational research: Planning, conducting and evaluating quantitative and qualitative research* (4th ed.). Pearson Education Limited.

Dandis, M.A. (2013). The assessment methods that are used in a secondary mathematics class. *Journal for Educators, Teachers and Trainers*, *4*(2), 133–143.

DiCicco-Bloom, B., & Crabtree, B.F. (2006). The qualitative research interview. *Medical Education*, *40(*4), 314–321. https://doi.org/10.1111/j.1365-2929.2006.02418.x

Dikli, S. (2003). Assessment at a distance: Traditional vs. alternative assessments. *The Turkish Online Journal of Educational Technology*, *2*(3), 13–19.

Dogan, M. (2011). Student teachers' views about assessment and evaluation methods in mathematics. *Educational Research and Reviews*, *6*(5), 417–431.

Duncan, C.R., & Noonan, B. (2007). Factors Affecting Teachers' Grading and Assessment Practices. *The Alberta Journal of Educational Research*, *53*(1), 1–21.

Esomonu, N.P., & Eleje, L.I. (2020). Effect of diagnostic testing on students' achievement in secondary school quantitative economics. *World Journal of Education, 10*(3), 178-187. https://doi.org/10.5430/wje.v10n3p178

Fraenkel, J.R., & Wallen, N.E. (2003). *How to design and evaluate research in education* (5th ed.). McGraw-Hill.

Gronlund, N.E. (2006). *Assessment of student achievement* (8th ed.). Pearson.

Guha, R., Wagner, T., Darling-Hammond, L., Taylor, T., & Curtis, D. (2018). *The promise of performance assessments: Innovations in high school learning and college admission*. Learning Policy Institute.

Guskey, T.R., & Link, L.J. (2019). Exploring the factors teachers consider in determining students' grades. *Assessment in Education: Principles, Policy & Practice*, *26*(3), 303-320.

Hess, K., Colby, R., & Joseph, D. (2020). *Deeper competency-based learning: Making equitable, student-centered, sustainable shifts*. Corwin.

Jiang, Y. (2020). Teacher classroom questioning practice and assessment literacy: Case studies of four English Language teachers in Chinese universities. *Frontiers in Education*, *5*(23), 1-17. https://doi.org/10.3389/feduc.2020.00023

Johnson, B., & Turner, L.A. (2003). Data collection strategies in mixed methods research. In A. Tashakkori & C. Teddie (Eds.), *Handbook of mixed methods in social and behavioural research* (pp. 297–319). Sage.

Koh, K. (2017). Authentic assessment. *Oxford Research Encyclopedia of Education*. Retrieved from https://oxfordre.com/education/view/10.1093/acrefore/9780190264093.001.0001/acrefore-9780190264093-e-22.

Kubiszyn, T., & Borich, G. (2013). *Educational testing and measurement: Classroom application and practice* (10th ed.). John Wiley & Sons, Inc.

Marshall, C., & Rossman, G.B. (2016). *Designing qualitative research* (5th ed.). Sage.

McMillan, J.H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice*, *20*(1), 20-32. https://doi.org/10.1111/j.1745-3992.2001.tb00055.x

McMillan, J.H. (2003). Understanding and improving teachers' classroom assessment decision-making: Implications for theory and practice. *Educational Measurement: Issues and Practice*, *22*(4), 34–43. https://doi.org/10.1111/j.1745-3992.2003.tb00142.x

McMillan, J.H. (2014). *Classroom assessment: Principles and practice for effective standards-based instruction* (6th ed.). Pearson.

Miller, D., Linn, R., & Gronlund, N. (2013). *Measurement and assessment in teaching* (11th ed.). Pearson Education: Upper

Monteiro, V., Mata, L., & Santos, N. (2021) Assessment conceptions and practices: Perspectives of primary school teachers and students. *Frontiers in Education*, *6*, 631185. https://doi.org/10.3389/feduc.2021.631185

National Education Inspectorate [NEI]. (2013, November). Chief inspector's baseline report. http://www.nei.org.jm/Portals/0/Content/Documents/Chief%20Inspector's%20Report%20November%202013.pdf?ver=2015-04-08-111059-667

National Education Inspectorate [NEI]. (2014, June). Chief inspector's baseline report. http://www.nei.org.jm/Portals/0/Content/Documents/Chief%20Inspector's%20Report-%20June%202014%20Final.pdf

National Education Inspectorate [NEI]. (2015, September). Chief inspector's baseline report. http://www.nei.org.jm/Portals/0/Chief%20Inspector's%20Baseline%20Report%202015.pdf?ver=2015-09-30-125548-787

National Education Inspectorate [NEI] (2016). Chief inspector's report. http://www.nei.org.jm/Portals/0/Content/Documents/C2R1%20Chief%20Inspector's%20Report%202016%20Final.pdf?ver=2018-04-19-115528-887

National Education Inspectorate [NEI] (2017). Chief inspector's report. https://www.nei.org.jm/Portals/0/Content/Documents/Chief%20Inspector's%20Report%202017.pdf?ver=2018-11-30-102446-537&ver=2018-11-30-102446-537

Oluwatayo, J.A. (2012). Validity and reliability issues in educational research. *Journal of educational and social research*, *2*(2), 391–400.

Ong, S.L. (n.d.). *Profiling Classroom Teachers Assessment Practice*. Retrieved from www.iaea.info/documents/paper_4d32f2cd.pdf

Onyefulu, C. (2018). Assessment practices of teachers in selected primary and secondary schools in Jamaica. *Open Access Library Journal*, *5*(12), 1-25. https://doi.org/10.4236/oalib.1105038

Organisation for Economic Cooperation and Development (OECD) (2019). *OECD reviews of evaluation and assessment in education: Student assessment in Turkey*. Retrieved from https://www.oecdilibrary.org/sites/1807effcen/index.html?itemId=/content/component/1807effc-en

Popham, J.W. (2005). *Classroom assessment: What teachers need to know*. Pearson Education.

Popham, J.W. (2018). *Assessment literacy for educators in a hurry*. Alexandria. ASCD.

Saefurrohman. (2017). Indonesian EFL teachers' classroom assessment methods in reading. *Advances in Social Science, Education and Humanities Research (ASSEHR), 109.* 4th Asia Pacific Education Conference. https://doi.org/10.2991/aecon-17.2017.40

Saefurrohman, & Balinas, E. (2016). English Teachers Classroom Assessment Practices. *International Journal of Evaluation and Research in Education, 5*(1), 82 – 92.

Saldaña, J. (2016). *The coding manual for qualitative researchers* (3rd ed.). Sage.

Sewagegn, A.A. (2019). A study on the assessment methods and experiences of teachers at an Ethiopian university. *International Journal of Instruction, 12*(2), 605-622. https://doi.org/10.29333/iji.2019.12238a

Statistical Institute of Jamaica (2017). Education Statistics. Retrieved from https://statinja.gov.jm/Demo_SocialStats/Education.aspx

Stiggins, R.J, & Conklin, N.F. (1992*). In teachers' hands: Investigating the practices of classroom assessment*. State University of New York Press.

Taber, K.S. (2018). The use of cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*(6), 1273-1296. https://doi.org/10.1007/s11165-016-9602-2

Thomas, D.R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation, 27*(2), 237-246. https://doi.org/10.1177/1098214005283748

Vlachou, M. (2018). Classroom assessment practices in middle school science lessons: A study among Greek science teachers. *Cogent Education, 5*(1), Article: 1455633. https://doi.org/10.1080/2331186X.2018.1455633

Williams-McBean, C. (2021). Contextual considerations: Revision of the Wiliam and Thompson (2007) formative assessment framework in the Jamaican context. *The Qualitative Report, 26*(9), 2943-2969. https://doi.org/10.46743/2160-3715/2021.4800

Wren, D., & Gareis, C.R. (2019). *Assessing deeper learning: Developing, implementing, and scoring performance tasks.* Rowman & Littlefield.

Yin, R.K. (2014). *Case study research design and methods* (5th ed.). Sage.

Zhang, Z., & Burry-Stock, J.A. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. *Applied Measurement in Education, 16*(4), 323–342.

Zohrabi, M. (2013). Mixed method research: Instruments, validity, reliability and reporting the findings. *Theory and Practice in Language Studies, 3*(2), 254-263. http://dx.doi.org/10.4304/tpls.3.2.254-262

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 4, 906–930
https://doi.org/10.21449/ijate.1109800
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

An observational look at classroom practices in the Turkish language teaching process

Mustafa Koroglu[1,*], **Ahmet Balci**[2]

[1]Hatay Mustafa Kemal University, Faculty of Education, Department of Turkish Language, Hatay, Türkiye
[2]Hatay Mustafa Kemal University, Faculty of Education, Department of Turkish Language, Hatay, Türkiye

ARTICLE HISTORY
Received: Apr. 27, 2022
Revised: Nov. 1, 2022
Accepted: Nov. 10, 2022

Keywords:
Reading,
Comprehension,
Turkish teacher,
Observation,
Teaching understanding.

Abstract: The aim of this study is to determine the practices of middle school 8th grade Turkish teachers towards comprehension (reading) teaching in the process of learning-teaching Turkish lessons and how long they allocate for these practices. The model of the research is the case study model, which is one of the qualitative research methods. The participants of the study consist of five Turkish teachers who gave eighth grade Turkish lessons in the 2019-2020 academic year and participated in the study voluntarily. In the study, an "Observation Form" developed by the researcher was used as the data collection tools. Literature review was used in the development of measurement tools, expert opinion was obtained, and the level of harmony between coders was examined. Descriptive statistical techniques (frequency, percentage, average, etc.) were used to analyze the data. In the study, each of the teachers wrote five to the poem text "Kaldırımlar" (Sidewalks) in the theme "Individual and Society", the informative text "Gündelik Hayatımızda E-Hastalıklar" (E-Diseases in Our Daily Life) in the "Science and Technology" theme and the narrative text "Göç Destanı" (Epic of Migration) under the theme "Our National Culture" lesson time (600 minutes) had been allocated. Accordingly, 75 lesson hours (3000 minutes) of five Turkish teachers were observed in total. As a result, it was seen that 8th grade Turkish lesson teachers, whose teaching process was observed in our study, used only the texts and activities related to the texts while applying comprehension (reading) strategies in the Turkish lesson learning-teaching process.

1. INTRODUCTION

The human being, who lives in the community and stands out with his social presence, takes this most important feature from interpersonal communication. Language is the element that enables people to communicate with those around them. In its simplest definition, language is a tool that provides communication and agreement between people (Ergin, 2003). In the Turkish Dictionary (2011) language is defined as “the agreement people make with words or signs to express their thoughts and feelings, language”. Language is the most important feature distingushing humans from other living things. Language is a human-specific feature that consistof comprehension and expression skills. People develop listening and speaking skills

*Corresponding Author: Mustafa Koroglu koroglumustafa_z@hotmail.com Hatay Mustafa Kemal University, Faculty of Education, Department of Turkish Language Education, Hatay, Türkiye.

e-ISSN: 2148-7456 /© **IJATE** 2022

starting from their early ages while they acquire the ability to read and write later. In this respect, the ability to understand language is important for human life.

Reading by which students can access new information in all processes of education is the most basic means of acquiring knowledge among the four basic language skills and is the transfer of symbols and signs perceived by the eye to the brain and their interpretation and interpretation by the brain. Thus, students acquire different sources of information by accessing various sources thanks to their reading skills. Many definitions of reading have been made, some of which are as follows:

"One of the ways of understanding and acquiring knowledge" (Özbay, 2014, p.10); "A meaningful interpretation of written language" (Haris & Sipay, 1990, p.18); and "An active process in which a person creates new meanings by combining what they already know with what they learn from the text" (Güneş, 2013, p.13).

As can be understood from these definitions, it is understood that the ultimate goal of reading is comprehension since reading has been seen as understanding and making sense and it has been accepted as linking people's previous knowledge with newly learned information. As reading is defined as the process of perceiving and interpreting words and sentences or a text as a whole (Temizkan, 2009), the purpose of reading is for the reader to comprehend the text, make sense of it, and make connections with the text (Pressley & Allington, 2015).

Comprehension is defined as perceiving the message that the text and the speaker want to say (Göğüş, 1978). According to Özbay (2014), comprehension takes place in a known language and knowing the word !!!read is not sufficient for comprehension. It is stated that the punctuation marks of that language are also necessary for understanding. The RAND Reading Working Group (2002) defined reading comprehension as the process of simultaneously extracting and constructing through interaction and participation with written language, and stated that it consists of three elements: reader, text and reading purpose.

Teaching reading comprehension strategies is defined as one of the five focal points of literacy programs, and reading for understanding is at the center of all educational reading programs. (National Reading Panel, 2000). This involves teaching and applying strategies that develop students' ability to extract meaning from what they read (Pressley, 2006; Rand Reading Study Group, 2002). The act of reading is meaningless if students can decode words, read fluently, but cannot make sense of what is being read. Comprehension education should therefore be an integral component of teaching reading.

Although many innovations have been made in the field of education in our country in recent years, it is seen that the reading comprehension levels of the students are still below the expected level. This result is also seen in the scores obtained from the national and international exams. The correct answer average of the 40-question Turkish course questions, mostly consisting of reading comprehension questions, in the 2019 higher education institutions entrance exam is 14.67. While this average was 16.18 in the exam in 2018, it was determined as 17.28 in 2017. According to these results, it can be concluded that the correct answer averages in Turkish lessons have been gradually decreasing.

According to Programme for International Student Assessment (PISA), fluent reading is the ability of students to read the given text easily and effectively, that is, to be able to read, analyze and express the text correctly so that they understand the meaning of the text. The process of accessing information was divided into two sub-headings. The first of these is scanning and finding information in the text. Accordingly, although the requested information is included in the text, the person reading the text is required to scan and find this information. Secondly, it is emphasized that the reader should work with more than one text in the process of accessing information, and this process is more needed, especially during the digital reading. As the

reason for this, it was emphasized that the reason for these readers would encounter too many texts in the digital environment and that they should search for and select these texts. Another cognitive process in the reading skills assessment process is comprehension. Comprehension is divided into two sub-processes: expressing literal meaning and combining inferences. Accordingly, it is required that the readers of the text should be able to interpret sentences and short paragraphs and make inferences about what is being in the given text. In the last skill of evaluation and reflection, the readers of the text are asked to evaluate the content, quality and reliability of the text and go beyond the real meaning and inferences in the text. Understanding questions constitute 45% of the questions asked in PISA 2018 (OECD, 2019). Table 1 shows Türkiye's performance in reading skills by years.

Table 1. *Turkey's reading skills performance between PISA 2003-2018.*

Years	Score
PISA 2003	441
PISA 2006	447
PISA 2009	464
PISA 2012	475
PISA 2015	428
PISA 2018	466

In Table 1, it can be seen that the average scores of Türkiye in the field of reading skills varied between 428 and 475 between the years 2003 and 2018, the average score increased from 2003 to 2012, but there was a significant decrease in 2015 and the average score started to rise again as of 2018. According to the preliminary report of the last 2018 PISA, although students' reading skills performance scores increased, this result was below the average of OECD countries. According to the results of national and international exams, it is seen that there are problems in terms of teaching reading and reading comprehension in our country, as in most countries in the world.

Therefore, such results nationally and internationally raised the question, "What happens in the classroom when teaching reading comprehension?". Durkin (1978-79), who made the first landmark study in this field, revealed the lack of understanding teaching in classrooms through the observation of behaviors of the students as well as the teacher's practices in 39 classes in Reading and Social Studies courses in 14 different schools. Durkin (1978-79) stated in his study that teachers do not teach comprehension, however they evaluate using the question strategy, they spend too much time on applications that are not related to comprehension, and they neglect teaching comprehension in their lessons. In many studies conducted in the following years, similar results to the study of Durkin (1978) were obtained (Ateş, 2011; Brevik, 2015, 2017; Dole et al., 1991; Dole et al., 1996; Ness, 2009, Pressley & Allington, 2015; Pressley et al., 2006; Pressley et al., 2007; Taylor, Pearson, Peterson, & Rodriguez, 2003). On the other hand, many researchers working in the field of reading agree that teachers can help their students understand the text while reading (Stahl, Jacobsen, Davis, & Davis, 1989; Taylor et al., 2003). Despite this, many teachers do not implement practices that improve reader's comprehension in their classrooms (Pressley, 2006).

After Durkin's (1978-79) study, more than one strategy was developed by researchers working in the field of reading and many studies were conducted on the effects of these strategies on reading comprehension (Duke & Martin, 2015; Emre, 2014; Epçaçan, 2008; Garner, 1987; Karatay, 2007; Luttenegger, 2012; McCown & Thomason, 2014; McIntyre & Hulan, 2013; Neuman & Gambrell, 2013; Palincsar & Schutz, 2011; Pearson, 2009). However, very little attention has been given to observational studies on what happens in classrooms related to the process of teaching comprehension (reading). The only study conducted in this area in our

country is Ateş (2011) on the teaching process of the 4th grade Turkish lesson. In the literature review conducted by the researcher, it has been determined that such a study has not been carried out at the secondary and higher grade levels in Turkey. In this study, it has been tried to determine what the 8th grade Turkish teachers' practices for teaching comprehension (reading) in the process of learning-teaching Turkish lessons are and how much time they devote to these practices.

2. METHOD

2.1. Research Design

The case study model, which aims to examine the practices of secondary school 8th grade Turkish teachers in teaching comprehension (reading) in the Turkish lesson teaching process. Case studies are defined as "the method in which one or more events, environments, programs, social groups, or other interconnected systems are examined in depth." (Büyüköztürk et al., 2013).

Creswell (2018) defined the case study as a multifaceted study in the qualitative tradition. Yin (2014), on the other hand, defined case studies as identifying and capturing the conditions of a daily situation. A case study is also known as a case study. Case studies have begun to be recognized as a more valid research method today. Flyvbjerg (2006) explains the value of case studies as follows:

A scientific discipline without many in-depth case studies is one without the systematic production of examples, and a discipline without examples is ineffective. Social sciences can be strengthened by conducting many good case studies (Flyvbjerg, 2006, p.221).

Yin (2014) pointed out that the case study is a preferred research method "when examining current events, but when relevant behaviors are not manipulated". In this study, the researcher examined the practices of secondary school 8th grade Turkish teachers in teaching comprehension (reading) in the Turkish lesson teaching process; observed how they teach comprehension (reading) strategies in poetry, informative and narrative text types. These observations helped answer the research questions. Furthermore, it provided guidance on how to implement the current curriculum, which includes poetry, informative and narrative text at the 8th grade level of secondary school.

2.2. Participants

The participants of the research were five Turkish teachers who taught eighth grades in the 2019-2020 academic year and participated in the study voluntarily. The Turkish language teachers participating in the study worked in public schools with the same socio-economic level in the same district in Hatay, Türkiye. Table 2 presents information about the teachers who took part in the study in their own time.

Table 2. *Information about the participants of the study.*

Participants	Working year (Seniority)	Number of students
Teacher A	12	32
Teacher B	8	16
Teacher C	6	25
Teacher D	5	28
Teacher E	5	21

The names of the Turkish teachers participating in the research were coded and given as Teacher A, B, C, D and E. In addition, the names of teachers will be mentioned in this way in the following parts of the study. Detailed information about the teachers participating in the study is presented below:

2.2.1. *Teacher A*

As of the 2019-2020 academic year, Teacher A has been working as a Turkish teacher for 12 years. He stated that he took Turkish lessons in 8th grades throughout his working life. There are 32 students studying in Teacher A's class. The characteristics of the class taught by Teacher A are as follows:

• The teacher's desk is located directly opposite the entrance door of the classroom, on the left side of the classroom according to the seating plan of the students.

• The smart board was hung centered on the wall directly opposite the students.

• It has been observed that the students sit in the classical seating arrangement due to the large class size.

• Teacher A said that since there is a library in the school, there is no library in the classroom.

2.2.2. *Teacher B*

Teacher B has been working as a Turkish teacher for 8 years as of the 2019-2020 academic year. She stated that she attended 8th grade Turkish lessons throughout her working life. 16 students are studying in Teacher B's class. Teacher B stated that she had a master's degree with a thesis regarding her education level. The characteristics of the class taught by Teacher B are as follows:

• The teacher's desk is located directly opposite the entrance door of the classroom, on the left side of the classroom according to the seating plan of the students.

• The smart board was hung centered on the wall directly opposite the students.

• Students sit according to the classical classroom seating plan.

• There is a library on the right and left of the teacher's desk.

• There is also a cassette player in the classroom.

• There are boards on the back and right walls of the classroom. There is a poetry corner for Turkish lessons on the back panel.

2.2.3. *Teacher C*

As of the 2019-2020 academic year, Teacher C has been working as a Turkish teacher for 6 years. She has been taking Turkish lessons for the 8th grades since the last three years of her working life. 25 students are studying in the classroom where the application is made. The characteristics of the class taught by Teacher C are as follows:

• The teacher's desk is located directly opposite the entrance door of the classroom, on the left side of the classroom according to the seating plan of the students.

• The blackboard was hung centered on the wall directly opposite the students.

• Students sit according to the classical classroom seating plan.

• There is a library in the middle of the classroom on the right side.

2.2.4. *Teacher D*

Teacher D has been working as a Turkish teacher for 5 years as of the 2019-2020 academic year. She has been taking Turkish lessons for the 8th grades since the last two years of her working life. 28 students are studying in the classroom where the application is made. The characteristics of the class taught by Teacher D are as follows:

• The teacher's desk is located directly opposite the entrance door of the classroom, on the left side of the classroom according to the seating plan of the students.
• The smart board was hung centered on the wall directly opposite the students.
• Students sit according to the classical classroom seating plan.
• There is a library on the right side of the teacher's desk.

2.2.5. *Teacher E*

Teacher E has been working as a Turkish teacher for 5 years as of the 2019-2020 academic year. She has been taking Turkish lessons for the 8th grades since the last three years of her working life. 21 students are studying in the classroom where the application is made. The characteristics of the class taught by Teacher E are as follows:

• The teacher's desk is located directly opposite the entrance door of the classroom, on the left side of the classroom according to the seating plan of the students.
• The smart board was hung centered on the wall directly opposite the students.
• Students sit according to the classical classroom seating plan.
• There is a library on the right side of the teacher's desk.
• There are boards on the back and right walls of the classroom.

2.3. Data Collection

2.3.1. *Observation Form*

With the semi-structured observation technique, it was aimed to determine the practices of the 8th grade Turkish teachers in secondary school for comprehension (reading) education and the time they allocated to these practices. The applications made by the teacher during the observation were coded into the observation form by the researcher. In addition, the researcher made a sound recording during the observation.

Turkish lessons in our country are carried out through the texts included in the themes in the textbooks. A total of 5 Turkish teachers were observed for 75 lesson hours (3000 minutes) in the study. The process of processing poetry, informative and narrative text types of all teachers participating in the research was observed by the researcher.

An observation form (Appendix 1) was developed to observe the practices of the 8th grade Turkish lesson teachers in the reaching of Turkish lesson comprehension (reading) lesson. The procedures performed during the development phase of the observation form created to have information about the practices carried out by the Turkish lesson teachers are as follows:

a) Research obtained from the literature in the development of this form (Ateş, 2011; Brevik, 2015, 2017; Dole et al., 1991; Dole et al., 1996; Durkin, 1978-1979; Durkin, 1989; Ness, 2009; Pearson, 2010b; Pressley & Allington, 2015; Pearson et al., 2009; Pressley et al., 1998; Pressley et al., 2006; Pressley et al., 2007; Taylor et al., 2003) were examined so that the content validity of the observation form was tried to be ensured.

b) The draft observation form was presented to 10 experts in the field. For the expert opinion, an information note was created that briefly introduces and exemplifies the units of the observation form, and the experts were asked to evaluate the draft observation form accordingly.

c) Based on the expert's opinions, the observation form was given its final form. All these stages were deemed sufficient for the validity of the observation form.

d) The last stage of qualitative data analysis is that of checking the accuracy of the findings. Confirmation of the findings can be achieved by testing the results obtained (Merriam, 1998). In qualitative research, the results obtained after controlling the codes and categories can determine the level of representation of the data included in the analysis (Poggenpoel &

Myburgh, 2003). In order to get rid of the researcher's own influence and to make a coding, it is important that different coders code for the same data set. According to Fidan and Öztürk (2015), it is important that different coders encode the same data set and that this coding has a high similarity rate. The closeness of this similarity ratio is important in determining the reliability of qualitative research. In order to ensure the reliability of the form, the method of "consistency among the evaluators" was used. This method, also called inter-rater agreement, is used to examine the reliability of the scores given by two or more independent observers regarding the degree to which a large number of objects possess a certain feature. It can be said that reliability will increase as the scores given by the observers get closer to each other (Büyüköztürk et al., 2013:114).

For this purpose, the researcher studied the qualities of the observation form with the second observer who is an expert in the field of Turkish teaching. For the reliability study of the content analysis codes in the evaluation of the observation form, the formula $\Delta = C \div (C + \partial) \times 100$ developed by Miles and Huberman (1994) to determine the reliability level between the coders in qualitative studies was used. In the formula, Δ: Reliability coefficient, C: Number of subjects/terms on which consensus is reached, ∂: Number of subjects/terms on which there is no consensus. According to the coding control, which gives internal consistency, the consensus among coders is expected to be at least 80% (Miles & Huberman, 1994; Patton, 2002).

The reliability result of the observation form (.93) was found by using the reliability formula developed by Miles and Huberman (1994). Based on these results, it was accepted that the agreement between the coders was sufficient in the observation form.

2.4. Data Analysis

The case study is a multidimensional research. As the type of research in the data analysis process, the researcher has followed several consecutive steps from specific to general in qualitative data analysis. These steps are as follows:

Step-1: Observation data were arranged and prepared for analysis. This step is the stage of recording the voice recordings of teacher observations on the computer, writing, categorizing and classifying the notes kept in the field.

Step 2: All data were read and analyzed by the researcher and an expert in his field. This step allowed us to reveal the general structure of the research. It gave the researcher an idea about which comprehension (reading) methods the participants used and what their applications were in this subject.

Step 3: The researcher started to encode all the data he collected into the observation form. At this stage, the audio recordings were also printed in written form and the data was organized by marking the words representing a comprehension (reading) category.

Step 4: The researcher created themes for teaching comprehension (reading) and explained the information to be encoded in these themes. Themes in this category; It has been examined under two headings as comprehension-related and non-comprehension categories.

Step 5: In this last step, the researcher has revealed the value of his original work by interpreting the coded form of the data obtained as a result of his observations. He interpreted what the data he obtained meant, what strategies were used by 8th grade Turkish teachers in teaching comprehension (reading), and how much time he spent on these strategies.

3. RESULTS

The researchers observed teachers' (Teachers A, B, C, D and E) applications of Kaldırımlar (poetry), Gündelik Hayatımızda E-Hastalıklar (Informative), and Göç Destanı (Narrative) and obtained data as to their classroom practices, the course hours when they performed the practices, and the time they allocated for such practices.

3.1. Findings About to Teacher A's In-Class Practices

Under this title, Teacher A's in-class practices, the course hours he performed the practices while he was teaching the texts "Kaldırımlar" (poetry), "Gündelik Hayatımızda E-Hastalıklar" (informative) and "Göç Destanı" (narrative) in the Turkish lesson teaching process, and how they were taught. There are findings about the time allotted.

The findings regarding the classroom practices of Teacher A in the teaching-learning process of the Turkish lesson of the poem "Kaldırımlar" and the time allocated to these practices are presented in Table 3:

Table 3. *In-class practices of Teacher A's poem "Kaldırımlar" and the time allocated to these practices.*

In-class activities	Time	Lesson hours
Taking attendance before starting the lesson, introductory speech	5 minutes	Lesson 1
Preparatory work (activating prior knowledge)		
The 1st preparatory study question in the textbook (the teacher's reading of poems about expatriation)	20 minutes	
Informing the teacher about the author	5 minutes	
Teacher listening to the poem in N. Fazıl Kısakürek's own voice	5 minutes	
Teacher's vocalization of the poem	5 minutes	
Student reading the poem	5 minutes	Lesson 2
Asking inference questions about poetry by the teacher	10 minutes	
The teacher gives short information about the poet and reads examples from her other poems.	10 minutes	
Student on duty coming to the classroom and making an announcement, Students' complaints about each other, talking about other lessons, talking about the practice exam, talking about football	15 minutes	
Rereading the poem Kaldırımlar by another student	5 minutes	Lesson 3
Activity 1 in the textbook (Vocabulary Teaching)	10 minutes	
Activity 2 in the textbook (Questions about poetry)	15 minutes	
The teacher's re-information about the poet	6 minutes	
Extracurricular conversations about football	4 minutes	
Taking attendance, speaking on a subject outside the class, opening the smart board.	5 minutes	Lesson 4
Don't talk about the Kara Tren folk song about expatriate	5 minutes	
The teacher gives information about the poet	5 minutes	
3rd activity in the textbook (Theme and main emotion of the poem)	10 minutes	
Talking about extracurriculars and exams	5 minutes	
4th activity item a in the textbook (verbal arts)	10 minutes	
Teacher entering the lesson, preparing lesson material	3 minutes	Lesson 5
Solving multiple choice exam questions for LGS exam	37 minutes	
Total	200 minutes	5 Lessons

Table 3, displays that Teacher A was observed during 5 lesson hours (200 minutes) while lecturing the poem "Kaldırımlar" in the Turkish lesson learning-teaching process. In the process of learning-teaching the poem "Kaldırımlar" Teacher A spent the most time on the 1st activity in the textbook (37 minutes) to solve the multiple-choice exam questions for LGS in the 5th lesson hour and to activate the students' prior knowledge in the 1st lesson hour (20 minutes). When the 8th grade Ministry of Education (MEB) Publications Turkish textbook is examined, it is seen that there are seven activities related to the poem "Kaldırımlar".

Table 4. *In-class practices and the time allocated to these practices by Teacher A in the text "Gündelik Hayatımızda E-Hastalıklar".*

In-class activities	Time	Lesson hours
Attendance and extracurricular speaking before starting the class	5 minutes	Lesson 1
Preparatory work in the textbook (activating prior knowledge-talking about technology addiction and technology usage time)	20 minutes	
Estimating the content of the text based on the images in the text and the title of the text	10 minutes	
The time given for the activities (preparatory work) in the textbook in the 1st lesson	5 minutes	
Reading the text by students (silent-aloud)	15 minutes	Lesson 2
The teacher interrupts and explains while the text is being read.	20 minutes	
Time given for the 1st activity in the textbook (Vocabulary Teaching)	5 minutes	
Speech	5 minutes	Lesson 3
Activity 1 in the textbook (Vocabulary Teaching-unknown words)	10 minutes	
Activity 2 in the textbook (Questions about the text)	25 minutes	
Extracurricular speech	5 minutes	Lesson 4
The time given to the students for the 3rd activity in the textbook	6 minutes	
Activity 3 in the textbook (Inference questions)	10 minutes	
The time given for the 4th activity in the textbook	5 minutes	
Activity 4 in the textbook (Determination of words and phrases)	15 minutes	
Talking about an extracurricular topic	4 minutes	
Attendance, speaking on an extracurricular topic	4 minutes	Lesson 5
Time given for the 5th activity in the textbook	5 minutes	
5th activity in the textbook (determining the text type)	14 minutes	
The time given for the 6th activity in the textbook	5 minutes	
6th activity in the textbook (visual-graphic interpretation)	12 minutes	
Total	200 minutes	5 Lessons

According to Table 4, teacher A was observed for a total of 5 lesson hours (200 minutes) while he was processing the text "Gündelik Hayatımızda E-Hastalıklar" in the Turkish lesson learning-teaching process. According to Table 4, Teacher A spent the most time in the process of learning-teaching the text "Gündelik Hayatımızda E-Hastalıklar", in the 3rd lesson hour, on questions about the text, which is the second activity in the textbook (25 minutes), and on the preparation in the textbook in the 1st lesson hour (20 minutes) and making explanations (20 minutes) while reading the text in the 2nd lesson hour.

Table 5. *In-class practices and the time allocated to these practices by Teacher A in the text of "Göç Destanı".*

In-class activities	Time	Lesson hours
Attendance, speaking before starting the lesson	5 minutes	Lesson 1
Preparatory work in the textbook	18 minutes	
Silent reading of the text by the students	7 minutes	
Reading aloud by students	10 minutes	
Studies of summarizing the text	10 minutes	Lesson 2
The time given for the 1st activity in the textbook	6 minutes	
Activity 1 in the textbook (vocabulary teaching)	20 minutes	
Talking about an extracurricular topic	4 minutes	
Reading aloud by students	10 minutes	Lesson 3
Silent reading of the text by the students	10 minutes	
The time given for the second activity in the textbook	10 minutes	
2nd activity in the textbook (questions about the text) (for lack of time)	10 minutes	

Table 5. *Continued*

Class start, attendance	5 minutes	
Activity 2 in the textbook (questions about the text)	10 minutes	
The time given for the 3rd activity in the textbook	5 minutes	Lesson 4
3rd activity in the textbook (topic-main idea)	10 minutes	
homework, extracurricular speaking	5 minutes	
Starting the lesson, preparing for the lesson	4 minutes	
The time given for the 4th activity in the textbook	6 minutes	
Activity 4 in the textbook (detection of real and fictional elements)	12 minutes	
The time given for the 5th activity in the textbook	5 minutes	Lesson 5
5th activity in the textbook (activity related to the text type)	10 minutes	
Assigning homework for the next lesson	3 minutes	
Total	200 minutes	5 Lessons

According to Table 5, teacher A was observed for a total of 5 lesson hours (200 minutes) while he was processing the text "Göç Destanı" in the Turkish lesson learning-teaching process. It is understood that in the learning-teaching process of this narrative text type, Teacher A spent the most time for the 1st activity in the textbook (20 minutes) in the 2nd lesson hour and the preparatory studies (18 minutes) in the 1st lesson hour.

3.2. Findings About Teacher B's In-Class Practices

Under this title, Teacher B's in-class practices and practices in the text processing process of "Kaldırımlar" (poetry), "Gündelik Hayatımızda E-Hastalıklar" (informative) and "Göç Destanı" (narrative) in the Turkish lesson teaching process, and at what time There are findings regarding the time devoted to the implementations. The findings regarding the classroom practices of Teacher B in the teaching-learning process of the Turkish lesson of the poem "Kaldırımlar" and the time allocated to these practices are presented in Table 6.

Table 6. *In-class practices of Teacher B's poem "Kaldırımlar" and the time allocated to these practices.*

In-class activities	Time	Lesson hours
Attendance before starting the class, extracurricular speaking	4 minutes	
Homework (for the next lesson)	4 minutes	
The teacher gives brief information about the type of poetry.	2 minutes	
Having the teacher listen to the composed version of the poem	10 minutes	Lesson 1
Talking about how the composed poem makes students feel	5 minutes	
Before starting to read the poem, the time given to underline the places where the words given in the first activity are used in the poem.	5 minutes	
Reading the poem aloud (each student read a stanza)	10 minutes	
The teacher's reading of the poem	8 minutes	
Activity 1 in the textbook (word meaning-prediction)	15 minutes	Lesson 2
Giving time to do the 2nd activity in the textbook	4 minutes	
Activity 2 in the textbook (questions about poetry)	13 minutes	
Preparation to start the lesson	5 minutes	
3rd activity in the textbook (detection of subject-main emotion)	10 minutes	
The time given for the 4th activity in the textbook	4 minutes	Lesson 3
4th activity item a in the textbook (verbal arts)	15 minutes	
4th activity item b in the textbook (comment on the contribution of rhetoric to the meaning)	6 minutes	
Preparation for starting the lesson, speaking on extracurricular topics	3 minutes	
5th activity item a in the textbook (talk about urban life, modernization and neighborhood)	10 minutes	Lesson 4
Solving multiple choice questions to prepare for LGS exams	27 minutes	
Teacher coming to class, speaking on an extracurricular subject	4 minutes	Lesson 5
Solving multiple choice questions to prepare for LGS exams	36 minutes	
Total	200 minutes	5 Lessons

According to Table 6, while Teacher B was processing the text of the poem "Kaldırımlar" in the Turkish lesson learning-teaching process, it was observed for a total of 5 lesson hours (200 minutes). It is seen that Teacher B spends the most time (63 minutes) on multiple choice questions to prepare for the LGS exams in the 5th and 4th lesson hours in the course of the "Kaldırımlar" text.

Table 7. *In-class practices and the time allocated to these practices by Teacher B in the text "Gündelik Hayatımızda E-Hastalıklar".*

In-class activities	Time	Lesson hours
Attendance, speaking on extracurricular topics, checking the class library before starting the class	10 minutes	
Preparatory work in the textbook (activating prior knowledge) (talking about technology addiction and technology use time)	20 minutes	Lesson 1
Estimating the content of the text based on the images in the text and the title of the text	10 minutes	
Text reading and reading (silent-aloud)	23 minutes	
The time given for the 1st activity in the textbook (Vocabulary Teaching)	5 minutes	Lesson 2
Activity 1 in the textbook (Vocabulary Teaching-unknown words)	12 minutes	
Speaking on extracurricular topics	5 minutes	
The time given for the 2nd activity in the textbook	7 minutes	Lesson 3
Activity 2 in the textbook (Questions about the text)	28 minutes	
Speaking on extracurricular issues, preparing for the lesson, extracting materials	5 minutes	
The time given to the students for the 3rd activity in the textbook	5 minutes	
Activity 3 in the textbook (inference questions)	10 minutes	Lesson 4
The time given for the 4th activity in the textbook	8 minutes	
Activity 4 in the textbook (detection of words and phrases)	12 minutes	
Speaking on extracurricular topics	4 minutes	
Course preparation, preparation of course materials	5 minutes	
Announcement by the student on duty and speaking about the announcement	5 minutes	
The time given for the 5th activity in the textbook	3 minutes	Lesson 5
5th activity in the textbook (determining the text type)	10 minutes	
The time given for the 6th activity in the textbook	7 minutes	
6th activity in the textbook (visual-graphic interpretation)	10 minutes	
Total	200 minutes	5 Lessons

According to Table 7, while teacher B was processing the text "Gündelik Hayatımızda E-Hastalıklar" in the Turkish lesson learning-teaching process, it was observed for a total of 5 lesson hours (200 minutes). It is seen that Teacher B spends the most time in the text processing process for text reading and reading (23 minutes) in the 2nd lesson and for the preparation work in the textbook (20 minutes) in the 1st lesson.

According to Table 8, Teacher B was observed for a total of 5 lesson hours (200 minutes) while he was processing the "Göç Destanı" text in the Turkish lesson learning-teaching process. It is seen that in the process of processing the "Göç Destanı" text, Teacher B devoted the most time to questions about the text in the textbook in the 4th lesson (25 minutes) and to the preparatory work in the textbook (20 minutes) in the 1st lesson.

Table 8. *In-class practices and the time allocated to these practices by Teacher B in the text of "Göç Destanı".*

In-class activities	Time	Lesson hours
Preparation of materials, attendance, waiting for the smart board to be opened to start the lesson	8 minutes	Lesson 1
Preparatory work in the textbook	20 minutes	
Students reading the text silently	12 minutes	
Students reading the text aloud	11 minutes	Lesson 2
Students summarizing the text	9 minutes	
The time given for the 1st activity in the textbook	12 minutes	
Activity 1 in the textbook (half)-(vocabulary teaching)	8 minutes	
Course preparation, attendance, preparing course material	5 minutes	Lesson 3
Teacher reading the text aloud	10 minutes	
Students reading the text silently	12 minutes	
Activity 1 (half) in the textbook- (vocabulary teaching)	13 minutes	
Lesson start preparation	5 minutes	Lesson 4
The time given for the 2nd activity in the textbook	10 minutes	
Activity 2 in the textbook (questions about the text)	25 minutes	
Starting the lesson, preparing for the lesson	5 minutes	Lesson 5
The time given for the 3rd activity in the textbook	4 minutes	
3rd activity in the textbook (topic-main idea)	10 minutes	
The time given for the 4th activity in the textbook	4 minutes	
Activity 4 in the textbook (detection of real and fictional elements)	10 minutes	
The time given for the 5th activity in the textbook	3 minutes	
5th activity in the textbook (activity related to the text type)	4 minutes	
Total	200 minutes	5 Lessons

3.3. Findings About Teacher C's In-Class Practices

Under this title, while teaching the texts “Kaldırımlar” (poetry), "Gündelik Hayatımızda E-Hastalıklar" (informative) and “Göç Destanı” (storyteller) in the Turkish lesson teaching process, teacher C, in-class applications, the lesson time he performed the applications and the information on these applications. There are findings about the time he spends.

The findings regarding the classroom practices of Teacher C in the learning-teaching process of the Turkish lesson of the poem "Kaldırımlar" and the time allotted to these practices are presented in Table 9:

Table 9. *In-class practices of Teacher C's poem "Kaldırımlar" and the time allocated to these practices.*

In-class activities	Time	Lesson hours
Attendance before starting the lesson, speaking on extracurricular issues, waiting for the smart board to open	10 minutes	Lesson 1
Question about the theme and subject to be covered, giving information about the author	10 minutes	
Student reading the poem	7 minutes	
Talking about unknown words in the poem	13 minutes	
Playing a recorded voiceover of the poem	6 minutes	Lesson 2
Guess the unknown words in the poem	5 minutes	
Speaking on extracurricular topics	5 minutes	
The time given for the 1st activity in the textbook	5 minutes	
Activity 1 in the textbook (Vocabulary Teaching)	13 minutes	
The time given for the second activity in the textbook	6 minutes	

Table 9. *Continued*

Activity 2 in the textbook (questions about poetry)	15 minutes	Lesson 3
The time given for the 3rd activity in the textbook	5 minutes	
Explaining the difference between main emotion and main idea in poetry	3 minutes	
3rd activity in the textbook (Theme and main emotion of the poem)	12 minutes	
The time given for the 4th activity in the textbook	5 minutes	
Giving information about the arts of speech	10 minutes	Lesson 4
4th activity item a in the textbook (verbal arts)	10 minutes	
4th activity item b in the textbook (contribution of rhetoric to expression)	10 minutes	
Assigning homework for the next lesson	5 minutes	
Talking about extracurricular topics, talking about the exam	5 minutes	
Teacher coming to class, attendance	4 minutes	Lesson 5
Solving multiple choice questions to prepare for LGS exams	36 minutes	
Total	200 minutes	5 Lessons

According to Table 9, Teacher C was observed for a total of 5 lesson hours (200 minutes) while he was processing the text of the poem "Kaldırımlar" in the Turkish lesson learning-teaching process. In the process of teaching the poem "Kaldırımlar" by Teacher C, the most time was spent on solving multiple-choice questions (36 minutes) to prepare for LGS exams in the 5th lesson and answering questions about the text, which is the 2nd activity in the textbook, in the 3rd lesson (15 minutes). It appears to be separated.

Table 10. *In-class practices and the time allocated to these practices by Teacher C in the text "Gündelik Hayatımızda E-Hastalıklar".*

In-class activities	Time	Lesson hours
Teacher entering the class, preparing to start the lesson, talking about an extracurricular subject, attendance	10 minutes	Lesson 1
Introducing the subject and theme to be covered, opening the textbooks	3 minutes	
Preparatory work in the textbook	15 minutes	
Informing the teacher about current e-diseases	12 minutes	
Teacher entering the lesson, starting the lesson	4 minutes	Lesson 2
The teacher asks the students to guess the content of the text and the mutual guesses are spoken (based on the pictures and the title)	12 minutes	
Reading the text aloud by the teacher	11 minutes	
Making students read the text paragraph by paragraph	13 minutes	
Teacher coming to class, material preparation, attendance	5 minutes	Lesson 3
Silent reading of the text by the students	8 minutes	
Detection of unknown / incomprehensible words by students	10 minutes	
The time given for the 1st activity in the textbook	5 minutes	
Activity 1 in the textbook (Vocabulary Teaching-unknown words)	12 minutes	
Teacher coming to class, talking about extracurricular	3 minutes	Lesson 4
The time given for the second activity in the textbook	5 minutes	
Activity 2 in the textbook (Questions about the text)	20 minutes	
Time given to do the 3rd activity in the textbook	4 minutes	
Activity 3 in the textbook (Inference questions)	8 minutes	
The teacher's arrival in the classroom, preparation for the lesson	4 minutes	Lesson 5
The time given for the 4th activity in the textbook	3 minutes	
Activity 4 in the textbook (Determination of words and phrases)	6 minutes	
The time given for the 5th activity in the textbook	2 minutes	
5th activity in the textbook (determining the text type)	3 minutes	
The time given for the 6th activity in the textbook	4 minutes	
6th activity in the textbook (visual-graphic interpretation)	10 minutes	
10th activity in the textbook grammar subject processing (subject-verb)	8 minutes	
Total	200 minutes	5 Lessons

According to Table 10, while Teacher C was processing the text "Gündelik Hayatımızda E-Hastalıklar" in the Turkish lesson learning-teaching process, it was observed for a total of 5 lesson hours (200 minutes). It is seen that Teacher C spends the most time for the 2nd activities (20 minutes) in the textbook in the 4th lesson and the preparatory work (15 minutes) in the 1st lesson in the process of processing the text "Gündelik Hayatımızda E-Hastalıklar".

Table 11. *In-class practices and the time allocated to these practices by Teacher C in the text of "Göç Destanı".*

In-class activities	Time	Lesson hours
The teacher's coming to the classroom, attendance, preparation of the course material	10 minutes	Lesson 1
The teacher tells the students about the subject to be covered and the page in the textbook.	4 minutes	
Explain the meaning of the word millet and epic before starting the text	5 minutes	
Predicting the content of the text based on the visuals and title of the text	8 minutes	
Silent reading of the text by the students	13 minutes	
Teacher entering the class, filling the class notebook	3 minutes	Lesson 2
Reading the text aloud by the students (by split reading method)	12 minutes	
The teacher informs the students about the text type	8 minutes	
The study of determining the keywords of the text	7 minutes	
Summing up the text by the students	10 minutes	
Preparation to start the lesson, the teacher coming to the class	3 minutes	Lesson 3
Giving time to do the 1st activity in the textbook	5 minutes	
Activity 1 in the textbook (vocabulary teaching)	18 minutes	
Giving time to do the 2nd activity in the textbook	5 minutes	
2nd activity in the textbook (first 2 questions) (questions about poetry)	9 minutes	
The teacher's arrival in the classroom, the start of the lesson	3 minutes	Lesson 4
Activity 2 in the textbook (continued) (questions about poetry)	25 minutes	
The time given for the 3rd activity in the textbook	4 minutes	
3rd activity in the textbook (topic-main idea)	8 minutes	
Teacher entering the class, attendance	4 minutes	Lesson 5
The time given for the 4th activity in the textbook	5 minutes	
Activity 4 in the textbook (detection of real and fictional elements)	12 minutes	
The time given for the 5th activity in the textbook	4 minutes	
5th activity in the textbook (activity related to the text type)	6 minutes	
Assigning homework for the next lesson	5 minutes	
Talking about an extracurricular topic	4 minutes	
Total	200 minutes	5 Lessons

According to Table 11, teacher C was observed for a total of 5 lesson hours (200 minutes) while lecturing the text "Göç Destanı" in the Turkish lesson learning-teaching process. In the process of processing the text of "Göç Destanı", it is seen that Teacher C spent the most time for the 2nd activity (25 minutes) in the textbook in the 4th lesson and the silent reading activity (13 minutes) of the students in the 2nd lesson hour.

3.4. Findings About Teacher D's Classroom Practices

Under this heading, Teacher D's classroom practices and the time he devoted to these practices while he was teaching the texts “Kaldırımlar” (poetry), "Gündelik Hayatımızda E-Hastalıklar" (informative) and “Göç Destanı” (storyteller) in the Turkish lesson teaching process are included.

The findings regarding the classroom practices of Teacher D in the teaching-learning process of the Turkish lesson "Kaldırımlar" and the time allocated to these practices are presented in Table 12:

Table 12. *In-class practices of Teacher D's poem "Kaldırımlar" and the time allocated to these practices.*

In-class activities	Time	Lesson hours
Teacher entering the class, taking attendance before starting the lesson, speaking about extracurricular	10 minutes	Lesson 1
Teacher playing the recorded composition of the poem	5 minutes	
Teacher's explanation about expatriate	5 minutes	
The teacher asks the students questions about the visuals in the poem	10 minutes	
Silent reading of the poem by students	10 minutes	
Teacher entering the lesson, starting the lesson	4 minutes	Lesson 2
Reading the poem aloud (students reading by sharing the quatrains)	7 minutes	
Identifying keywords	8 minutes	
Teacher reading the poem aloud	6 minutes	
Giving time to do the 1st activity in the textbook	4 minutes	
Activity 1 in the textbook (Vocabulary Teaching)	11 minutes	
The teacher's arrival in the classroom, the start of the lesson	4 minutes	Lesson 3
Reading aloud by students	6 minutes	
Giving time to do the 2nd activity in the textbook	5 minutes	
Activity 2 in the textbook (Questions about poetry)	15 minutes	
Giving time for the 3rd activity in the textbook	4 minutes	
3rd activity in the textbook (Theme and main emotion of the poem)	6 minutes	
Teacher entering the lesson, starting the lesson	3 minutes	Lesson 4
Giving time for the 4th activity in the textbook	5 minutes	
4th activity item a in the textbook (verbal arts)	10 minutes	
4th activity item b in the textbook (contribution of rhetoric to expression)	5 minutes	
Grammar topic (elements of the sentence)	17 minutes	
The teacher enters the lesson, starts the lesson	4 minutes	Lesson 5
Solving multiple choice questions to prepare for LGS exams	36 minutes	
Total	200 minutes	5 Lessons

According to Table 12, Teacher D was observed for a total of 5 lesson hours (200 minutes) while teaching the poem "Kaldırımılar" in the Turkish lesson learning-teaching process. It is seen that in the process of teaching the poem "Kaldırımılar", Teacher D spends the most time on solving multiple-choice questions (36 minutes) in preparation for the LGS exam in the 5th lesson and on the grammar subject (elements of the sentence) in the 4th lesson (17 minutes).

Table 13. *In-class practices and the time allocated to these practices by Teacher D in the text "Gündelik Hayatımızda E-Hastalıklar".*

In-class activities	Time	Lesson hours
Teacher entering the class, taking attendance before starting the lesson, speaking about extracurricular	10 minutes	Lesson 1
Talking about technology and addiction	7 minutes	
Teacher talking about the topic to be covered	3 minutes	
Preparatory work in the textbook	15 minutes	
Informing the teacher about technology diseases	5 minutes	

Table 13. *Continued*

The teacher's arrival in the classroom, the start of the lesson	2 minutes	Lesson 2
Making guesses about the content of the text based on the images in the text and the title of the text	10 minutes	
Silent reading of the text by the students	10 minutes	
Writing unknown words in the text on the board	4 minutes	
Reading the text aloud	10 minutes	
Identifying keywords	4 minutes	
Arrival of the teacher, starting the lesson	5 minutes	Lesson 3
Silent reading of the text	10 minutes	
Summing up the text by the students	5 minutes	
Giving students time to do the 1st activity in the textbook	5 minutes	
Activity 1 in the textbook (Vocabulary Teaching-unknown words)	10 minutes	
Giving students time to do the 2nd activity in the textbook	5 minutes	
Teacher coming to class, talking about extracurricular	3 minutes	Lesson 4
Activity 2 in the textbook (questions about the text)	17 minutes	
Giving time to do the 3rd activity in the textbook	4 minutes	
Activity 3 in the textbook (Inference questions)	8 minutes	
Giving time to do the 4th activity in the textbook	3 minutes	
Activity 4 in the textbook (Determination of words and phrases)	6 minutes	
The teacher's arrival in the classroom, the preparation to start the lesson	3 minutes	Lesson 5
5th activity in the textbook (determining the text type)	5 minutes	
Giving time to do the 6th activity in the textbook	4 minutes	
6th activity in the textbook (visual-graphic interpretation)	6 minutes	
Conversations about where it falls in Turkish grammar (transition between activities)	2 minutes	
Solving multiple choice questions to prepare for LGS exam	20 minutes	
Total	200 minutes	5 Lessons

According to Table 13, teacher D was observed for a total of 5 lesson hours (200 minutes) while processing the text "Gündelik Hayatımızda E-Hastalıklar" in the Turkish lesson learning-teaching process. In the process of teaching the text "Gündelik Hayatımızda E-Hastalıklar", Teacher E spent the most time on solving multiple-choice questions (20 minutes) to prepare for the LGS exam in the 5th lesson and a question-answer activity related to the text, which is the 2nd activity in the textbook, in the 4th lesson (17 minutes).

Table 14. *In-class practices and the time allocated to these practices by Teacher D in the text of "Göç Destanı".*

In-class activities	Time	Lesson hours
Teacher's arrival, attendance, preparation of course material	10 minutes	Lesson 1
Informing the teacher about the concepts of nation and nationality, introduction to the subject	6 minutes	
Preparatory work in the textbook	10 minutes	
Giving information about epic	5 minutes	
Silent reading of the text by students	9 minutes	
Teacher entering the class, preparing material	3 minutes	Lesson 2
Reading the text aloud by the students (Unknown words in the text were asked to be underlined and the text was divided into parts and read)	13 minutes	
Giving time to do the 1st activity in the textbook	5 minutes	
Activity 1 in the textbook (vocabulary teaching)	10 minutes	
Silent reading of the text by students	9 minutes	

Table 14. *Continued*

Teacher entering the class, attendance	3 minutes	Lesson 3
Giving time to do the 2nd activity in the textbook	5 minutes	
Activity 2 in the textbook (questions about the text)	18 minutes	
Giving time to do the 3rd activity in the textbook	5 minutes	
The 3rd activity in the textbook (detection of the subject-main idea)	9 minutes	
Teacher entering the class, preparing material	5 minutes	Lesson 4
Silent reading of the text by students	10 minutes	
Giving time to do the 4th activity in the textbook	5 minutes	
Activity 4 in the textbook (detection of real and fictional elements)	8 minutes	
Giving time to do the 5th activity in the textbook	4 minutes	
5th activity in the textbook (activity related to the text type)	8 minutes	
Teacher entering the class, preparing material	4 minutes	Lesson 5
Solving multiple choice questions for grammar teaching (verb topic) and preparation for LGS exam	36 minutes	
Total	200 minutes	5 Lessons

According to Table 14, Teacher D was observed for a total of 5 lesson hours (200 minutes) while lecturing the "Göç Destanı" text in the Turkish lesson learning-teaching process. In the process of teaching the text of "Göç Destanı", Teacher D spent the most time on grammar teaching (verb in verb) and solving multiple-choice questions (36 minutes) for preparation for the LGS exam in the 5th lesson, and with the text, which is the 2nd activity in the textbook, in the 3rd lesson. It is seen that he allocates (18 minutes) to the relevant question-answer activity.

3.5. Findings about Teacher E's classroom practices

Under this title, Teacher E's classroom practices and the findings of the time he devoted to these practices while he was teaching the texts "Kaldırımlar" (poetry), "Gündelik Hayatımızda E-Hastalıklar" (informative) and "Göç Destanı" (storyteller) in the Turkish lesson teaching process are included.

The findings regarding the classroom practices of Teacher E in the teaching-learning process of the Turkish course "Kaldırımlar" text and the time allocated to these practices are presented in Table 15:

Table 15. *In-class practices of Teacher E's poem "Kaldırımlar" and the time allocated to these practices.*

In-class activities	Time	Lesson hours
The teacher's arrival in the classroom, attendance before starting the lesson, opening the smart board, preparing the course materials	7 minutes	Lesson 1
Giving information about the poem "Kaldırımlar" to be processed and the poet	6 minutes	
Predicting the content of the text based on the visuals in the text of the poem.	5 minutes	
Preparatory work in the textbook	5 minutes	
Silent reading of the poem by students	7 minutes	
Poetry teacher reading aloud (with attention to emphasis and intonation)	5 minutes	
Students reading the poem aloud (each stanza was read by a student)	5 minutes	

Table 15. *Continued*

The teacher's arrival in the classroom, the preparation to start the lesson	4 minutes	Lesson 2
Giving time to do the 1st activity in the textbook	5 minutes	
Activity 1 in the textbook (Vocabulary Teaching)	10 minutes	
Giving time to do the 2nd activity in the textbook	5 minutes	
Activity 2 in the textbook (Questions about poetry)	12 minutes	
Giving homework (Teacher asked students to find and bring poems and songs about expatriate)	4 minutes	
The teacher's coming to the classroom, preparation to start the lesson, taking attendance	5 minutes	Lesson 3
Silent reading of the poem by students	6 minutes	
Reading aloud by the student	5 minutes	
3rd activity in the textbook (detecting the subject and main emotion of the poem)	5 minutes	
Informing the teacher about rhetoric	5 minutes	
Giving time to do the 4th activity in the textbook	4 minutes	
4th activity item a in the textbook (finding the rhetoric)	6 minutes	
4th activity item b in the textbook (contribution of rhetoric to expression)	4 minutes	
The teacher's arrival in the classroom, the preparation to start the lesson	3 minutes	Lesson 4
Grammar teaching	37 minutes	
The teacher's arrival in the classroom, the preparation to start the lesson	5 minutes	Lesson 5
Solving multiple choice questions to prepare for LGS exam	35 minutes	
Total	200 minutes	5 Lessons

According to Table 15, the teacher was observed for a total of 5 lesson hours (200 minutes) while teaching the poem "Kaldırımlar" in the learning-teaching process of the Turkish lesson. It is seen that Teacher E spends the most time on grammar teaching (37 minutes) in the 4th lesson and solving multiple-choice questions (35 minutes) in preparation for the LGS exam in the 5th lesson.

Table 16. *In-class practices and the time allocated to these practices by Teacher E in the text "Gündelik Hayatımızda E-Hastalıklar".*

In-class activities	Time	Lesson hours
Teacher's attendance, attendance, preparation of course materials, speaking about extracurricular	10 minutes	Lesson 1
Preparatory work in the textbook (talk about technology and addiction)	15 minutes	
Estimating the content of the text based on the visuals in the text and the title of the text	15 minutes	
The teacher enters the lesson, the lesson begins	3 minutes	Lesson 2
Silent reading of the text by students	10 minutes	
Reading aloud by students	10 minutes	
Summarizing the text	7 minutes	
Writing keywords on the board and guessing the meanings of unknown words	10 minutes	
Teacher entering the lesson, starting the lesson, attendance	4 minutes	Lesson 3
Silent reading of the text by the students	6 minutes	
Giving time to do the 1st activity in the textbook	4 minutes	
Activity 1 in the textbook (Vocabulary Teaching-unknown words)	8 minutes	
Activity 2 in the textbook (Questions about the text)	18 minutes	

Table 16. *Continued*

The teacher enters the lesson, the lesson begins	4 minutes	Lesson 4
Giving time to do the 3rd activity in the textbook	4 minutes	
Activity 3 in the textbook (Inference questions)	8 minutes	
Giving time to do the 4th activity in the textbook	5 minutes	
Activity 4 in the textbook (Determination of words and phrases)	7 minutes	
5th activity in the textbook (determining the text type)	8 minutes	
Talking about the exam	4 minutes	
The teacher enters the lesson, the lesson begins	3 minutes	Lesson 5
Making the 6th activity in the textbook (visual-graphic interpretation)	8 minutes	
Talking about where to stay on grammar topics	3 minutes	
Solving multiple choice questions to prepare for LGS exam	26 minutes	
Total	200 minutes	5 Lessons

According to Table 16, Teacher E was observed for a total of 5 lesson hours (200 minutes) while he was processing the text "Gündelik Hayatımızda E-Hastalıklar" in the Turkish lesson learning-teaching process. In the process of teaching the text "Gündelik Hayatımızda E-Hastalıklar", Teacher E spent the most time on solving multiple-choice questions (26 minutes) to prepare for the LGS exam in the 5th lesson, the preparatory work activity in the textbook (15 minutes) in the 1st lesson, and the visuals in the text. It is seen that he allocates (15 minutes) to the activity of estimating the content of the text.

Table 17. *In-class practices and the time allocated to these practices by Teacher E in the text of "Göç Destanı".*

In-class activities	Time	Lesson hours
Teacher's attendance, attendance, preparation of course materials, talking to students about extracurricular issues	8 minutes	Lesson 1
Giving information about the "Göç destanı" to be processed, giving information about the epic type	5 minutes	
Preparatory work in the textbook	6 minutes	
Silent reading of the text by students	9 minutes	
Reading aloud by students	8 minutes	
Determining the meanings of unknown words in the text	4 minutes	
The teacher arrives in the classroom and the lesson begins	3 minutes	Lesson 2
Activity 1 in the textbook (vocabulary teaching)	15 minutes	
Giving students time to do the 2nd activity in the textbook	5 minutes	
Activity 2 in the textbook (questions about the text)	17 minutes	
The teacher enters the lesson and the lesson begins	5 minutes	Lesson 3
Silent reading of the text by students	10 minutes	
Carrying out the 3rd activity in the textbook (detection of the subject and main idea)	5 minutes	
Giving time to do the 4th activity in the textbook	4 minutes	
Activity 4 in the textbook (detection of real and fictional elements)	7 minutes	
Giving time to do the 5th activity in the textbook	3 minutes	
5th activity in the textbook (activity related to the text type)	6 minutes	
Teacher entering the class and starting the class and attendance	5 minutes	Lesson 4
Giving homework (related to the 7th Activity)	4 minutes	
Conversation about where you fall in grammar	2 minutes	
Grammar lecture	29 minutes	
Teacher entering the lesson, starting the lesson	3 minutes	Lesson 5
Solving multiple choice questions to prepare for LGS exam	37 minutes	
Total	200 minutes	5 Lessons

According to Table 17, Teacher E was observed for a total of 5 lesson hours (200 minutes) while he was processing the text of "Göç destanı" in the Turkish lesson learning-teaching process. It is seen that Teacher E spends the most time on solving multiple-choice questions (37 minutes) to prepare for the LGS exam in the 5th lesson and grammar lectures in the 4th lesson (29 minutes) in the process of processing the "Göç destanı" text.

Teacher E had his students do only the first five of the nine activities in the textbook during the process of processing the "Göç destanı" text. For teaching comprehension; the practice of activating the prior knowledge (11 minutes), the reading-to-speech practice (27 minutes), the vocabulary teaching activity (15 minutes), the question-answer practice about the text (17 minutes), the practice of determining the subject/main idea of the text (5 minutes) and It has been determined that he allocates time (6 minutes) to the activity related to the detection of the text type.

In addition, according to Table 17, it was observed that Teacher E devoted 29 minutes of the fourth lesson and 37 minutes of the fifth lesson to grammar and multiple-choice problem solving practices for preparation for the LGS exam in the process of processing the "Göç destanı" text.

4. DISCUSSION and CONCLUSION

In this study, it was aimed to examine the practices of 8th grade Turkish teachers in secondary school in teaching comprehension (reading) in the process of learning-teaching Turkish lessons. In this direction, first of all, the classroom practices of the 8th grade Turkish teachers in the Turkish lesson teaching process, the lesson time they performed the applications and the time they allocated for these applications were determined. Each of the teachers spent five lesson hours on the "Kaldırımlar" poem in the "Individual and Society" theme, the informative text “Gündelik Hayatımızdaki E-Hastalıklar” in the "Science and Technology" theme, and the “Göç Destanı” narrative text under the "National Culture" theme. 600 minutes) time. Accordingly, a total of five Turkish teachers were observed for 75 lesson hours (3000 minutes).

Research in the field of reading comprehension strategy education can be divided into intervention and observation studies. Although most of the studies (Boardman et al., 2017; Meyer, Wijekumar, & Lei, 2018; Plonsky, 2011) have focused on the effectiveness of strategy teaching, classroom observation remains an unexplored area (Pearson & Cervetti, 2017).

In recent years, researchers working in the field of reading have examined comprehension instruction in detail and published a list of comprehension strategies that have proven to be effective (Duke & Martin, 2015; Dymock & Nicholson, 2010). The following strategies are included in this the list published by the researchers: Bringing students' prior knowledge into the reading environment, teaching with text structure, practicing for the mental preparation process, and summarizing. In our study, it was observed that the teachers who participated in the practice used these strategies. In addition, it has been observed that teachers use other comprehension (reading) strategies that are not included in this list. An evaluation has been made about the classroom practices of the teachers participating in the study for teaching comprehension (reading), how much time they spare for these practices, and how they perform these practices.

The only study conducted in our country that overlaps with the purpose and results of our study is Ateş's doctoral thesis in 2011. In his study, Ateş (2011) observed five primary school teachers teaching Turkish in the fifth grade for 74 lesson hours (2960 minutes). In Ateş (2011) study, teachers; They concluded that they could not use the teaching time efficiently, that the strategy they used the most was the question-answer strategy, that they did not teach comprehension strategies, and that they conducted their lessons according to the textbook.

Outside of Turkey, studies on teaching comprehension started with Durkin (1978-79). Durkin (1978-79) observed 39 classes in reading and social studies lessons in 14 different schools. While doing this research, he observed the behaviors of the students as well as the teacher practices. In this study, he stated that teachers do not teach comprehension, they evaluate using the question strategy, they spend too much time on practices that are not related to comprehension, and they neglect teaching comprehension in their lessons. Durkin's (1978-79) work formed the basis of many studies on teaching comprehension. Other studies related to our study in the literature are those of Rieckhoff (1997) and Ness (2009).

Rieckhoff (1997) stated that he carried out his study to determine whether there is understanding teaching in classroom practices, as in Durkin's (1978-79) study. He made his observations with 872 minutes of observation in social studies and reading lessons in 20 classes in four different schools. The results of Rieckhoff's (1997) study also overlap with the results of Durkin's (1978-79) study. Rieckhoff (1997) concluded that in the lessons in which he observed 872 minutes, the duration of teaching comprehension was 112 minutes, which corresponds to 12% of the total time. He also stated that this result, which he reached according to his observations, did not reflect the real time for teaching comprehension, and that most of the comprehension practices he observed were related to the evaluation of comprehension.

Ness (2009) made observations for 2400 minutes in her study with teachers attending secondary school science and social studies classes. He concluded that the teachers allocated only 3% to teaching comprehension in their lessons, 12% to non-teaching activities and 12% to uninstructed transitions. Looking at the results obtained from the Ness (2006) study, it is seen that little change has occurred in the classroom practices for teaching comprehension after Durkin's (1978-79) study.

The findings and results obtained from the above studies carried out in different countries and different cultures; are similar to the findings and results of teachers not including teaching comprehension in the teaching process, the question-answer strategy being the most used strategy for teaching comprehension, neglecting comprehension teaching, and giving too much space to non-teaching practices that are not related to comprehension.

As a result, it was seen that the 8th grade Turkish language teachers, whose teaching process was observed in our research, only benefited from the texts in the textbook and activities related to the texts while applying the comprehension (reading) strategies in the Turkish lesson learning-teaching process. Temizkan (2009) emphasized that mother tongue education is done with texts, and this education should be a skill lesson, not a knowledge lesson. It is an indisputable fact that textbooks are the most effective tool for acquiring skills in schools. In the studies carried out, it was concluded that the teachers who carry out the educational activities in Turkey stick to the textbooks (Akyol, 2005; Yalçın, 1996), while the Turkish lessons are carried out according to the textbooks at a rate of 94.44% (Özbay, 2003). All these results show that the texts and text activities in the textbooks should be carefully prepared and selected.

Acknowledgments

This study was produced from the doctoral thesis supported by the Scientific Research Projects Coordination Unit of Hatay Mustafa Kemal University within the scope of the project numbered 17.D.005. This paper was produced from the first author's doctoral 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**: Hatay Mustafa Kemal University/ Social Science Institution, 01-10-2020/08.

Authorship Contribution Statement

Mustafa Koroglu: Investigation, Resources, Methodology, Visualization, Software, Formal Analysis, and Writing -original draft. **Ahmet Balci:** Investigation, Resources, Methodology, Visualization, Software, Formal Analysis, Writing -original draft, Supervision, and Validation.

Orcid

Mustafa Koroglu https://orcid.org/0000-0003-4701-8120
Ahmet Balci https://orcid.org/0000-0002-7424-592X

REFERENCES

Akyol, H. (2005). *Türkçe İlk Okuma Yazma Öğretimi [Turkish primary reading and writing teaching]*. Pegem Akademi.

Ateş, S. (2011). *İlköğretim beşinci sınıf Türkçe dersi öğrenme-öğretme sürecinin anlama öğretimi açısından değerlendirilmesi [Evaluation of fifth-grade Turkish course learning and teaching process in terms of comprehension instruction]* [Unpublished doctoral dissertation]. Gazi University.

Brevik, L.M. (2015). *How teachers teach and readers read. Developing reading comprehension in English in Norwegian upper secondary school.* [Unpublished doctoral dissertation]. University of Oslo.

Brevik, L.M. (2017). Strategies and shoes: Can we ever have enough? Teaching and using reading comprehension strategies in general and vocational programmes. *Scandinavian Journal of Educational Research*, *61*(1), 76-94.

Büyüköztürk, Ş., Çakmak, E.K., Akgün, Ö.E., Karadeniz, Ş., & Demirel, F. (2013). *Bilimsel araştırma yöntemleri [Scientific research methods]* (1. Eds). Pegem Akademi.

Cervetti, G., & Hiebert, E.H. (2015). Knowledge, literacy, and the Common Core. *Language Arts*, *92*(4), 256-269.

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

Creswell, J.W., & Poth, C.N. (2018). *Qualitative inquiry and research design: Choosing among five approaches (4th ed.).* Sage.

Boardman, A.E., Greenberg, D.H., Vining, A.R., & Weimer, D.L. (2017). *Cost-benefit analysis: concepts and practice.* Cambridge University Press.

Dole, J.A., Brown, K.J., & Trathen, W. (1996). The effects of strategy instruction on the comprehension performance of at-risk students. *Reading Research Quarterly*, *31*(1), 62-88. https://doi.org/10.1598/RRQ.31.1.4

Dole, J.A., Duffy, G.G., Roehler, L.R., & Pearson, P.D. (1991). Moving from the old to the new: Research on reading comprehension instruction. *Review of Educational Research*, *61*(2), 239-264. https://doi.org/10.3102/00346543061002239

Duke, N.K., & Martin, N.M. (2015). Best practices in informational text comprehension instruction. In L.B. Gambrell & L.M. Morrow (Eds.), *Best practices in literacy instruction* (pp. 249-267). Guilford.

Durkin, D. (1978-1979). What classroom observations reveal about reading comprehension instruction. *Reading Research Quarterly*, *14*, 481-533.

Durkin, D. (1989). *Teaching Them to Read*. (Fifth Edition). Allyn and Bacon.

Dymock, S., & Nicholson, T. (2010). "High 5!" Strategies to enhance comprehension of expository text. *The Reading Teacher*, *64*(3), 166-178. https://doi.org/10.1598/RT.64.3.2

Rieckhoff, B.S. (1997). *An assessment of current practices in reading comprehension instruction*. Loyola University Chicago.

Emre, Y. (2014). *Farklı akademik seviyedeki 4. sınıf öğrencilerinin okuma stratejilerini kullanma durumları [Utilization of reading strategies among 4th grade students with*

different academic levels] [Unpublished master's thesis]. Kütahya Dumlupınar University.

Epçaçan, C. (2008). *Okuduğunu anlama stratejilerinin bilişsel ve duyuşsal öğrenme ürünlerine etkisi [Effects of reading comprehension strategies on product of cognitive and affective learning]* [Unpublished doctoral dissertation]. Hacettepe University.

Ergin M (1998). *Türk Dil Bilgisi [Turkish Grammar].* Bayrak Publishing.

Fidan, T., & Öztürk, İ. (2015). Perspectives and expectations of union member and non-union member teachers on teacher unions. *Journal of Educational Sciences Research*, *5*(2), pp.191-220.

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative inquiry*, *12*(2), pp.219-245.

Frankel, G., Louizos, C., & Austin, Z. (2012). Canadian educational approaches for the advancement of pharmacy practice. *American journal of pharmaceutical education*, *78*(7). https://doi.org/10.5688/ajpe787143

Garner, R. (1987). *Metacognition and reading comprehension*. Ablex Publishing.

Göğüş, B. (1978). *Orta dereceli okullarımızda Türkçe ve yazın eğitimi [Turkish and literature education in secondary schools]*. Kadıoğlu Publishing.

Güneş, F. (2013). *Türkce ögretimi yaklaşımlar ve modeller [Turkish teaching approaches and models]*. Pegem Akdemi.

Harris, A.J. ve Sipay, E.R. (1990). *How to increase reading ability* (9. Edition). Longman.

Karatay, H. (2007). *İlköğretim Türkçe öğretmeni adaylarının okuduğunu anlama becerileri üzerine alan araştırması [A field study on the reading comprehension skills of elementary school Turkish teacher candidate]* [Unpublished doctoral dissertation]. Gazi University.

Luttenegger, K. (2012). Explicit strategy instruction and metacognition in reading instruction in preservice teachers' elementary school classrooms. *Journal of Reading Education*, *37*(3), 13-20.

McCown, M., & Thomason, G. (2014). Informational text comprehension: Its challenges and how collaborative strategic reading can help. *Reading Improvement*, *51*(2), 237-253.

McIntyre, E., & Hulan, N. (2013). based, culturally responsive reading practice in elementary classrooms: A yearlong study. *Literacy Research and Instruction*, *52*(1), 28-51. https://doi.org/10.1080/19388071.2012.737409

Merriam, S.B. (1998). *Qualitative research and case study applications in education*. Jossey-Bass.

Plonsky, L. (2011). The effectiveness of second language strategy instruction: *A meta-analysis. Language learning*, *61*(4), 993-1038. https://doi.org/10.1111/j.1467-9922.2011.00663.x

Meyer, B.J., Wijekumar, K., & Lei, P. (2018). Comparative signaling generated for expository texts by 4th–8th graders: Variations by text structure strategy instruction, comprehension skill, and signal word. *Reading and Writing*, *31*(9), 1937-1968. https://doi.org/10.1007/s11145-018-9871-4

Miles, M.B., & Huberman, A.M. (1994). *Qualitative data analysis: A sourcebook of new methods.* Sage.

National Reading Panel (US), National Institute of Child Health, & Human Development (US). (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups.* National Institute of Child Health and Human Development, National Institutes of Health.

Ness, M.K. (2009). Reading comprehension strategies in secondary content area classrooms: Teacher use of and attitudes towards reading comprehension instruction. *Reading Horizons, 49*(2), 143-166.

Neuman, S.B., & Gambrell, L.B. (2013). Challenges and opportunities in the implementation of Common Core State Standards. In S.B. Neuman & L.B. Gambrell (Eds.), *Quality reading instruction in the age of Common Core standards* (pp. 1–12). International Reading Association.

OECD (2019). *PISA 2018 results volume I: What students know and can do.* OECD Publishing.

Özbay, M. (2003). *Öğretmen görüşlerine göre ilköğretim okullarında Türkçe öğretimi [Teaching Turkish in Primary Schools According to Teachers' Opinions]*. Gölge Publishing.

Özbay, M. (2014). *Anlama Teknikleri I: Okuma Eğitimi [Comprehension Techniques I: Reading Education]*. Öncü Publishing.

Palincsar, A.S., & Schutz, K.M. (2011). Reconnecting strategy instruction with its theoretical roots. *Theory Into Practice*, *50*(2), 85-92.

Patton, M.Q. (2002). *Qualitative research and evaluation methods (3rd ed.).* Sage.

Pearson, P.D. (2009). The roots of reading comprehension instruction. In S. E. Israel & G. G. Duffy (Eds.), *Handbook of research on reading comprehension* (pp. 3-31). Routledge.

Pearson, P.D. (2010a). Reading First hard to live-or without. *Journal of Literacy Research*, *42*(1), 100-108. https://doi.org/10.1080/10862961003613782

Pearson, P.D. (2010b). The roots of reading comprehension. In K. Ganske & D. Fisher (Eds.), *Comprehension across the curriculum: Perspectives and practices K-12* (pp. 279-321). Guilford.

Poggenpoel, M., & Myburgh, C. (2003). The researcher as research instrument in educational research: A possible threat to trustworthiness? *Education*, *123*(2), 418-421.

Pressley, M. (2006). *What the future of reading research could be.* Paper presented at the International Reading Association Reading Research Conference, Chicago.

Pressley, M., & Allington, R.L. (2015). *Reading instruction that works: The case for balanced teaching.* Guilford.

Pressley, M., Gaskins, Solie, K., & Collins, S. (2006). A portrait of Benchmark School: How a school produces high achievement in students who previously failed. *Journal of Educational Psychology*, *98*, 282-306. https://psycnet.apa.org/doi/10.1037/0022-0663.98.2.282

Pressley, M., Mohan, L., Raphael, L.M., & Fingeret, L. (2007). How does Bennett Woods Elementary School produce such high reading and writing achievement? *Journal of Educational Psychology, 99*(2), 221-240. https://doi.org/l0.1037/0022-0663.92.2.221

Pressley, M., Wharton-McDonald, R., Hampston, J.M., & Echevarria, M. (1998). The nature of literacy instruction in ten grade-4 and-5 classrooms in upstate New York. *Scientific Studies of Reading, 2*. 159-191. https://doi.org/10.1207/s1532799xssr0202_4

RAND, S., & Catherine chair of RAND Reading Study Group. (2002). *Reading for Understanding. Toward an R&D Program in Reading Comprehension.* Santa Monica.

Stahl, S.S., Jacobson, M.G., Davis, C.E., & Davis, R.L. (1989). Prior knowledge and difficult vocabulary in the comprehension of unfamiliar text. *Reading Research Quarterly*, *24*(1), 27-43.

Taylor, B., Pearson, P., Peterson, D., & Rodriguez, M. (2003). Reading growth in highpoverty classrooms: The influence of teacher practices that encourage cognitive engagement in literacy learning. *Elementary School Journal*, *104*(1). https://doi.org/10.1086/499740

Temizkan, M. (2009). *Metin türlerine göre okuma eğitimi [Reading education according to text types]*. Nobel Akademi.

Tierney, W.G., & Lincoln, Y. S. (1994). Teaching qualitative methods in higher education. *The Review of Higher Education*, *17*(2), 107-124.

Yalçın, A. (1996). Türkçe ders kitaplarının planlanması ve yazılması [Planning and writing of Turkish textbooks]. *Journal of Türk Yurdu*, *107*, 24-27.

Yin, R.K. (2014). *Case study research (5th ed.).* Sage.

APPENDIX

Appendix 1. Observation Form

Observed School: Observed Class: Observed Teacher: Class size:		Theme of the Observed Text: Observed Text: Date and time: Course start-end time:	
	Sub Categories	Encodings	Time
Understanding Practices	Practices for the Mental Preparation Process	- Prediction applications based on the title and images of the text - Reading intent building apps - Applications to bring prior knowledge to the reading environment - Applications of detecting keywords in the text	
	Reading-Reading Applications	- Teacher reading aloud - Student reading aloud - Silent reading by students - Students reading by sharing - reading by discussion - reading by marking	
	Practices for understanding the text (Strategy Usage/Teaching)	- Practice for teaching vocabulary - Teaching applications with text structure - Practice for summarizing text - Practice to determine the main idea / main emotion - inference practice - Forecasting practice	
Non-Comprehension Practices	Time without instruction	- This code was used when the teacher did not engage in any teaching behavior.	
	Activities outside of the classroom	- This code was used when the teacher was talking to students on a topic other than Turkish. (For example, when he talks about football, basketball, or an event at school.)	
	Technology-based activities	- In this coding, the teacher uses technology as a teaching resource to expand and reinforce comprehension (reading) education. This code contains technology-based instructions such as internet searches and computer games usage. If the teacher tries to find an activity by opening the smart board or cassette player, this category is coded.	
	Grammar teaching/test solving	- This code was coded when the teacher made applications for the central exam that the students would take at the end of the year, apart from the text that the teacher was teaching.	
	Time given to students for activities	- This code was coded when the teacher gave students extra time to do the activities in the textbook.	

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 4, 931–948
https://doi.org/10.21449/ijate.946609
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

The development and validation of a scale measuring mobile phone use in an academic environment

Nehir Yasan Ak[1,*], **Soner Yildirim**[2]

[1]Akdeniz University, University, Faculty of Social Sciences and Humanities, Department of Management Information Systems, Antalya, Türkiye
[2]Middle East Technical University, Faculty of Education, Department of Computer Education and Information Systems, Ankara, Türkiye

ARTICLE HISTORY

Received: June 01, 2021
Revised: Aug. 19, 2022
Accepted: Nov. 17, 2022

Keywords:
Mobile phone affinity,
Educational mobile phone use,
Scale development,
Smartphone,
Educational technology,
Academic environment.

Abstract: The purpose of this study was to bridge the gap in current research on educational mobile phone use within the framework for the rational analysis of the mobile education (FRAME) model. The paper developed and validated the Mobile Phone Use in Academic Environment Scale (MPUAES) to measure both positive and negative aspects of educational use of mobile phones. The participants were 1887 undergraduate students enrolled in all faculties and grade levels of Middle East Technical University in Ankara, Türkiye. The inclusion criterion for the participation in the study was owning a smartphone. The exploratory and confirmatory factor analyses were run with two different samples. Three factors structure with 18 items were obtained, which were labeled as facilitator, distractor, and connectedness. These three factors explained 63.42% of the total variance. For confirmation of the factor structure, confirmatory factor analysis was performed with the second sample. Cronbach alpha coefficient of each factor ranged between .90 and .74. To conclude, the findings of the study proposed that the scores obtained from the developed scale were valid and reliable in measuring undergraduate students' mobile phone use in an academic environment.

1. INTRODUCTION

The use of mobile phones among college students has increased rapidly in recent years. The "mobility" and "highly customizable" features of the mobile phones enable learners to take control of their own learning and engage in learning activities according to their own needs, interests, and curiosity (Kukulska-Hulme & Shield, 2008). Despite providing such opportunities in learning environments, the opinions on the use of mobile devices in education vary. In other words, there are both proponents and opponents of the educational use of mobile devices in the literature. Correspondingly, Obringer and Coffey (2007) stated "although mobile devices are the central of the students' life in terms of personal and educational purposes, they face inconsistent attitudes among teachers and administrators with regard to use in the school" (p. 43). Bernacki et al. (2020)'s study also showed that mobile technologies can be used to improve learning processes. Additionally, Crompton (2017) refers to supportive role of mobile technologies in terms of

*CONTACT: Nehir YASAN AK ✉ nehiryasanak@akdeniz.edu.tr Akdeniz University, Faculty of Social Sciences and Humanities, Management Information Systems, Antalya, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

collaboration However, opponents consider those devices as disruptive and unsuitable tools in an educational context, which causes a challenge for the universities' adoption and use of mobile device in education (Losh, 2014). Regarding this issue, a study conducted by Purba and Setyarini (2020) found that students encountered some concentration problems while using the mobile application in language learning. Some scholars, on the other hand, hold a more holistic perspective and suggest that mobile devices are both a distractor and facilitator in learning environments (Lockhart, 2016). Quaglia and Corso (2014) have a similar opinion and claim that:

> In this era of prolific use and debate regarding the utility, integration, and efficacy of educational technology devices such as tablets and smartphones, one constant that is frequently missing from the purported ideologies and opinionated inferences is the perspective of the learner or user (p.21).

As Quaglia and Corso (2014) highlighted, there was a need to investigate how undergraduate students use their mobile phones for educational purposes in detail. Thus, this study will shed light on the learner perspective on the use of mobile phones in an academic environment. Furthermore, most of the studies of using mobile phones for educational purposes were conducted by using qualitative analyses in the literature (Ford, 2016; Huang, 2016; Dukic & Chiu, 2015; Gikas & Grant, 2013). On the other hand, when the quantitative studies were examined in the field, it was seen that the majority of them were carried out through acceptance models such as TAM and UTAUT (e.g., Han & Yi, 2019; Bryant, 2016; Cheon et al., , 2012; Abu-Al-Aish & Love, 2013; Pan et al., 2013; Iqbal & Qureshi, 2012; Venkatesh et al., 2012; Lowenthal, 2010; Wang et al., 2009). The present study was an attempt to offer a new measurement approach for the assessment of educational mobile phone use. Thus, the purpose of this study was to develop a valid and reliable instrument measuring both positive and negative aspects of mobile phone use of undergraduate students in the academic environment.

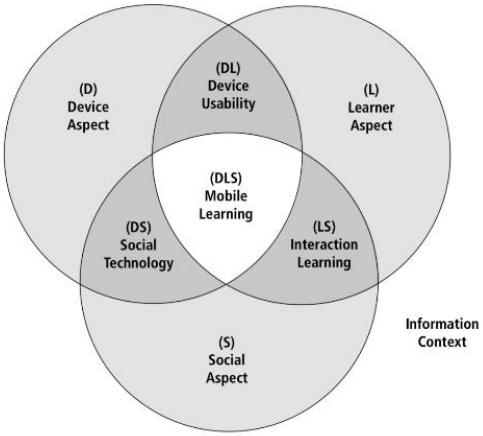
1.1. Mobile Learning

The term "mobile learning" refers to the use of mobile technologies to deliver learning materials to learners (Parsons & Ryu, 2006). Cell phones, smartphones, palmtops, handheld computers, tablet PCs, laptops, and personal media players are typical examples of mobile devices. Since the definition of mobile learning varies among researchers, it is important to clarify how the term is defined in the literature. According to Keegan (2005), mobile learning is "the provision of education and training on smartphones and mobile phones" (p. 3). Similarly, Peters (2007) defined mobile learning as a form of learning supported by mobile technologies. However, these definitions were considered technology-centric by some researchers (Traxler, 2007; Vosloo, 2012). Another definition was provided by Motiwalla (2007), who described mobile learning as individualized learning from anywhere at any time. On the other hand, mobile learning is not regarded as one type of learning in some studies. Indeed, it was defined as learning facilitated by mobile devices (Herrington & Herrington, 2007; Valk et al., 2010).

1.2. The Framework for The Rational Analysis of Mobile Education (FRAME) Model

In order to understand each component of mobile learning, the present study needed an over-arching framework. For this purpose, the FRAME model was chosen, which was developed by Koole (2006) and Koole and Ally (2006). This model was accepted as the first comprehensive theoretical framework for mobile learning. In this model, mobile learning was defined as a process resulting from the convergence of mobile technologies, human learning capacities, and social interaction. It is helpful for educators in terms of planning and designing mobile learning environments (Park, 2011). A Venn diagram was used to represent the FRAME model (Koole, 2009) (see Figure 1).

Figure 1. *The FRAME model.*



The three circles represent three main aspects, namely Device Aspect (D), Learner Aspect (L), and Social Aspect (S). There are also three intersection areas, which are comprised of two different aspects. Device Aspect (D) represents the mobile devices and their technical, physical features, and capabilities. This aspect is important due to behaving as a bridge between the learner and the learning task(s) (Koole, 2009). Learner Aspect (L) refers to the situations and tasks that the student wants or needs to succeed. The learner aspect highlights the learner characteristics that include cognitive ability, memory, prior knowledge, emotions, and possible motivations (Koole, 2009). Social Aspect (S) defines social interaction and cooperation. Device Usability Intersection (DL) includes the elements of both Device Aspect (D) and Learner Aspect (L). This intersection corresponds to the characteristics of mobile devices which influence the learners' psychological comfort and satisfaction while interacting with them. Its functions like a bridge between the characteristics and needs of the learner and the technical features of the mobile device. Social Technology Intersection (DS) includes both Device Aspect (D) and Social Aspect (L). This intersection refers to how mobile devices provide communication and collaboration among multiple learners through multiple systems, and it is mostly based on the philosophy of social constructivism. Learner Aspect (L) and Social Aspect (S) constitute Interaction Learning Intersection (LS). According to Koole (2006), this intersection includes learning and instructional theories, but is largely based on the philosophy of social constructivism. As the primary intersection of the FRAME model, Mobile Learning Process (DLS) contains three elements that belong to Device Aspect (D), Learner Aspect (L), and Social Aspect (S). In an effective mobile learning process, it is expected to provide cognitive environments where learners can appropriately interact with each other, instructors, and course materials (Koole, 2006). In this way, the time for searching information and efforts spend for the evaluation of it are reduced.

2. METHOD

2.1. Instrument Development

The Mobile Phone Use in Academic Environment Scale (MPUAES) was adapted from the Mobile Phone Affinity Scale (MPAS) (Bock et al., 2016). The MPAS scale assessed both negative and positive aspects of mobile phone use in the work environment. Thus, 6-factor of the MPAS was assigned as follows: Connectedness, Productivity, and Empowerment as positive sub-dimensions; Anxious Attachment, and Addiction as negative sub-dimensions; and Continuous Use as a neutral sub-dimension. The present study aimed to develop the Mobile Phone Use in Academic Environment Scale (MPUAES) based on 24 items of the MPAS, which was adapted to the academic environment.

The necessary permissions were taken before starting work on this scale development study. To ensure content validity, the researchers worked with three experts in the field of Computer Education and Instructional Technology Department, one expert in the field of Curriculum and Instruction Department, and one expert in the English Language Department. Besides excluding some words related to the work environment, some words were included to make it suitable for an instructional environment. Furthermore, cognitive interviews with three undergraduate students were conducted before piloting the scale, which was important for detecting possible response errors and finding the reasons for these errors in the survey (Willis, 2004). The students evaluated the items to avoid misunderstanding and hence unintended responses. With the guidance of student comments, some items were revised by adding a more prevalent verb near the less-known words to ease the understanding of participants and make sure that all the items were clear to them. For example, in one of the items, the phrase "keep track of" was used and it was clarified by adding the word "follow" as seen in the following: "My phone helps me keep track of *-follow-* my academic life". Moreover, an operational definition of the concept of "academic life" was given at the beginning of the survey to clarify its meaning and share a common understanding with the students.

2.2. Participants

Data was collected during the fall semester of 2016-2017 and the spring semester of 2017-2018 from all faculties of Middle East Technical University (METU). It was assumed that those familiar with technology would be more willing to fill out the online survey compared to the others who were not quite familiar with it. To ensure common conditions for the completion of the survey, the researchers handed out a hand-delivered questionnaire and the online survey form was not preferred to prevent low internal validity owing to the possibility of a selection threat (selection bias) (Kite & Whitley, 2018). The inclusion criteria for participation in this study were defined as any undergraduate student who was still studying in any department of METU and owned a smartphone. In the demographics section, information regarding gender, current GPA, age, faculty, department, and graduate level was collected.

In the first stage, the factorial structure of the instrument was explored with 240 undergraduate students. The second stage comprised of 1647 participants. In both stages, the data were collected from all faculties and all grade levels of METU (Table 1).

Table 1. *Distribution of the participants in the pilot study and validation study by departments and study year.*

	Pilot study		Validation study	
	Sample1 (n_1 = 240)		Sample2 (n_2 = 1647)	
	f	*%*	*f*	*%*
Gender				
Female	140	58.3	832	50.5
Male	100	41.7	815	49.5
Faculty				
Architecture	-	-	98	6.0
Arts & Science	84	35.0	325	19.7
Economics & Administrative Sciences	14	5.8	231	14.0
Education	72	30.0	207	12.6
Engineering	70	29.2	786	47.7
Study Year				
Freshman	70	29.2	447	27.1
Sophomore	70	29.2	468	28.4
Junior	70	29.2	421	25.6
Senior & Senior (+)	70	29.2	311	18.9

2.3. Data Analysis

Initially, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA)were performed through SPSS and AMOS for the development of MPUAES. In addition to EFA and CFA analyses, structural model validation and convergent and divergent construct validity were applied for the validation and confirmation of the factor structure. A pilot study was carried out with 240 undergraduate students. Then, cross-validation analysis was performed for the validation of the three-factor structure of the scale with a sample of 1867 undergraduate students. According to Byrne (2010), this type of analysis offers the advantage of examining the factorial structure of the scale across different samples of the same population. Thus, the sample in the present study was split into two random samples for conducting both EFA and CFA analyses based on the suggestion of Cudeck and Browne (1983).

3. RESULTS

3.1. Findings on Content Validity

The items in the study were generated based on 24 items of the MPAS. The researchers worked with five experts to ensure content validity; three of them were from the Computer Education and Instructional Technology Department; one expert was from the Curriculum and Instruction Department; and one expert was from the English Language Teaching Department. Based on the suggestions of experts, while some words were excluded from the items, some were added to be suitable for an academic environment. Before piloting the study, cognitive interviews were conducted with three undergraduate students. In this way, the possible response errors were detected.

3.2. Findings on Construct Validity

3.2.1. *Exploratory factor analysis*

Before performing the EFA, missing data was examined in the data. Due to the less than five percent on a single variable, it was ignored based on the suggestion of Hair et al. (2010). The sample size for conducting the EFA was checked in two ways. Firstly, 10:1 rule, which means ten cases for each item, or being above 100 cases (Hatcher, 1994) was acceptable to run the EFA. The rules were met for 24 items with 240 cases. Secondly, Kaiser-Meyer-Olkin (KMO) was checked. Since KMO value (.92) was above .60, it was accepted as a great value for sampling adequacy according to Hutcheson and Sofroniou (1999). On the other hand, the data were screened to detect univariate outliers and multivariate outliers. Although some cases were found, as the recommendation of Tabachnick and Fidell (2013), the researcher examined whether the cases were suitably part of the sample and decided not to remove them. As another assumption, univariate normality was checked by Skewness and Kurtosis values, Kolmogorov Smirnov and Shapiro Wilk tests, histograms, and Q-Q plots. The normality assumption was met based on Skewness and Kurtosis values, histogram, and Q-Q plots. Multivariate normality was also checked through Mardia's Test. It was found significant (p = .00), which means the multivariate normality was violated. Lastly, the appropriateness of EFA was checked through a correlation matrix and Barlett's test of sphericity. According to Tabachnick and Fidell (2009), if correlation coefficients are under .30, there is no need to conduct EFA. When the correlation matrix was examined, it was seen that many correlations exceeded this threshold. Moreover, Barlett's test of sphericity was found significant (χ^2 (153) = 2252.40, $p < 0.05$) at the .05 level, which indicates the presence of nonzero correlations. Both the results of the correlation matrix and Barlett's test of sphericity were the indicators of suitability for performing EFA. After all, the preliminary analysis showed that it was appropriate to conduct factor analysis. Since the multivariate normality assumption was not met, Principal Axis Factoring (PAF) was selected as the extraction method (Costello & Osborne, 2005). Moreover, oblique rotation, more specifically direct oblimin, was chosen as a factor rotation method owing to the presence of correlated factors (Preacher & McCallum, 2003).

In order to determine the number of factors, the scree-test and eigenvalues were checked. In the first run of EFA with 24 items, a pattern matrix with 5 factors was observed. With the rule of .30 factor loadings (Fidell, 2006; Hair et. al, 2010), Item 12 and Item 13 were deleted. After removing those 2 items, the EFA was run again. Item 18 and Item 21 were omitted because their communality values were lower than .40 based on the suggestions of Costello and Osborne (2005). Since Item 14 and Item 17 had similar meanings, the lower-loaded one, Item14, was deleted. Although its factor loading was above .30, Item 8 was also deleted since it was loaded on the first factor for which it is not suitable. After omitted the aforementioned items, the EFA was performed with 18 items for the last time. The pattern matrix was screened, and it was observed that all factor loadings were above .40, and there was not any cross-loaded item. The scree pilot indicated the presence of three factors. Eigenvalues were also examined to decide a reliable estimation on the number of factors. According to Tabachnick and Fidell (2013), eigenvalues less than 1 are not important for variance. There were three factors explaining 63.42% of the total variance in the study (see Table 2). Factor 1, 2, and 3 accounted for 41.93, 13.28, and 8.22 of the total variance, respectively.

Table 2. *Pattern coefficient for mobile phone use in academic environment scale*

		Factor			
	Item	1	2	3	Communality
Facilitator	i16. My phone is necessary for my academic life	**.86**	-.08	-.02	.68
	i1. I feel in control of my academic life when I have my phone with me	**.83**	.07	-.14	.63
	i22. In my academic life, my phone gives me a sense of comfort.	**.79**	-.07	.06	.64
	i17. Without my mobile phone, I feel detached -*out of touch, isolated*- to my academic life.	**.72**	.15	-.18	.50
	i11. Having my phone with me makes it easier to sort out -*resolve, handle*- the critical situations related to my academic life.	**.71**	-.09	.20	.63
	i7. For my academic life, I feel dependent on my phone.	**.69**	.09	.03	.54
	i23. My phone helps me be more organized for my academic life.	**.68**	.01	.20	.64
	i4. When it comes to the academic life, my phone is my personal assistant.	**.61**	-.07	.26	.56
	i6. I feel more comfortable in doing my school work when I have my phone with me.	**.57**	.08	.14	.47
Distractor	i5. When I should be doing the school work, I find myself occupied with my phone.	.02	**.80**	-.03	.64
	i10. I find myself occupied on my phone even when I'm with my classmates or instructors (during the class or studying).	.07	**.73**	.01	.58
	i9. In class or whenever I study, I read/send text messages that are not related to what I am doing.	.04	**.72**	.05	.57
	i3. I would get more school work done if I spent less time on my phone.	-.12	**.66**	.00	.40
	i24. I find myself engaged with my mobile phone for longer than I intended	.09	**.58**	.11	.44
Connectedness	i2. I use my phone to connect with my classmates or instructors	-.12	.05	**.79**	.57
	i1. My phone helps me keep track of -*follow*- my academic life.	.18	.00	**.62**	.53
	i19. My phone helps me stay close to my classmates and instructors.	.16	.14	**.58**	.53
	i20. My phone makes it easy to cancel the arranged plans withclassmates or instructors.	.23	.17	**.53**	.56
	Eigenvalues	7.55	2.40	1.48	
	% of Variance	41.93	13.28	8.22	
	Cronbach's α	.92	.84	.81	

Note. Extraction Method: Principal Axis Factoring. Rotation Method: Oblimin with Kaiser Normalization. The items above .30 were signed in bold.

Based on the aforementioned rules, it was concluded that the number of factors to be retained was three. Items 16, 15, 22, 17, 11, 23, 4, and 6 were loaded on Factor 1 labeled as Facilitator; items 5, 10, 9, 3, and 24 were loaded on Factor 2 labeled as Distractor; items 2, 1, 19, and 20 were loaded on Factor 3 labeled as Connectedness.

Kaiser's eigenvalue-greater-than-one rule, namely the Kaiser criterion, is seen as the most approved method in practice (Fabrigar et. al, 1999) and it is also accepted as the most accurate method to reveal the relationships between the items (Büyüköztürk, 2007). Nonetheless, some researchers found this rule problematic and inefficient in determining the number of factors (Ladesma & Pedro, 2007). Therefore, the parallel analysis has been proposed as the best alternative and appropriate method in some studies (Humphreys & Montanelli, 1975; Zwick & Velicer, 1986). Both Kaiser's eigenvalues in the first column and the PA eigenvalues in the third column are seen in Table 3. According to these results, none of the eigenvalues of PA was greater than Kaiser's eigenvalues. This means that there was not a factor obtained by the chance. To conclude, the Kaiser criterion was supported by the results of the parallel analysis upon which the number of factors to be retained was three.

Table 3. *The Results of the Parallel Analysis.*

Factor	Kaiser's eigenvalues	Mean of eigenvalues	PA eigenvalues
1*	7.55	1.51	1.61
2*	2.40	1.41	1.48
3*	1.48	1.33	1.39

**The retained factor according to the results of the parallel analysis.*

3.2.2. *Structural model validation*

A measurement model refers to the linear or nonlinear statistical functions involving the relation between items and constructs to be measured (Yurdugül & Aşkar, 2008). In order to evaluate the proposed measurement model and alternative models, first-order confirmatory factor analysis was performed. As an estimation method, the maximum likelihood (ML) was chosen upon the recommendation of Tabachnick and Fidell (2013) for medium to large sample sizes and plausible assumptions. The data consisted of 240 undergraduate students. In order to investigate factorial validity, five measurement models were used and given in the explanations below.

- Model I indicated 24 items with a unidimensional construct measurement model.
- Model II indicated a six-factor measurement model as proposed in the original scale. These factors were as follows: Connectedness, Productivity, Empowerment, Anxious Attachment, and Continuous Use.
- Model III indicated a three-factor measurement model which was obtained in the present study. Principal Axis Factoring was selected as the extraction method. The model included 18 items, and the factors were as follows: Facilitator, Distractor, and Connectedness. In this model, the three factors were considered to be correlated.
- Model IV indicated a three-factor measurement model which was obtained in the present study, where the latent factors were considered to be uncorrelated.
- Model V (Empirical Measurement Model) indicated a three-factor measurement model which was obtained in the present study; and the factors were correlated. Differently, in order to improve model-fit, some error variances were allowed to covary in this model

The following fit indices were chosen to compare alternative models (Yurdugül, 2007): root mean square error of approximation (RMSEA), goodness of fit index (GFI), comparative fit index (CFI), and non-normed fit index (NNFI). The model-data fits were computed for all the measurement models as depicted in Table 4. The criteria for good-fit-indices are also illustrated in the table.

Table 4. *Good-of-fit indices and comparison of the measurement models.*

		RMSEA	GFI	CFI	NNFI
		<0.08	≥0.90	≥0.90	≥0.90
Model I:	Unidimensional Model	.12	.66	.71	.68
Model II:	Six-Factor Structure	.10	.77	.81	.78
Model III:	3-factor Structure (Correlated)	.10	.82	.87	.85
Model IV:	3-factor Structure (Uncorrelated)	.12	.78	.80	.77
Model V:	3-factor Structure (correlated- covaried)	.06	.92	.96	.95

Note. References*:* Hair et al. (2010). Kline (2011).

Firstly, Model I was built, which was a unidimensional model with 24 items. According to fit indices of the model, Model I showed a poor model fit. This can be interpreted as an indicator that the scale consisting of 24 items did not confirm the one-factor structure model, but it should have more than one sub-construct. Secondly, Model II was based on the six-factor structure model as the original scale, which included 24 items. Although an improvement was observed in the fit indices compared to Model I, it was not sufficient for a good model fit. This was also proof that the scale was not suitable for the six-factor structure model with 24 items. Thirdly, the present study proposed Model III, in which a three-factor structure (correlated) model was obtained from the pilot study. In this model, the number of items dropped from 24 to 18 items. Although the fit indices showed an improvement, they were not in the acceptable range. Similar to Model III, Model IV indicated a three-factor structure model obtained from the present study, but the latent factors were assumed to be uncorrelated. As seen in Table 4, a decline was observed in the good-of-fit indices of the model. Finally, Model V was built, which was a three-factor measurement model with 18 items. The latent factors were correlated; and some error variances which were found highly correlated were allowed to covary in the model. According to the fit indices, Model V was found as the most appropriate among five measurement models. Consequently, it was continued with Model V based on these results in the current study.

3.2.3. *Convergent and discriminant validity*

In the present study, construct validity was also examined by two ways: (1) convergent validity, and (2) discriminant validity. (Yurdugül & Sırakaya, 2013). The present study used three measures to estimate convergent validity of the model. The first rule was that factor loadings should be greater than .050 (Hair et al., 2010). They were between .51 and .82, which met the rule. Secondly, average variance extracted (AVE) was calculated and obtained above .50, which was acceptable according to the rule of thumb greater than .50. Lastly, composite (construct) reliability (CR) was calculated as an indicator of convergent validity. As seen in Table 5, CR values were obtained between .80 and .91, which were acceptable according to the rule of thumb greater .70.

Table 5. *Convergent validity for the measurement model.*

	L Interval (a)	AVE (b)	CR (c)
Facilitator	.61 – .80	.56	.92
Distractor	.51 – .82	.50	.83
Connectedness	.58 – .81	.51	.80

Note. L = Factor Loadings. AVE = Average Variance Extracted. CR = Composite Reliability

For discriminant validity, the correlations among the subscales of the MPUAES and the square root of AVE were used. According to this, the square root of AVE calculated for each dimension

must be greater than correlations coefficients between the corresponding sub-dimension and remaining sub-dimensions and must be higher than .50 as well (Fornel & Larcker, 1981). As seen in Table 6, the discriminant validity was ensured.

Table 6. *Discriminant validity for the measurement model.*

	Facilitator (1)	Distractor (2)	Connectedness (3)
Facilitator (1)	**(.75)**		-
Distractor (2)	.43	**(.71)**[b]	-
Connectedness (3)	.71[a]	.55	**(.71)**

Note. The values in parentheses are the square roots of AVE. *a = .7090. b = .7135.*

3.2.4. *Confirmatory factor analysis*

In order to confirm a three-factor structure of MPUAES, CFA was performed with the rest of the data which consisted of 1647 students. Before performing confirmatory factor analysis, the following assumptions were checked, separately: sample size, normality, and absence of outliers (Tabachnick & Fidell, 2013) Firstly, the adequacy of sample size was checked. The thumb rule 1:10 was met with 18 items and 1647 participants (Hair, et al., 2010). Secondly, both univariate and multivariate outliers were screened. For univariate outliers, standardized z-scores and box-plot were checked. 10 cases were detected which exceeded the absolute value of 3.29. Regarding box-plot representations, a few univariate outliers were observed, which were possible for the studies with the large sample size (Pallant, 2007; Tabachnick & Fidell, 2007). As being a multivariate analysis, SEM studies take into consideration multivariate outliers instead of univariate ones. Thus, they were not deleted. For multivariate outliers, Mahalanobis distance (D^2) was calculated for each case. Out of 1647, thirty-seven cases were detected as multivariate outliers with the critical value of 42.312 (df = 18, p = .001). After omitting these cases, the analysis was performed again. It was observed that the results were not substantially affected. That is, 37 cases were determined as possible outliers, which were remained in the data. Thirdly, univariate normality was also checked. Kolmogorov-Smirnov and Shapiro-Wilk test results were found significant, which was a sign of non-normal distribution. However, these tests cannot be considered as only indicators for normality because of being very sensitive to sample size. Skewness and kurtosis values were also checked, which were between -3 and +3. The visual inspection of histogram and Q-Q plots were also observed, in which there was not any evidence for violation of normality. Thus, the univariate normality of the data was assured by skewness and kurtosis values, histogram, and Q-Q plots. As an estimation method, the maximum likelihood (ML) was chosen upon the recommendation of Tabachnick and Fidell (2013) for medium to large sample sizes and plausible assumptions. The following fit indices were selected to assess the goodness-of-fit of the model: Chi-square (χ^2), comparative fit index (CFI), adjusted goodness of fit index (AGFI), goodness of fit index (GFI), non-normed fit index (NNFI), normed fit index (NFI), root mean square error of approximation (RMSEA), root mean square residual (RMR), and standardized root mean square residual (SRMR) (Jöreskog & Sörbom, 1993; Kline, 2011). The model fit indices selected for the current study are presented in Table 7, in which the references for each fit index are also shown.

The second-order CFA resulted a significant chi-square, χ^2 (132, n = 1647) = 1684.21, p = .00, which indicated an unacceptable model. However, according to Tabachnick and Fidel (2013), chi-square is sensitive to sample size. Thus, other fit indices were examined, and the following results were found: CFI = .89, NNFI = .87, GFI = .89, AGFI = .86, RMR = .08, RMSEA = .09, and SRMR = .06. CFI, and NNFI values showed poor model fitting, which should be greater than .95 for a perfect model fit, and at least .90 for a good model fit (Tabachnick & Fidell, 2013; Jöreskog & Sörbom, 1993; Kline, 2011). The same rule was in use for the values of GFI and

AGFI, which also showed poor fitting due to being less than .90 (Hair et al., 2010). In addition, RMSEA value greater than .08 indicates a poor fitting model (Browne & Cudeck, 1993). The values SRMR and RMR were only indicatives of a good fit (Jöreskog & Sörbom, 1993; Kline, 2011). Thus, the researchers examined the error covariances (i.e., modification indices of errors). Eight error covariances (ε4-ε13, ε16-ε17, ε12-ε13, ε7-ε13, ε3-ε5, ε8-ε9, ε1-ε14, and ε1-ε15) were found highly relatively in the program output. As seen in Figure 2, the items related to these error covariances were loaded on the same factors. Before covarying, the relevant items were checked by two experts from the Computer Education and Instructional Technology Department. In the first factor, namely facilitator, item 13 "Without my mobile phone, I feel detached -out of touch, isolated- to my academic life." was related to item 4, item 7, and item 12. When these three items were examined (see Table 9 in Appendix), it was seen that they highlighted the necessity of mobile phones in an academic life. Thus, the experts allowed them to covary in the model. Similarly, under the facilitator factor, the following item pairs, namely item 16 and item 7, were also allowed to covary since both pointed out that it was a great convenience using mobile phones in an academic life. In the distractor factor, one of the error covariances was observed between item 8 and item 9. The experts allowed to covary these errors because both items implied that mobile phones could be a distraction while studying. The other item pairs were item 3 and item 5. They were also allowed to covary since "school work" was the focus in both items. The other two modification errors were under the connectedness factor. Item 1 "My phone helps me keep track of -follow- my academic life" was related to item 14 and item 15. When these two items were checked, "follow academic life" and "keep in touch with classmates and instructors" might be perceived as similar, thus the experts allowed them to covary as well.

Table 7. *The model fit indices used for confirmatory factor analysis.*

Model Fit Index	Acceptable Fit		Sample Statistics	Decision	References*
	Moderate Fit	Good Fit			
NNFI	.95 - .97	.97 – 1.00	.91	Moderate	a, b, e
CFI	.90 - .95	.95 – 1.00	.92	Moderate	a, b, d, e, f,
GFI	.90 - .95	.95 – 1.00	.92	Moderate	d, f
AGFI	.90 - .95	.95 – 1.00	.90	Moderate	b, e, f,
SRMR	.05 - .08	≤ .05	.06	Moderate	c, d
RMR	.05 - .08	≤ .05	.08	Moderate	c, d
RMSEA	.05 - .08	≤ .05	.07	Moderate	c, f

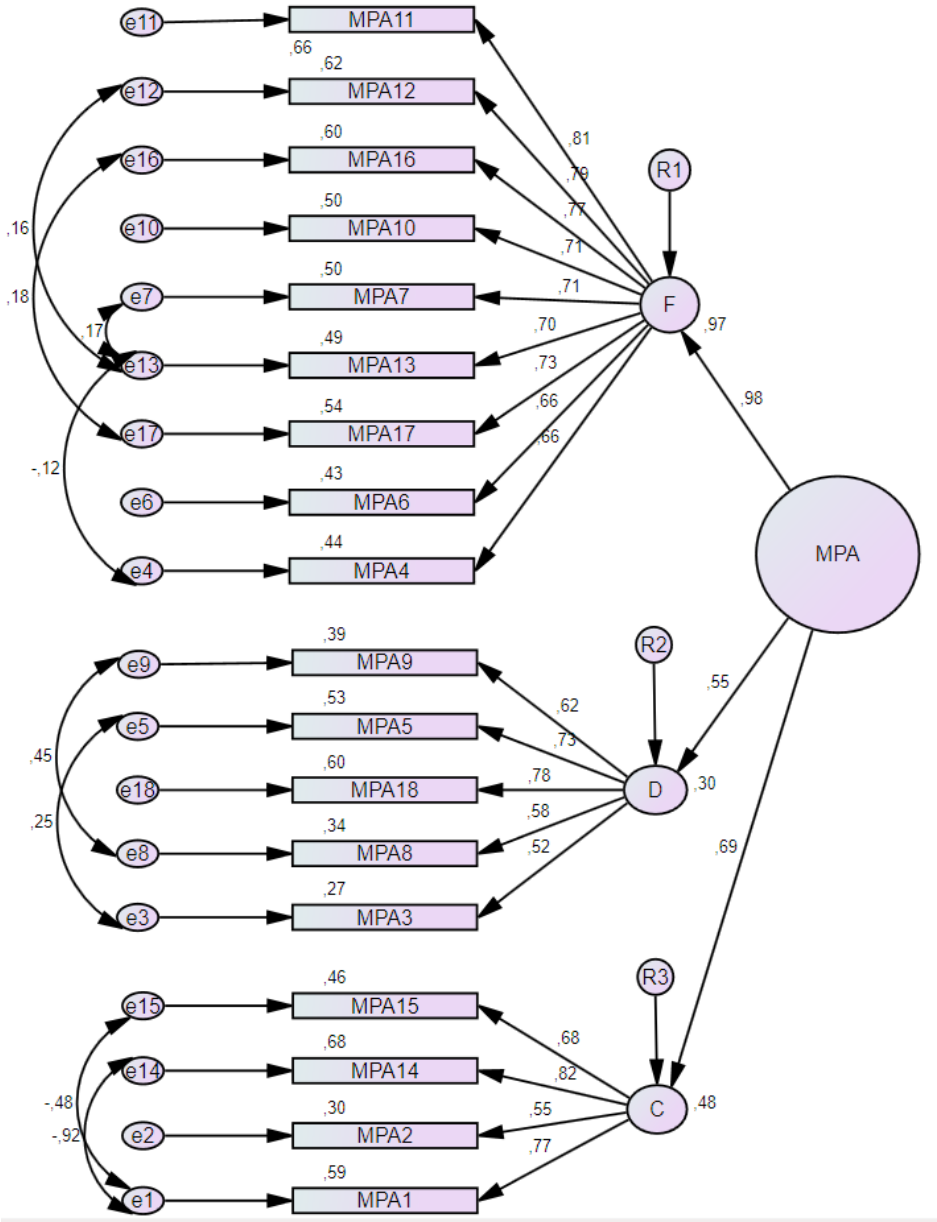
Note. * References: a = Tabachnick and Fidell (2013). b = Jöreskog and Sörbom (1993). c = Browne and Cudeck (1993). d = Hu and Bentler (1999). e = Kline (2011). f = Hair et al. (2010).

The results revealed a close fit model. The fit indices of the model were as follows: CFI = .92, NNFI = .91, GFI = .92, AGFI = .92, RMR = .08, SRMR = .06 and RMSEA = .07. Chi-square was found significant despite of decreasing the value $\chi 2$ (129, n = 1647) = 1199.574, p = .00. Since chi-square (χ^2) is expected to be significant for large sample sizes, other fit indices should be taken into consideration (Tabachnick & Fidell 2013). All other fit indices, except SRMR value, indicated a good model fit. The SRMR value was found .05, which was an indicator of the perfect fitting model (Hu & Bentler, 1999).

The proposed second-order factor model of MPUAES is shown in Figure 2. The standardized estimates of the second-order factors were .98, .55, and .69. Their standardized factor loadings varied between .66 and .81 for the facilitator factor, varied between .52 and .78 for the distractor factor, and .55 and .82 for connectedness factor. Thus, it can be concluded that all items had a

significant contribution to the proposed model since the cut-off point of the standardized estimates of the items was .40 (Stevens, 2002).

Figure 2. *The factor structure of MPUAES with standardized estimates.*



3.3. Findings on Reliability

For internal consistency, Cronbach alpha coefficients were examined for each factor, which was found as .92 for facilitator factor (9 items), .82 for distractor factor (5 items), and .73 (4 items) for connectedness factor. Being greater than .70, these values were acceptable (Nunally, 1978).

3.4. Interpretation of Mobile Phone Use in Academic Environment Scale Scores

The Mobile Phone Use in Academic Environment (MPUAES) comprised of 16 items. A 5-point Likert-type grading scale [Extremely true (5) → Not at all true (1)] was applied on the scale. Three proposed dimension and their items are shown in Table 8: facilitator (9 items), distractor (5 items), and connectedness (4 items). Therefore, possible scores for each dimension range as follows: between 9 and 45 for facilitator; between 5 and 25 for distractor; and between

4 and 20 for connectedness factor. Since a second-order CFA was performed, the total score of the scale was calculated as well. Accordingly, it ranges between 18 to 90 for the whole mobile phone use in an academic environment scale.

Table 8. *The dimensions and items of MPUAES.*

Dimensions	Number of items	Items
Facilitator	9	i11, i2, i16, i10, i17, i7, i13, i6, i4
Distractor	5	i9, i18, i5, i8, i3
Connectedness	4	i15, i14, i12, i1

The evaluation of the MPUAES scores was performed according to both the scores from the subscales and the total score of the scale. This means that besides the dimensions of the scale, the total score related to mobile phone use in an academic environment can be obtained on the scale as well. If the students' scores from the subscales are high, their mobile phone use in terms of relevant dimensions is also high. Likewise, a high total score indicates that students' mobile phone use in an academic environment is high.

4. DISCUSSION and CONCLUSION

The MPUAES was developed based on the 24 items of the MPAS scale with a six-factor structure (Bock et al., 2016). The original scale was developed for a work environment, which was adapted to the academic environment in this study. First, a pilot study was carried out with 240 students and the EFA was run several times to diagnose the problematic items. As a result of this process, six problematic items were omitted and a three-factor structure with 18 items was obtained. The number of factors was decided based on scree plot, Kaiser's eigenvalues, and the parallel analysis. Then, the validation of the three-factor structure of the scale was performed with 1647 students. To sum up, the MPUAES proposed a three-factor structure with 18 items: facilitator (9 items), distractor (5 items), and connectedness (4 items) (see Table 9 in Appendix). Cronbach alpha coefficients were examined for each factor, which was found as .92, .82, and .73, respectively. Being greater than .70, these values were acceptable (Nunally, 1978).

According to the results of factor analysis, three factors were obtained, which were labeled as facilitator, distractor, and connectedness, upon FRAME model developed by Koole (2006). According to this model, mobile learning consists of three aspects: (1) Device, (2) Learner, and (3) Social. That is, besides the technical specifications of the mobile devices, social and personal dimensions of learning should be considered in the context of mobile learning. Furthermore, in the FRAME model, each aspect intersected with the other one and formed three intersections, which are device usability (device and learner aspect), social technology (social and device aspect), and interaction learning (learner and social aspect). The intersections of these three aspects lead to the ideal mobile learning. In the MPUAES, the three factors, namely facilitator, distractor, and connectedness, covered the aforementioned three main aspects and three intersections of the FRAME model. More specifically, the factors were assigned as follows: technical features of smartphones as device aspect; facilitator and distractor sub-dimensions as learner aspect; and connectedness sub-dimension as the social aspect. For instance, item 2 "I use my phone to connect with my classmates or instructors" corresponds to the social aspect of the FRAME model. Apart from the association of the items with the main aspects of the model, they were also related to the intersections. For instance, item 23 loaded on facilitator factor "My phone helps me more organized for my academic life" consisted of both device and learner aspect, so it corresponds to the intersection of device usability, as well. Similarly, item 9 under distractor factor "In class or whenever I study, I read/send text messages that are not related to what I am doing" was associated with all three intersections due to including functionality of the device, social relationship, and learner characteristics. Although all items were

associated with all aspects and intersections of the model in some way, the learner aspect was essential for the MPUAE scale because of focusing on students' experiences with their mobile phones in the academic environment such as prior knowledge, skills, emotions, and motivations, etc. Thus, it can be concluded that MPUAES was primarily based on the learner aspect of the FRAME model, and also as the characteristics of the FRAME model, the scale was a convergence of mobile technologies, learner characteristics, and social interaction.

To conclude, the results of the study indicated that the scores obtained from the developed scale MPUAES were valid and reliable in assessing undergraduate students' mobile phone use in an academic environment. The study had some significant implications which should be considered by researchers interested in mobile technologies usage in higher education. The present study provided a comprehensive perspective on undergraduate students' educational mobile phone use by considering both positive and negative aspects. Apart from the technology acceptance models, the current study offered a new measurement approach for the assessment of educational mobile phone use. Yet, the inclusion of only one university was one of the limitations of this study. To enhance generalizability and external validity (Merriam & Tisdell, 2015), the study might further be conducted with different universities from different regions of Turkey. Moreover, the criterion-based validity could not be checked due to the absence of an educational mobile phone use scale that can be used as a criterion. Thus, this can be further analyzed in the future studies. Lastly, this study focused especially on the learner aspect. Further studies might focus on other aspects of the FRAME model.

Acknowledgments

The research was supported by Scientific Research Project Coordinator of METU as an FDP Project (*No:1416*).

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s). **Ethics Committee Number**: 2017-FEN-003, İnsan Araştırmaları Etik Kurulu (İAEK)

Authorship Contribution Statement

Nehir YASAN AK: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing - original draft. **Soner YILDIRIM**: Investigation, Methodology, Supervision, and Validation.

Orcid

Nehir Yasan Ak https://orcid.org/0000-0003-4801-2740
Soner Yildirim https://orcid.org/0000-0002-3167-2112

REFERENCES

Abu-Al-Aish, A., & Love, S. (2013). Factors influencing students' acceptance of m-learning: an investigation in higher education. *The International Review of Research in Open and Distributed Learning*, *14*(5). https://doi.org/10.19173/irrodl.v14i5.1631

Bernacki, M.L., Greene, J.A., & Crompton, H. (2020). Mobile technology, learning, and achievement: Advances in understanding and measuring the role of mobile technology in education. *Contemporary Educational Psychology*, *60*, 101827. https://doi.org/10.1016/j.cedpsych.2019.101827

Bock, B.C., Lantini, R., Thind, H., Walaska, K., Rosen, R.K., Fava, J.L., ... & Scot Sheldon, L. A. (2016). The mobile phone affinity scale: enhancement and refinement. *JMIR mHealth and uHealth*, *4*(4). https://doi.org/10.2196/mhealth.6705

Bryant, E.C., (2016). *Graduate student perceptions of multi-modal tablet use in academic environments* [Doctoral dissertation, University of South Florida]. The USF Libraries. https://core.ac.uk/download/pdf/154477153.pdf

Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Sage. https://psycnet.apa.org/record/1993-97481-000

Büyüköztürk, Ş. (2007). *Sosyal bilimler için veri analizi el kitabı (*7. Baskı). [Data analysis handbook for social sciences]. Pegem Akademi Yayınları.

Byrne, B.M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (2nd ed.). Routledge. https://doi.org/10.4324/9780203805534

Cheon, J., Lee, S., Crooks, S.M., & Song, J. (2012). An investigation of mobile learning readiness in higher education based on the theory of planned behavior. *Computers and Education*, *59*(3), 1054–1064. http://doi.org/10.1016/j.compedu.2012.04.015

Costello, A.B., & Osborne, J.W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment Research and Evaluation, 10*(7). http://pareonline.net/pdf/v10n7.pdf

Crompton, H. (2017), Moving toward a mobile learning landscape: presenting a mlearning integration framework. *Interactive Technology and Smart Education*, *14*(2), 97-109. https://doi.org/10.1108/ITSE-02-2017-0018

Cudeck, R., & Browne, M.W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research, 18*(2), 147-167. https://doi.org/10.1207/s15327906mbr1802_2

Fabrigar, L.R., Wegener, D.T., MacCallum, R.C., & Strahan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods*, *4*(3), 272-299. https://doi.apa.org/doi/10.1037/1082-989X.4.3.272

Field. A. (2009). *Discovering statistics using by SPSS* (3rd ed.). London: Sage Publication. https://uk.sagepub.com/en-gb/eur/discovering-statistics-using-sas/book234095

Ford, J.R. (2016). *Learners' Perspectives on How Mobile Computing Devices Usage Interacts with Their Learning* (Order No. 10168369). Available from ProQuest Dissertations & Theses Global. (1846531179). https://www.proquest.com/dissertations-theses/learners-perspectives-on-how-mobile-computing/docview/1846531179/se-2

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

Gikas J., & Grant, M.M. (2013). Mobile computing devices in higher education: Student perspectives on learning with cellphones, smartphones & social media. *The Internet and Higher Education. 19*, 18-26. https://doi.org/10.1016/j.iheduc.2013.06.002

Hair, J.F., Anderson, R.E., Tatham, R.L., & Black, W.C. (2010). *Multivariate Data Analysis* (7th ed.). Prentice Hall, Inc. https://doi.org/10.1016/j.iheduc.2013.06.002

Han, S., & Yi, Y.J. (2019). How does the smartphone usage of college students affect academic performance? *Journal of Computer Assisted Learning*, *35*(1), 13-22. https://doi.org/10.1111/jcal.12306

Hatcher, L. (1994). *A step-by-step approach to using the SAS® system for factor analysis and structural equation modeling.* Cary, NC, USA: SAS Institute, Inc. https://www.sas.com/storefront/aux/en/spsxsfactor/61314_excerpt.pdf

Herrington, A., & Herrington, J. (2007, November). *Authentic mobile learning in higher education* [Paper presentation]. Australian Association for Research in Education Conference, Fremantle, Western Australia. https://www.aare.edu.au/07pap/her07131.pdf

Hu, L.H. & 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), 1-15. https://doi.org/10.1080/10705519909540118

Huang, R.T., Jang, S.J., Machtmes, K., & Deggs, D. (2012). Investigating the roles of perceived playfulness, resistance to change and self-management of learning in mobile English learning outcome. *British Journal of Educational Technology*, *43*(6), 1004-1015. https://doi.org/10.1111/j.1467-8535.2011.01239.x

Humphreys, L.G. & Montanelli, R.G. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research*, *10*, 193-206. https://doi.org/10.1207/s15327906mbr1002_5

Hutcheson, G., & Sofroniou, N. (1999). *The multivariate social scientist: Introductory statistics using generalized linear models.* Sage. https://uk.sagepub.com/en-gb/eur/the-multivariate-social-scientist/book205684

Iqbal, S., & Qureshi, I.A. (2012). M-learning adoption: A perspective from a developing country. *The International Review of Research in Open and Distributed Learning*, *13*(3), 147-164. https://doi.org/10.19173/irrodl.v13i3.1152

Jöreskog, K.G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language.* Chicago: Scientific Software International. https://psycnet.apa.org/record/1993-97878-000

Keegan, D. (2005, October). *The incorporation of mobile learning into mainstream education and training* [Paper presentation]. 4th World Conference on mLearning, Cape Town, South Africa. Google Scholar

Kite, M.E., & Whitley, B.E. (2018). *Principles of Research in Behavioral Science* (4th ed.). Routledge. https://doi.org/10.4324/9781315450087

Kline, R.B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York: The Guilford Press. ftp://158.208.129.61/suzuki/PP_SEM_3e.pdf

Koole, M.L. (2006). *The framework for the rational analysis of mobile education (frame) model: an evaluation of mobile devices for distance education* (Doctoral dissertation). Athabasca University, Athabasca, AB, Canada, 2006. http://hdl.handle.net/2149/543

Koole, M., & Ally, M. (2006, April). *Framework for the rational analysis of mobile education (FRAME) model: Revising the ABCs of educational practices* [Paper presentation]. International Conference on Networking, International Conference on Systems and International Conference on Mobile Communications and Learning Technologies, Morne, Mauritius. https://doi.org/10.1109/ICNICONSMCL.2006.103

Koole, M.L. (2009). A model for framing mobile learning. In M. Ally (Ed.), *Mobile learning: Transforming the delivery of education and training* (pp. 25-47). Athabasca University Press. https://www.aupress.ca/app/uploads/120155_99Z_Mohamed_Ally_2009-MobileLearning.pdf#page=45

Kukulska-Hulme, A., & Shield, L. (2008). An overview of mobile assisted language learning: From content delivery to supported collaboration and interaction. *ReCALL*, *20*(3), 271-289. https://doi.org/10.1017/S0958344008000335

Lockhart, K.S. (2016). *A comparison of the attitudes of administrators and teachers on cell phone use as an educational tool* (Order No. 10075104). Available from ProQuest Central; ProQuest Dissertations & Theses Global. (1777350856). https://www.proquest.com/dissertations-theses/comparison-attitudes-administrators-teachers-on/docview/1777350856/se-2

Losh, E. (2014). *The war on learning: Gaining ground in the digital university*. USA: MIT Press. https://mitpress.mit.edu/books/war-learning

Lowenthal, J.N. (2010). Using mobile learning: Determinates impacting behavioral intention. *The American Journal of Distance Education*, *24*(4), 195-206. https://doi.org/10.1080/08923647.2010.519947

Merriam, S.B., & Tisdell, E.J. (2015). *Qualitative research: A guide to design and implementatio*n. John Wiley & Sons. GoogleScholar

Motiwalla, L. (2007). Mobile learning: A framework and evaluation. *Computers and Education, 49*(3). 581-596. https://doi.org/10.1016/j.compedu.2005.10.011

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

Obringer. J. & Coffey. K. (2007). Cell phones in American high schools: A national survey. *The Journal of Technology Studies, 33*(1), 41-47. https://eric.ed.gov/?id=EJ847358

Purba, M., & Setyarini, S. (2020, October). Mobile Learning through WhatsApp: EFL Students' Perceptions. *12th International Conference on Education Technology and Computers* (pp. 27-32). https://doi.org/10.1145/3436756.3437016

Park, Y. (2011). A pedagogical framework for mobile learning: Categorizing educational applications of mobile technologies into four types. *The International Review of Research in Open and Distributed Learning*, *12*(2), 78-102. https://doi.org/10.19173/irrodl.v12i2.791

Parsons, D., & Ryu, H. (2006, April). *A framework for assessing the quality of mobile learning* [Paper presentation]. 11th International Conference for Process Improvement, Research and Education, Southampton, UK. Google Scholar

Peters, K. (2007). M-Learning: Positioning educators for a mobile, connected future. *International Review of Research in Open and Distance Learning*, *8*(2). https://doi.org/10.19173/irrodl.v8i2.350

Preacher, K.J., & MacCallum, R.C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding statistics: Statistical Issues in Psychology, Education, and the Social Sciences, 2*(1), 13-43. https://doi.org/10.1207/S15328031US0201_02

Quaglia, R., & Corso, M. (2014). *Student voice: The instrument of change*. Corwin: A Sage Company: Thousand Oaks, CA. https://us.corwin.com/en-us/nam/student-voice/book243538

Stevens, J.P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Hillsdale, NJ: Erlbaum. https://psycnet.apa.org/record/1992-98099-000

Tabachnick, B.G., & Fidell, L.S. (2013). *Using multivariate statistics.* Allyn and Bacon. https://www.pearsonhighered.com/assets/preface/0/1/3/4/0134790545.pdf

Traxler, J. (2007). Defining, discussing and evaluating mobile learning: The moving finger writes and having writ... *The International Review in Open and Distance Learning, 8*[2], 1–13. Google Scholar

Wang, Y.S., Wu, M.C., & Wang, H.Y. (2009). Investigating the determinants and age and gender differences in the acceptance of mobile learning. *British Journal of Educational Technology*, *40*(1), 92–118. https://doi.org/10.1111/j.1467-8535.2007.00809.x

Willis, G.B. (2004). *Cognitive interviewing: A tool for improving questionnaire design*. Sage Publications. https://www.jstor.org/stable/27641121

Valk, J.H., Rashid, A.T., & Elder, L. (2010). Using mobile phones to improve educational outcomes: An analysis of evidence from Asia. *The International Review of Research in Open and Distributed Learning*, *11*(1), 117-140. https://doi.org/10.19173/irrodl.v11i1.794

Venkatesh, V., & Davis, F.D. (2000). A theoretical extension of the technology acceptance model: four longitudinal field studies. *Management Science, 46*(2), 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

Vosloo, S. (2012). *Mobile learning and policies: Key issues to consider*. Paris, France: UNESCO. http://unesdoc.unesco.org/images/ 0021/002176/217638E.pdf

Yurdugül, H. (2007). The Effects of Different Correlation Types on Goodness-of-Fit Indices in First Order and Second Order Factor Analysis for Multiple Choice Test Data. *İlköğretim Online*, *6*(1), 154-179. https://ilkogretim-online.org/?mno=121223

Yurdugül, H., & Aşkar, P. (2008). An investigation of the factorial structures of pupils' attitude towards technology (PATT): A Turkish sample. *Elementary Education Online*, *7*(2), 288–309. http://ilkogretim-online.org/fulltext/218-1596636637.pdf?1612882858

Yurdugül, H., & Sırakaya, D.A. (2013). The scale of online learning readiness: A study of validity and reliability. *Education and Science*, *38*(169), 391-406. https://hdl.handle.net/20.500.12513/1731

APPENDIX

Table 9. *The last version of the mobile phone use in academic environment scale (MPUAES).*

Please use the 1-5 scale provided ("Not at all true" to "Extremely true") to rate how TRUE for YOU the following statements are.	1 – Not at all true	2 – A little true	3 – Somewhat true	4 – Very true	5 – Extremely true
1. My phone helps me keep track of -*follow*- my academic life.	□	□	□	□	□
2. I use my phone to connect with my classmates or instructors	□	□	□	□	□
3. I would get more school work done if I spent less time on my phone.	□	□	□	□	□
4. When it comes to the academic life, my phone is my personal assistant.	□	□	□	□	□
5. When I should be doing the school work, I find myself occupied with my phone.	□	□	□	□	□
6. I feel more comfortable in doing my school work when I have my phone with me.	□	□	□	□	□
7. For my academic life, I feel dependent on my phone.	□	□	□	□	□
8. In class or whenever I study, I read/send text messages that are not related what I am doing.	□	□	□	□	□
9. I find myself occupied on my phone even when I'm with my classmates or instructors (during the class or studying).	□	□	□	□	□
10. Having my phone with me makes it easier to sort out –*resolve, handle*- the critical situations related to my academic life.	□	□	□	□	□
11. I feel in control of my academic life when I have my phone with me.	□	□	□	□	□
12. My phone is necessary for my academic life.	□	□	□	□	□
13. Without my mobile phone, I feel detached -*out of touch, isolated*- to my academic life.	□	□	□	□	□
14. My phone helps me stay close to my classmates and instructors.	□	□	□	□	□
15. My phone makes it easy to cancel the arranged plans with classmates or instructors.	□	□	□	□	□
16. In my academic life, my phone gives me a sense of comfort.	□	□	□	□	□
17. My phone helps me be more organized for my academic life.	□	□	□	□	□
18. I find myself engaged with my mobile phone for longer than I intended.	□	□	□	□	□

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 4, 949–963
https://doi.org/10.21449/ijate.1002824
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Adaptation of Motivation to Read Profile Scale to Turkish

Zeynep Aydemir[1,*], **Ergun Ozturk**[2]

[1]Marmara University, Ataturk Faculty of Education, Department of Elementary Education, Istanbul, Türkiye
[2]Erciyes University, Faculty of Education, Department of Elementary Education, Kayseri, Türkiye

ARTICLE HISTORY
Received: Sep. 30, 2021
Revised: Oct. 02, 2022
Accepted: Nov. 23, 2022

Keywords:
Reading motivation,
Reading profile,
Motivation to read profile scale,
Scale adaptation,
Elementary education.

Abstract: The purpose of this research is to adapt the "Motivation to Read Profile Scale" developed by Malloy et al. (2013) into Turkish. Within the framework of adaptation studies, firstly, the items of the scale were translated into Turkish by the researchers, then ten experts were consulted for the Turkish and English forms of the scale, and amendments to the translation were made in line with their opinions. The scale was administered to 317 students for validity and reliability studies. Confirmatory factor analysis was performed directly on the two-factor scale, as the experimental evidence regarding the construct validity of the scale in the original culture was determined. As a result of the general confirmatory factor analysis, the two-factor structure can be characterized as having values that can be acceptable. Cronbach's alpha internal consistency coefficient for the Turkish form of the scale was 0.86. As a result, it was seen that the Turkish form of the scale was valid and reliable for this research group.

1. INTRODUCTION

Reading is a skill that affects the individual in primary education and further educational life, and it is accepted as an act of good behavior in the social environment. For this reason, beyond just reading the letters, it is an action that affects the education and social life of the person. According to Castleman and Littky (2007), the main factor underlying success in any academic field and lifelong learning is reading. The act of reading is a process that starts with the person making sense of the letters, and it is associated with making an effort and internalizing and enjoying it. When the concepts of loving and appreciating are brought together with the act of reading, conceptual structures such as the love of reading and the individual's appreciation of reading emerge. The element that includes these concepts is the concept of motivation. Motivation is defined as an impulse that activates purposeful behaviors and intentions (Ames, 1990; 1992).

Studies indicate that motivation is influenced by affective, social, and cognitive factors (Relan, 1992) and intertwined with interest, curiosity, and the desire to achieve something (Williams & Burden, 1997). To like something is not enough for motivation. At the same time, this interest

*CONTACT: Zeynep Aydemir ✉ zeynep.aydemir@marmara.edu.tr Marmara University, Faculty of Education, Department of Elementary Education, Istanbul, Türkiye

e-ISSN: 2148-7456 /© IJATE 2022

should be continuous/sustainable. Motivation is one of the cornerstones of learning. Therefore, it is one of the factors affecting reading. The sustainability of the process of making an effort to read and appreciating reading requires reading motivation. Reading motivation is a situation that affects individuals' behaviors enabling them to take action, interest, and desire to read (Mckenna, Kear & Ellsworth, 1995). The equivalent of the word profile in our language is stated as attitude or tendency (Turkish Language Association, 2005).

Motivation and profile concepts are two important factors that feed and affect each other in the development of reading skills (Marinak, et al., 2015). Individuals with high positive attitudes towards reading can read longer and more efficiently, as their curiosity and interest will be high throughout the reading process (Başaran, 2021). The well-being of the relationship between reading and the child is a phenomenon that emerges with the determination of the reading profile. Determining the reading profile of children at an early age can positively affect reading motivation. Although there are different types of reader profiles, there is a reading profile for motivation to read, too. (Marinak et al., 2015).

When the literature was examined, scales that were developed directly and indirectly related to the reading profile were found. One of the indirectly related scales is the reading self-concept questionnaire developed by Chapman and Tunmer (1995), which consists of three dimensions. The dimensions are stated as perceiving reading proficiency, perceiving reading difficulty, and attitude towards reading. The other scale belongs to McKenna et al. (1995) and is a 20-item scale for reading attitude that measures how much students read in their spare time and at school. The scale that includes the concept of reading profile, which is directly related to and more comprehensive than both scales, is the "Motivation to Read Profile (MRP)" developed by Gambrell et al. (1996). This scale is used to determine students' self-concepts as readers, their interests, and the value they attach to reading. The scale consists of 20 items under 2 sub-dimensions namely Self-Concept as a reader and Value of Reading. Another scale that is directly related is “Motivation to Read Profile-Revised” developed by Malloy et al. (2013). The MRP scale consists of 2 dimensions and 20 items: students’ Self-Concept as a Reader and Value of Reading, which includes items measuring how much students enjoy reading. This scale was chosen for the adaptation study because it is a comprehensive and updated version of other scales. One of the dimensions in the preferred scale is similar to the reading self-perception scale in Chapman and Tunmer's scale. McKenna et al. (1995) stay within the scope of the definition of the concept of profile with the scale he developed for the reading attitude.

In the aforementioned motivation to read profile, the expectations from the reader are self-awareness as a reader and value given to reading. The reading motivation profile includes the behaviors that students expect of themselves to be successful and motivated readers. In this dimension, there are questions about intrinsic and extrinsic motivation for students. The individual who defines the self-concept as a reader asks, “Am I a good reader? Am I a good reader according to my friends?” The individual who wants to measure her self-awareness as a reader thinks about and makes sense of her expectations and the expectations of her friends from her. It includes children's beliefs, expectations for success, and competencies. The question “Why do I want to be a good reader?” is about the reasons for the different activities that children do or cannot do. Competence and skill alone are not enough to increase success. The question “Do I want it?” is part of intrinsic and extrinsic motivation. The individual's expectations are related to the concept of self-efficacy (Wigfield & Guthrie, 1997). Students' self-efficacy beliefs are related to the performance-based environment (Eccles et al., 1993; Wigfield, Eccles & Rodriguez, 1998).

Self-concept as a reader includes how the individual does reading comprehension, what her interests and strategies are, and how to share them. Item 3, for example, asks students to decide how easily they can figure out new words, and items 7 and 13 tap into perceptions of reading

comprehension. Low scores for these items might suggest that individual or small-group follow-up is important to further isolate the difficulties experienced in decoding or comprehension strategy use that might lead to these perceptions of low self-efficacy for these tasks. Further exploration during the conversational interview might also help develop specific teaching plans for supporting these students. In the 17th question in the scale item, the student is asked to describe how he/she feels while talking about the books he/she reads with his/her friends. In the studies, talking about texts and supporting students on this subject are seen as a process that increases motivation for reading (Christie et al., 2009; Reznitskaya, 2012). Students may perceive their ability to read silently as very different from their ability to read aloud. Item 19 provides a window to student perceptions of reading aloud, and low scores here might suggest some need for the development of oral reading fluency, such as Readers Theatre, or practicing a piece for recording a VoiceThread or Podcast book recommendation.

The second important element for the motivation to read profile is the value given to reading. To understand the concept of value in the reading profile, the expectancy-value theory should be looked at. According to the expectancy-value theory of motivation, it can be said that the motivation affecting reading behaviors consists of expectations for reading. The individual's insistence, energy, performance, belief, interest, and value given to reading are important (Vroom, 1967). Studies have shown that children who appreciate reading have high reading motivation (Guthrie et al., 1996; Morgan & Fuchs, 2007). They also said that motivation is not only affected by pubertal (physical) changes, but also by the environment. It has been suggested that academic motivation, which is also the focus of the motivation to read profiles, emerges with the phenomenon called class context rather than individual structure. It is seen that especially teacher practices that affect the classroom context affect students positively (Urdan & Schönfelder, 2006). In the study of Bektaş, Okur, and Karadağ (2014), the concept of "reading a book" stands out in elementary school students' perceptions of the categories "helping to learn", "creating a fun environment", "providing freedom", "supporting" and "giving peace". It is seen that the metaphors that students attribute to the concept of reading and the scale items in the motivation to read profile overlap (items 4, 6, 10, and 16). Therefore, the purpose of this article is to lay emphasis on the Motivation to Read Profile (MRP) and to engage in a discussion of how periodic, classwide administration of the MRP can inform practices to support motivating classroom contexts. It is not enough to tell students that reading is valuable. It is necessary to be a practical role model for them and to create authentic environments. Roberts and Wilson's (2006) question "Do the teaching methods or materials we use to encourage students to read?" becomes important at this point. The studies in the literature show that interactions such as increasing students' interactions with the real world, using interesting books and materials, supporting their choices, increasing cooperation among students, creating a teacher-controlled classroom context, and increasing interest affect reading motivation, reading amount and text comprehension processes positively (Ateş, 2011; Guthrie & Alao, 1997; Guthrie & Davis, 2003; Guthrie & Wigfield, 2000; Hidi & Harackiewicz, 2000; Köroğlu, 2021; Reynolds & Symons, 2001; Schraw & Dennison, 1994; Skinner, Wellborn, & Connell, 1990; Wentzel, 1993). Students prefer to read texts in which heroes are similar to themselves, look at scenes similar to their environment, or read about problems similar to theirs (Başaran, 2007).

Reading can also be valued as an achievable goal that is important to a student's future perspective. In this sense, becoming a good reader is valued because it can lead to a career or professional interest (Malloy et al., 2013). Items 8 and 12, in particular, indicate a student's perception that becoming a good reader is valuable to their future goals. For example, if several students in the class respond to item 10 "I think libraries are_____," with "a boring place to spend time", then the teacher should carefully consider ways that students use the library (Malloy et al., 2013). Different methods and materials should be chosen that encourage

students to read more and make reading fun. Students should be invited to literacy activities to have fun, find what they want, share what they have read, to learn about life issues (Marinak et al., 2012; Malloy et al., 2013).

An integrated resilience approach that covers past experiences and plans for the future should be prioritized for the formation of a culture of reading and literacy. When children start school, they are eager to learn. However, as the grade levels progress, it is seen that their learning and academic motivation decrease in many subjects, including reading (Eccles et al., 2006; Edmunds & Bauserman, 2006). To investigate the reasons for the decrease in reading motivation as the grade level progresses and to meet the learning reading needs of the students effectively, the reading motivations and the reading profiles that allow for determining the reading motivations should be evaluated correctly.

According to Rueda, Au, and Choi (2004), the importance of evaluating reading motivation is to inform teachers about how students acquire their reading motivation and how to become active readers. It was necessary to develop measurement tools to determine the relations of the students with reading and to take precautions for the determined situations. The Motivation to Read Profile (MRP; Malloy et al. 2013) is a scale designed to guide teachers about the value their students place on reading and their reading self-concept as a reader. The scale, which is intended to be adapted, is used to determine students' self-concept as readers and the value given to reading. Determining the children's reading profiles at an early age and supporting measuring their reading motivation can be realized together with the increase in awareness of teachers, families, and schools on this issue. In addition, early detection of children's reading-related status is important in terms of intervening in their reading success, the value given to reading, and their competence in reading. It is thought that this scale will provide important findings in determining and increasing students' reading motivation and will help in the process. The study aims to adapt "The Motivation to Read Profile Scale" developed by Malloy et al. (2013) into Turkish and to determine the motivation to read profiles of second, third, fourth, fifth, and sixth grade students in elementary school with this adapted scale.

2. METHOD

The research is a scale development study. A total of 317 students from the second, third, fourth, fifth, and sixth grades of a primary school in Istanbul were selected as the study group in the adaptation studies of the motivation to read profile scale. For factor analysis, it is stated that when the sample size is 200, it is medium and 300 is good (Tabachnick & Fidell, 2007). The sample size in the study is seen as an appropriate number. Of the students participating in the scale adaptation study, 167 (52.7%) were female and 150 (47.3%) were male students. Of 317 students, 25 (7.9%) were second graders, 122 (38.5%) were third graders, 130 (41%) were fourth graders, 26 (8.2%) were fifth graders, and 14 (4.4%) were sixth graders is at the grade level.

2.1. Data Collection Tools and Analysis

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: 2021/24) at the beginning of the research. The Motivation to Read Profile Scale was developed by Malloy et al. in 2013 and its structure was tested with confirmatory factor analysis in a group of students from the second grade to the sixth grade. The scale, consisting of 20 items and 2 factors, was published in the journal "The Reading Teacher" published by the International Literacy Foundation in 2013, and the scale was obtained from this article. It was decided to adapt the examined scale. After obtaining the necessary permission for the adaptation of the scale from Jacquelynn B. Malloy,

Barbara A. Marinak, Linda B. Gambrell, and Susan A. Mazzoni, who developed the scale, via e-mail, adaptation studies for the scale started.

The items that received 100% trait agreement were included in the field testing of the original MRP with 330 students from third to fifth grades from 4 eastern U.S. schools. The scales were found to be reliable (self-concept = .75; value = .82). The reading survey was designed as a self-report instrument that could be administered to the whole class or a small group, depending on the teacher support required. The four-point ordinal scale includes ranked responses with 10 items for each subscale. Self-concept as a reader is assessed through items such as, "I think I am a ____ reader" and "When I have trouble figuring out a word I don't know, I…". Items that are designed to tap the value of reading include "Reading is something I like to do….", and "My friends think reading is…". The reading survey was administered to students in three schools in the mid-Atlantic and Southern regions of the United States—one in Virginia, one in Pennsylvania, and one in South Carolina. In all, 118 third graders, 104 fourth graders, and 54 fifth graders submitted permission to take the MRP-R, resulting in 281 students. Student scores were loaded into a spreadsheet, and validity and reliability testing was conducted using Mplus statistical software. Reliability testing using Cronbach's alpha revealed an $\alpha = .87$ for the full scale, an $\alpha = .85$ for the value subscale, and an $\alpha = .81$ for the self-concept scale. As the scale for the survey items was ordinal, it was decided to determine validity using a root mean square error of approximation (RMSEA). An RMSEA estimate of .089 was revealed with a confidence interval of .081 – .098. The probability of RMSEA $\leq = .05$ was .000. Considering the ordinal nature of the survey scale, reliability and validity estimates are judged to be well within acceptable ranges for both classroom use and research purposes.

The scale was administered to 118 third grade, 104 fourth grade, and 54 fifth grade students. The scale consists of 20 items under two sub-dimensions: self-concepts as a reader (10 items) and value of reading (10 items). The total reliability coefficient of the scale is .87. While the reliability coefficient for the value sub-dimension is .85, and it is .81 for the self-concept dimension. Non-parametric analyses were used when the questionnaire items were ordinal. The estimated RMSEA value is .089, and the confidence interval values are .081 - .098. It is stated that the RMSA value is significant at the .05 level. A variable response scale form was used to increase the reliability of the scale. The answers to the scale items were determined starting from the least motivation level to the maximum or vice versa. The scoring is 1-4.

2.2. The Adaptation Process of the Scale to Turkish

It is possible to examine the procedures for the adaptation of the scale to Turkish in two parts. The first part includes the process of translating the scale into Turkish and receiving expert opinions. In the second part, validity and reliability analyses were made by applying the scale to the student. The translation of the scale into Turkish was carried out by the researchers. After the translation by the authors, the scale, which was translated into a structure containing the original items, the translated items, and the suggestions to be made, was distributed to the experts to get their opinions. Academicians working in the fields of English (5), Turkish (3), and Measurement and Evaluation (2) were consulted for expert opinions.

For each item of the form given for the expert opinion, the expressions "not suitable", "partially appropriate", "appropriate", and "completely appropriate" were included and the experts were asked to mark whether each item was appropriate or not. For each item, 80% completely appropriate or appropriate expression was sought, and the items below this rate were corrected in line with the suggestions received from the experts. After the changes, the scale was redistributed to the same experts, and their opinions were taken, and it was concluded that all items were suitable by at least 80%. Turkish and English versions of the scales might be assumed equivalent because the correlations between the English and Turkish versions are found to be .89.

In cross-cultural scale adaptation studies, it may be recommended to start the tool with a direct confirmatory factor analysis for the factor pattern in the target culture. Because the factor pattern of the mentioned tool in the original culture has been revealed by many qualitative and quantitative studies, the empirical evidence for the construct validity of the tool has been determined. At this point, whether the factor pattern of the instrument is also preserved in the target culture can be questioned by testing it with confirmatory factor analysis. If the model related to the original factor pattern of the tool is not confirmed or does not give high fit indices in the confirmatory factor analysis to be made, then the factor pattern in the target culture can be explored with exploratory factor analysis (Çokluk, Şekercioğlu & Büyüköztürk, 2018, p.283). For this reason, the scale was applied to 317 students for validity and reliability studies, construct validity was analysed with confirmatory factor analysis, and reliability analysis was performed with Cronbach's alpha internal consistency coefficient. Confirmatory factor analysis processes were carried out with the help of the Lisrel 8.54 package program.

3. RESULT

While adapting the scale, confirmatory factor analysis was used to examine the compatibility of the scale's structure with the collected data in Turkish students.

3.1. Findings on Confirmatory Factor Analysis (CFA)

In the CFA, first of all, the compatibility of the two-factor model of the original scale with 20 items was tested. First of all, operations were carried out without limiting the model and adding a connection. The standard solution, T, and R^2 values of each item as a result of the DFA processes are given in Table 1.

Table 1. *CFA Sd, T and* R^2 *Results.*

Item	*Sd*	T	R^2	Item	*Sd*	T	R^2
I1	.71	13.75	.51	I13	.52	9.33	.27
I2	.59	10.64	.35	I14	.50	8.78	.25
I3	.40	6.94	.16	I15	.63	11.78	.40
I4	.31	5.22	.098	I16	.42	7.22	.18
I5	.54	9.78	.30	I17	.40	6.91	.16
I6	.41	7.03	.17	I18	.60	10.73	.36
I7	.38	6.55	.14	I19	.67	12.77	.45
I8	.32	5.40	.11	I20	.61	11.10	.38
I9	.77	15.45	.60				
I10	.52	9.02	.27				
I12	.50	8.73	.25				

The Items classified under two factors in CFA were observed to have standard solution values between .31 and .77. Besides, the items were found to have R^2 values between .098 and .60. Since these are of high standard solution values, the items under all factors were considered to be important for their factors. Item 11 was removed from the scale because its values were obtained low. Following the standard solutions, *t* values between factors and items were analysed. Jöreskog and Sörbom (1996) mentioned that the lack of red arrows regarding the *t* values shows that all items are significant at the level of .05. It was found that the items had *t* values between 5.22 and 15.45, and these values are significant at the level of .01 in Figure 1. As a result of the analysis, the fit indices were: $\chi^2 = 337.20$ ($p. = .00$), $\chi^2/sd = 2.23$ RMSEA = .063, SRMR = .056, GFI = .90, AGFI = .87, CFI = .95, NFI = .92 and NNFI = .95. The fit index values suggested by Schermelleh-Engel, Moosbrugger, and Müller (2003) were taken as the

basis for the evaluation of the results obtained for the model. The suggested values are given in Table 2.

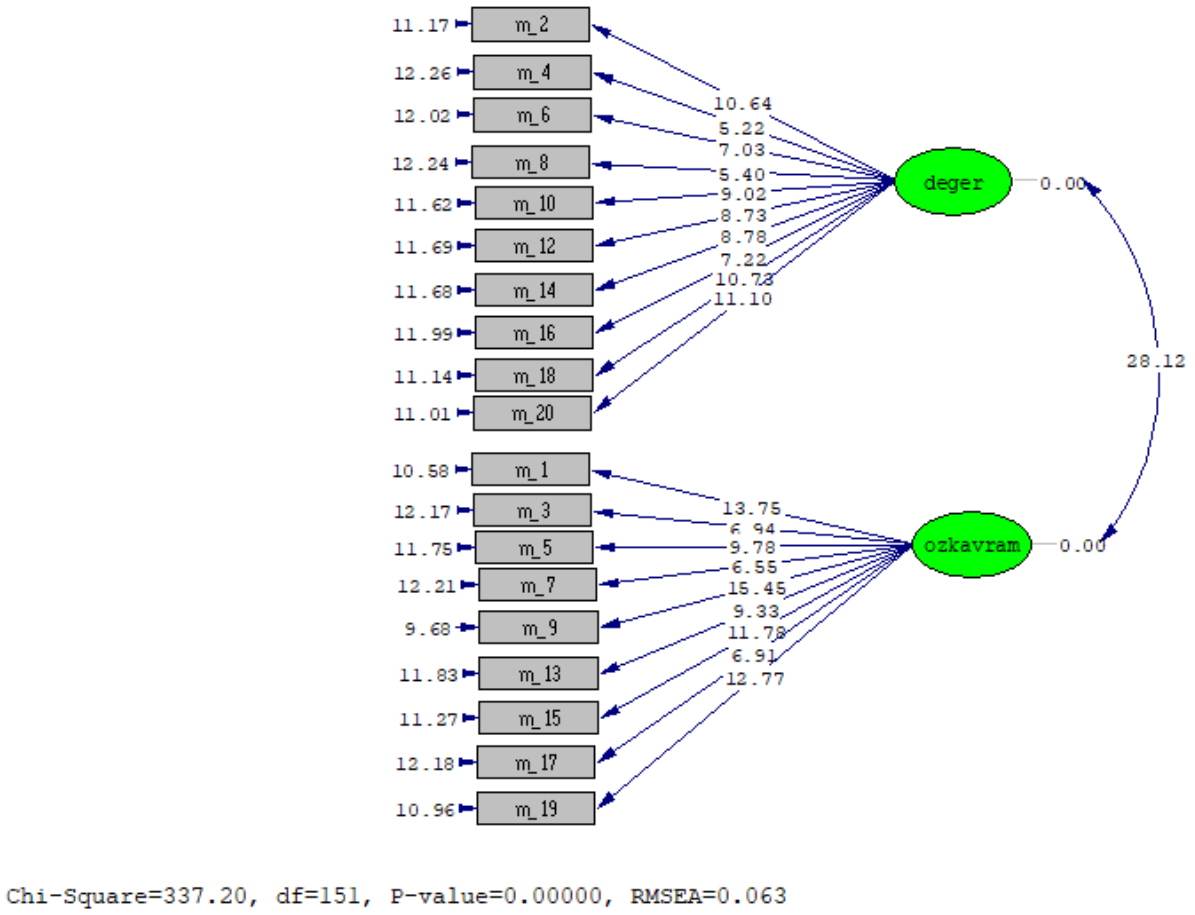
Table 2. *Model fit Indexes proposed by Schermelleh-Engel, Moosbrugger, and Müller (2003)*

Reviewed indices of fit	Perfect fit criteria	Acceptable fit criteria
χ^2/sd	$0 \leq \chi^2/sd \leq 2$	$2 < \chi^2/df \leq 3$
RMSEA	$0 \leq$ RMSEA $\leq .05$	$.05 <$ RMSEA $\leq .08$
SRMR	$0 \leq$ SRMR $\leq .05$	$.05 <$ SRMR $\leq .10$
CFI	$.97 \leq$ CFI ≤ 1	$.95 \leq$ CFI $< .97$
NFI	$.95 \leq$ NFI ≤ 1	$.090 \leq$ NFI $< .95$
NNFI	$.97 \leq$ NNFI ≤ 1	$.095 \leq$ NNFI $< .97$
GFI	$.95 \leq$ GFI ≤ 1	$.90 \leq$ GFI $< .95$
AGFI	$.90 \leq$ AGFI ≤ 1	$.85 \leq$ AGFI $< .90$

AGFI = Adjusted Goodness of Fit Index, CFI = Comparative Fit Index, GFI = Goodness of Fit Index, NFI = Normed Fit Index, NNFI = Nonnormed Fit Index, RMSEA = Root Mean Square Error of Approximation, SRMR = Standardized Root Mean Square Residual

When the fit index values of the scale, which were evaluated with a two-factor structure, were examined, it was found that χ^2/sd, good fit, SRMR, RMSEA, CFI, NFI, NNFI, GFI (.90), and AGFI (.87) indexes had acceptable fit values. In general, the two-factor structure can be characterized as having values that will show an acceptable fit.

Figure 1. *Measurement model for the scale.*



3.2. Reliability

The reliability of the scale was evaluated with Cronbach's alpha internal consistency coefficient. Cronbach's alpha value for the whole scale was found .86. This value for "Self-concept as a Reader", one of the sub-factors of the scale, was .81 and "Value of reading" was found to have a reliability value of .75. The fact that all internal consistency values are higher than .84 indicates that the reliability values of the scale are high, that is, it produces consistent data.

4. DISCUSSION and CONCLUSION

The aim of this research is to adapt the "Motivation to Read Profile Scale" developed by Malloy et al. (2013) into Turkish. For this purpose, the model fit of the Turkish form of the scale was examined by confirmatory factor analysis. As a result of the confirmatory factor analysis, the t values of the items were found acceptable except for the 11th item. After examining the 11th item, "I worry about what other kids think about my reading", it is thought that this item and the 1st item in the self-concept as a reader dimension are similar. The t values of the scale except for the 11th ranged between 5.22 and 15.45, and they were found to be significant at the .01 level as they were higher than 2.76. According to Jöreskog and Sörbom (1996), the absence of a red arrow related to t values indicates that the items are significant at the .05 level. In addition, it was found that the items other than item 11 had R^2 values between .098 and .60. Since these values have high solution values, it was decided that the items in all factors except the 11th item were important for the factors. As a result of the analysis, fit indices were χ^2 = 337. 20 (p. = .00), χ^2/sd = 2.23 RMSEA = .063, SRMR = .056, GFI = .90, AGFI = .87, CFI = .95, NFI = 0.92 and NNFI = .95. In the original form of the scale, the RMSEA estimated value is .089 and the confidence interval values are .081 - .098. It is stated that the RMSA value is significant at the .05 level. In Turkish, fit indices are acceptable (Byrne, 1998). In this respect, it has been revealed that the structure of the Turkish form of the scale has acceptable fit index values.

Cronbach's alpha internal consistency coefficients were checked for consistency in the reliability of the scale. Cronbach's alpha value for the entire scale was found .86. The coefficient for "Self-concept as a reader", one of the sub-factors of the scale, was .81, and for the "Value of reading", it was found to have a reliability value of .75. All internal consistency values of .84 and higher indicate that the scale has high-reliability values, that is, it produces consistent data. The total reliability coefficient in the original form of the scale is .87. While the reliability coefficient for the "value" sub-dimension is .85, it is .81 for the self-concept dimension (Malloy et al., 2013). The internal consistency coefficient of the Turkish version was .86, indicating that it is a good value for reliability (Green & Salkind, 2005). The internal consistency coefficients of the original form are close to the values obtained in the Turkish form.

As a result of the research, the "Motivation to Read Profile Scale" developed by Malloy et al. (2013) was adapted into Turkish. The adapted Turkish form was found to have a similar structure to the original form by removing only one item. Although the psychometric properties obtained from the Turkish form were quite suitable for a scale, some values were higher than the original form and some were lower. The "Motivational Profile (MRP)" scale developed by Gambrell, et al. (1996) and adapted by Yıldız (2013) originally consisted of 20 items, yet it was adapted into two sub-dimensions: the value of reading and the self-concept as a reader with 18 items. The reliability of the scale was found to be satisfactory (α = .81). Motivation to Read Profile-Turkish Form (MRP-TR) contained 9 items related to value of reading and 9 items related to self-concept as a reader. It is emphasized that the scales transferred from one language to another language undergo cultural changes, so they cannot be understood as in the original language, and their values may differ (Geisinger, 1994; Hambleton, Merenda, & Spielberger, 2005; Sireci & Berberoğlu, 2000). As a result, a 19-item scale consisting of two factors was

obtained. In this study, the meanings and contents attributed to the concept of reading were understood differently and applied to a different group from the original study group, which can be seen as the source of the difference. As a result of this study, a valid and reliably adapted scale emerged. It is recommended that the motivation to read profile scale be applied at the beginning and middle of each year from the second grade to the sixth grade levels to identify the factors that affect the reading motivation of the student and to guide the teacher (Appendix A-B). Just as an informal reading inventory or benchmark, assessment gives you a read on the pulse of what your students can do or already know, a quick check of their motivation at the beginning and midpoint of the school year may guide you in tailoring instruction that will support student motivation and engagement in literacy learning. The MRP is a tool available to teachers that will guide them in developing instructional practices that support students in becoming engaged and strategic readers for both personal and academic literacy needs.

Acknowledgments

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

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Erciyes University/Social and Humanities Ethics Committee, 2021/24.

Authorship Contribution Statement

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

Orcid

Zeynep Aydemir https://orcid.org/0000-0003-3002-1809
Ergun Ozturk https://orcid.org/0000-0002-4800-8437

REFERENCES

Ames, C.A. (1990). Motivation: What teachers need to know. *Teachers College Record, 91(3)*, 409-421.

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84(3),* 261-271. https://doi.org/10.1037/0022-0663.84.3.261

Ateş, S. (2011). *İlköğretim beşinci sınıf Türkçe dersi öğrenme-öğretme sürecinin anlama öğretimi açısından değerlendirilmesi* [Doctoral dissertation]. Gazi University.

Başaran, M. (2007). İlköğretim beşinci sınıf öğrencilerinin hikâye edici metinlere ilişkin tercihleri. [Doctoral dissertation, Gazi University].

Başaran, M. (2021). Okuduğunu anlayamayan öğrencilerin okuma esnasındaki bilişsel davranışları ve duygu durumları. *Ana Dili Eğitimi Dergisi, 9(1),* 45-58. https://doi.org/10.16916/aded.802475

Bektaş, M., Okur, A., & Karadağ, B. (2014). İlkokul ve Ortaokul Son Sınıf Öğrencilerinde Metaforik Algı Olarak Kitap [Book as a Metaphoric Perception in Last Class of'the Primary and Secondary Students]. *Türk Kütüphaneciliği,28(2)*, 154-168. http://dergipark.org.tr/tr/pub/tk/issue/48764/620447

Byrne, B.M. (1998). *Structural equation modeling with LISREL, PRELIMS and SIMPLIS: Basic concepts, applications, and programmings*. London: Lawrence Erlbaum Assocatiates, Publishers.

Chapman, J.W., & Tunmer, W.E. (1995). Development of young children's reading self-concepts: An examination of emerging subcomponents and their relationship with reading

achievement. *Journal of Educational Psychology, 87(1)*, 154-167. https://doi.org/10.1037/0022-0663.87.1.154

Castleman, B., & Littky, D. (2007). Learning to love learning. *Educational Leadership, 64(8)*, 58-61.

Christie, D., Tolmie, A., Thurston, A., Howe, C., & Topping, K. (2009). Supporting group work in Scottish primary classrooms: Improving the quality of collaborative dialogue. *Cambridge Journal of Education, 39(1)*, 141-156. https://doi.org/10.1080/03057640802702000

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2018). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları*. Pegem akademi.

Eccles, J., Wigfield, A., Harold, R.D., & Blumenfeld, P. (1993). Age and gender differences in children's self-and task perceptions during elementary school. *Child Development, 64(3),* 830-847. https://doi.org/10.2307/1131221

Eccles, J.S., Wigfield, A., Schiefele, U., Roeser R.W., & Kean, P.D. (2006). The development of achievement motivation. In W. Damon, Richard M. Lerner ve N. Eisenberg (Eds.). *Handbook of Child Psychology: Social, emotional, and personality development*. (Sixth Edition), pp. 934-988. Wiley.

Edmunds, K.M., & Bauserman, K.L. (2006). What teachers can learn about reading motivation through conversations with children. *The Reading Teacher, 59(5),* 414–424. https://doi.org/10.1598/RT.59.5.1

Gambrell, L.B., Palmer, B.M., Codling, R.M., & Mazzoni, S.A. (1996). Assessing motivation to read. *The Reading Teacher, 49(7)*, 518-533. https://doi.org/10.1598/RT.49.7.2

Geisinger, K.F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6(4),* 304-312. https://doi.org/10.1037/1040-3590.6.4.304

Green, S.B., & Salkind, N.J. (2005). *Using SPSS for Windows and Macintosh: Analyzing and Understanding Data (4th ed)*. Pearson.

Guthrie, J.T., & Alao, S. (1997). Designing contexts to increase motivations for reading. *Educational Psychologist, 32(2),* 95- 105. https://doi.org/10.1207/s15326985ep3202_4

Guthrie, J.T., & Davis, M.H. (2003). Motivating struggling readers in middle school through an engagement model of classroom practice. *Reading and Writing Quarterly, 19*, 59–85. https://doi.org/10.1080/10573560308203

Guthrie, J.T., Van Meter, P., Mccann, A. D., Wigfield, A., Bennett, L., Poundstone, C.C., Rice, M.E., Faibisch, F.M., Hunt, B., & Mitchell, A.M. (1996). Growth in literacy engagement: Changes in motivations and strategies during concept-oriented reading instruction. *Reading Research Quarterly, 31(3)*, 306-332. https://doi.org/10.1598/RRQ.31.3.5

Guthrie, J.T., & Wigfield, A. (2000). *Engagement and Motivation in Readin*g. In M.L. Kamil, P.B. Masenthal, P.D. Pearson & R. Bırr (Eds.), Reading Research Handbook (Vol III, pp.403-424), Mahwah, NJ: Erlbaum.

Hambleton, R.K., Merenda, P.F., & Spielberger, C.D., (2005). *Adapting educational and psychological tests for cross-cultural assessment.* Mahwah, NJ: Lawrence Erlbaum.

Hidi, S., & Harackiewicz, J.M. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research, 70,* 151–179. https://doi.org/10.3102/00346543070002151

Jöreskog, K., & Sörbom, D. (1996). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Chicago: Scientific Software International/Erlbaum.

Köroğlu, M. (2021). *Türkçe öğretmenlerinin Türkçe dersi öğretimi sürecinde anlama (okuma) öğretimine yönelik uygulamaları* [Doctoral dissertation]. Hatay Mustafa Kemal University.

Malloy, J.A., Marinak, B.A., Gambrell, L.B., & Mazzoni, S.A. (2013). Assessing motivation to read: The motivation to read profile–revised. *The Reading Teacher, 67(4),* 273–282. https://doi.org/10.1002/trtr.1215

Marinak, B.A., Malloy, J. B., Gambrell, L.B., & Mazzoni, S.A. (2015). Me and My reading profile: A Tool for Assessing Early Reading Motivation. *The Reading Teacher, 69(1),* 51-62. https://doi.org/10.1002/trtr.1362

Mckenna, M.C., Kear, D.J., & Ellsworth, R.A. (1995). Children's Attitudes Toward Reading: A National Survey. *Reading Research Quarterly, 30(4),* 934-956. https://doi.org/10.2307/748205

Morgan, P.L., & Fuchs, D. (2007). Is there a bidirectional relationship between children's reading skills and reading motivation? *Exceptional Children, 73(2),* 165-183. https://doi.org/10.1177/001440290707300203

Relan, A. (1992). Motivational Strategies in Computer-based Instruction: Some Lessons from Theories and Models of Motivation. In proceedings of selected research and development presentations at the Convention of the Association for Educational Communications and Technology,1994. ERIC Document Reproduction Service No. ED 348 017

Reynolds, P.L., & Symons, S. (2001). Motivational variables and children's text search. Journal of Educational Psychology, 93(1), 14–22. https://doi.org/10.1037/0022-0663.93.1.14

Reznitskaya, A. (2012). Dialogic teaching: Rethinking language use during literature discussions. *The Reading Teacher, 65(7*), 446-456. https://doi.org/10.1002/TRTR.01066

Roberts, M.S., & Wilson, J.D. (2006). Reading attitudes and instructional methodology: How might achievement become affected? *Reading Improvement, 43(2*), 64-69.

Wigfield A., Eccles J.S., & Rodriguez D. (1998). The Development of Children's Motivation in School Contexts. *Review of Research in Education, 23,* 73-118. https://doi.org/10.2307/1167288

Rueda, R., Au, J., ve Choi, S. (2004). Motivation to read: Comparing teachers' perceptions of students' motivation with students' self-reported motivation – A pilot study. In Y.B. Kafai, W.A. Sandoval, N. Enyedy, A. Scott Nixon, & F. Herrera (Eds.), Proceedings of the Sixth International Conference of the Learning Sciences: Embracing diversity in the Learning Sciences. (pp. 443-448). Mahwah, NJ: Lawrence Erlbaum Associates.

Sireci, S.G., & Berberoglu, G. (2000). Using bilingual respondents to evaluate translated-adapted items. *Applied Measurement in Education, 13(3*), 229-248. https://doi.org/10.1207/S15324818AME1303_1

Schraw, G., & Dennison, R.S. (1994). The effect of reader purpose on interest and recall. *Journal of Reading Behavior, 26(1),* 1-18. https://doi.org/10.1080/10862969409547834

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods Of Psychological Research Online, 8(2*), 23-74.

Skinner, E.A., Wellborn, J.G., & Connell, J.P. (1990). What it takes to do well in school and whether I've got it: A process model of perceived control and children's engagement and achievement in school. *Journal of Educational Psychology, 82(1),* 22–32. https://doi.org/10.1037/0022-0663.82.1.22

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

Turkish Language Association. (2005). *Turkish dictionary (10th ed).* Ankara: TDK.

Urdan, T., & Schöenfelder, E. (2006). Classroom effects on student motivation: Goal structures, social relationships, and competence beliefs. *Journal of School Psychology, 44 (5),* 331 – 349. https://doi.org/10.1016/j.jsp.2006.04.003

Vroom, V.H. (1967). *Work and motivation (3th ed).* John Wiley&Sons.

Wentzel, K.R. (1993). Motivation and achievement in early adolescence: The role of multiple classroom goals. *Journal of Early Adolescence, 13(1),* 4-20. https://doi.org/10.1177/0272431693013001001

Williams, M., & Burden, R.L. (1997). *Psychology for language teachers: A social constructivist approach*. Cambridge University Press.

Wigfield, A., & Guthrie, J.T. (1997). Relations of children's motivation for reading to the amount and breadth of their reading. *Journal of Educational Psychology, 89(3)*, 420-432. https://doi.org/10.1037/0022-0663.89.3.420

Yıldız, M. (2013). Adaptation of the motivation to read profile to Turkish. International Journal of Academic Research, *Part B, 5(4)*, 196-199. http://doi.org/10.7813/2075-4124.2013/5-4/B.29

APPENDIX

Appendix-A. Turkish Version of The Motivation to Read Profile Scale

OKUMA MOTİVASYONU PROFİLİ ÖLÇEĞİ
Hangi sınıftasın? ☐ **2. Sınıf** ☐ **3 . Sınıf** ☐ **4. Sınıf** ☐ **5. Sınıf** ☐ **6. Sınıf**
Cinsiyet ☐ **Kız** ☐ **Erkek**
1.Arkadaşlarım benim ……. olduğumu düşünür.
☐ Çok iyi bir okuyucu ☐ İyi bir okuyucu ☐ Ortalama okuyucu ☐ Zayıf bir okuyucu
2. Kitap okumak hoşlandığım bir etkinliktir.
☐ Asla ☐ Hemen hemen hiç ☐ Bazen ☐ Sık sık
3. Bilmediğim bir kelime ile karşılaştığımda, …….....
☐ Neredeyse her zaman bir anlam bulabilirim. ☐ Bazen anlam bulabilirim. ☐ Hemen hemen hiç anlam bulamam. ☐ Asla anlam bulamam.
4. **Arkadaşlarım okumanın …… düşünür**.
☐ Gerçekten eğlenceli olduğunu ☐ Eğlenceli olduğunu ☐ Kısmen eğlenceli olduğunu ☐ Hiç eğlenceli olmadığını
5. Ben …… okurum.
☐ Arkadaşlarım kadar iyi olmasa da ☐ Arkadaşlarımla aynı seviyede ☐ Arkadaşlarımdan biraz daha iyi ☐ Arkadaşlarımdan çok daha iyi
6. Arkadaşlarıma okuduğum güzel kitapları anlatırım.
☐ Hiç yapmam ☐ Neredeyse hiç yapmam ☐ Bazen yaparım ☐ Çok yaparım
7. Tek başıma okurken, ……
☐ Okuduğum her şeyi anlarım. ☐ Neredeyse okuduğum her şeyi anlarım. ☐ Neredeyse okuduğum şeylerin hiçbirini anlamam. ☐ Okuduğum şeylerin hiçbirini anlamam.
8. Çok okuyan insanlar ……..
☐ Çok ilginçtir. ☐ Biraz ilginçtir. ☐ Biraz sıkıcıdır. ☐ Çok sıkıcıdır.
9. Ben ……
☐ Zayıf bir okuyucuyum. ☐ Orta düzeyde bir okuyucuyum. ☐ İyi bir okuyucuyum. ☐ Çok iyi bir okuyucuyum.

10. Bence kütüphaneler ……. .

- ☐ Vakit geçirmek için kesinlikle harika bir yerdir.
- ☐ Vakit geçirmek için harika bir yerdir.
- ☐ Vakit geçirmek için sıkıcı bir yerdir.
- ☐ Vakit geçirmek için gerçekten sıkıcı bir yerdir.

***11. Arkadaşlarımın benim okumamla ilgili ne düşündüklerini merak ederim.**

Çok Bazen Neredeyse hiç merak etmem Asla merak etmem

12. İyi bir okuyucu olmanın …… düşünürüm.

- ☐ Hiç önemli olmadığını
- ☐ Biraz önemli olduğunu
- ☐ Önemli olduğunu
- ☐ Çok önemli olduğunu

13. Öğretmenim bana ne okuduğumu sorduğunda …….. .

- ☐ Asla bir cevap veremiyorum.
- ☐ Neredeyse hiçbir cevap veremiyorum.
- ☐ Bazen cevap verebilirim.
- ☐ Daima cevap verebilirim.

14. Okumak için zaman harcamanın ……..

- ☐ Gerçekten sıkıcı olduğunu düşünürüm.
- ☐ Sıkıcı olduğunu düşünürüm.
- ☐ Harika olduğunu düşünürüm.
- ☐ Gerçekten harika olduğunu düşünürüm.

15. Okuma benim için …… .

- ☐ Çok kolaydır.
- ☐ Biraz kolaydır.
- ☐ Biraz zordur.
- ☐ Çok zordur.

16. Öğretmenim kitapları sesli bir şekilde okuduğunda, …….. düşünürüm.

- ☐ Gerçekten harika olduğunu
- ☐ Harika olduğunu
- ☐ Sıkıcı olduğunu
- ☐ Gerçekten sıkıcı olduğunu

17. Arkadaşlarımla okuduğum kitaplar hakkında konuşurken …… .

- ☐ Fikirlerimi söylemekten nefret ederim.
- ☐ Fikirlerimi söylemekten hoşlanmıyorum.
- ☐ Fikirlerimi söylemekten hoşlanırım.
- ☐ Fikirlerimi söylemeye bayılırım.

18. Boş zamanım olduğunda, ……,…… .

- ☐ Zamanımı hiç okumakla geçirmem.
- ☐ Zamanımın çok azını okumakla geçiririm.
- ☐ Zamanımın bir kısmını okumakla geçiririm.
- ☐ Zamanımın çoğunu okumakla geçiririm.

19. Sesli okuma yaptığımda, ……. olurum.

- ☐ Zayıf okuyucu
- ☐ Kısmen iyi bir okuyucu
- ☐ İyi bir okuyucu
- ☐ Çok iyi bir okuyucu

20. Birisi bana hediye olarak kitap verdiğinde ……olurum.

- ☐ Çok mutlu
- ☐ Mutlu
- ☐ Mutsuz
- ☐ Çok mutsuz

* removed item

Appendix-B. Turkish Version of MRP Scoring Guidelines

Okuma Motivasyonu Profili Ölçeği Puanlama Tablosu

Ölçekte yer alan maddeler 1-4 arası puanlanmaktadır. Ölçek maddelerinin hangi alt boyutta yer aldığı göstermek için Okuyucu olarak öz kavram için (ÖK) ve Okumaya verilen değer için (D) kısaltmaları kullanılmıştır.

Madde numarası ve alt boyut	**1.Seçenek**	**2.Seçenek**	**3.Seçenek**	**4. seçenek**
1 ÖK	4	3	2	1
2 D	1	2	3	4
3 ÖK	4	3	2	1
4 D	4	3	2	1
5 ÖK	1	2	3	4
6 D	1	2	3	4
7 ÖK	4	3	2	1
8 D	4	3	2	1
9 ÖK	1	2	3	4
10 D	4	3	2	1
12 D	1	2	3	4
13 ÖK	1	2	3	4
14 D	1	2	3	4
15 ÖK	4	3	2	1
16 D	4	3	2	1
17 ÖK	1	2	3	4
18 D	1	2	3	4
19 ÖK	1	2	3	4
20 D	4	3	2	1

International Journal of Assessment Tools in Education
2022, Vol. 9, *No. 4*, 964–997

https://doi.org/10.21449/ijate.948489

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

A rating scale development study for the evaluation of lesson plans and teaching practices on argumentation-based inquiry

Funda Yesildag Hasancebi[1,*], **Busra Tuncay Yuksel**[2], **Gunkut Mesci**[3]

[1]Giresun University, Faculty of Education, Department of Science Education, Giresun, Türkiye.
[2]Giresun University, Faculty of Education, Department of Science Education, Giresun, Türkiye.
[3]Giresun University, Faculty of Education, Department of Science Education, Giresun, Türkiye.

ARTICLE HISTORY

Received: June 6, 2022
Revised: Dec. 3, 2022
Accepted: Dec. 6, 2022

Keywords:
Argumentation-based inquiry,
Science writing heuristic,
Lesson plans,
Rating scale.

Abstract: The purpose of this study was to develop a reliable and valid rating scale for the use of the assessment and evaluation of lesson plans and teaching practices that are based on argumentation-based inquiry (ABI). The study covered two academic years (four academic semesters). Qualitative and quantitative methods were utilized throughout the development of the rating scale including data collection and data analyses. A purposive sample of 72 pre-service science teachers (PSTs) who were enrolled in a public university located in East Black Sea region of Turkey constituted the sample of the study. Content Validity Ratio (CVR=.80) and Content Validity Index (CVI=.94) values were calculated as measures of content validity. Pearson Correlation Coefficient (r=.96) and Cohen's Kappa value (κ value was between .60 and 1.00) were calculated to test inter-rater reliability of the scores obtained by the rating scale. Findings provided evidence for the reliability and the validity of the ABI rating scale. ABI lesson plan template and ABI rating scale developed for the assessment and evaluation of ABI lesson plans and subsequent teaching practices are provided to the readers. Contributions to the field are discussed.

1. INTRODUCTION

Countries should focus on training qualified people to have a word in scientific and economic fields and capture future changes and developments that will occur in these fields (Stohlmann, Moore & Roehring, 2012; Şahin, Ayar & Adıgüzel, 2014; Tunkham, Donpudsa & Dornbundit, 2016; Turkish Industry and Business Association [TÜSİAD], 2017). From this point of view, it has become important to raise individuals who are responsible for their own learning and who can investigate and question various issues they are confronted with. Moreover, it has also become very important to educate citizens who can express their opinions on controversial contemporary issues and persuade others by presenting logical arguments instead of rejecting every other opinion/idea or directly accepting them as they are. The primary way to raise individuals who have the desired characteristics described above is to make necessary changes in education systems. In this context, cultivation of higher-level thinking skills, such as 21st

*CONTACT: Funda Yesildag-Hasancebi ✉ funda.hasancebi@giresun.edu.tr Giresun University, Faculty of Education, Department of Educational Sciences, Giresun, Türkiye.

e-ISSN: 2148-7456 /© **IJATE** 2022

century skills, is one of the emphasized educational goals that take place in educational reform documents (Leou, et. al., 2006).

Problem solving, critical thinking, reflective thinking, collaboration, and entrepreneurship are some of the skills included in 21st century skills (National Research Council [NRC], 2011) and they have a natural and strong connection with science education. For instance, Nature of Science (NOS) views and their development are proposed to be related to many of the 21st century skills (NGSS Lead States, 2013). Argumentation-based teaching practices are also recommended as an effective teaching approach to improve students' 21st century skills (Ecevit & Kaptan, 2021). Considering group work and small/large group active negotiation processes that include social interaction, argumentation improves communication skills, collaboration, critical thinking and decision-making skills which are listed among 21st century skills (Driver, Newton & Osborn, 2000; Ecevit & Kaptan, 2021; Kabataş Memiş, 2017; Nam, Choi & Hand, 2011; Sevgi & Şahin, 2017; Yeşildağ-Hasancebi & Günel, 2014). Based on these, this study focused on the development of a rating scale that may be used for the assessment and evaluation of argumentation-based inquiry (ABI) lessons. Details of the ABI teaching approach and its utilization in science education and the necessity of developing a rating scale that is based on ABI teaching approach are given in the following sections.

1.1. Theoretical Framework

This study is theoretically grounded by argumentation-based inquiry (ABI), which is based on Science Writing Heuristic (SWH) approach. The SWH approach is proposed as a way to help students gain deeper understanding about the big ideas of science by planning, constructing and testing questions, justifying their claims with the evidences they have gathered, making comparisons with others' ideas, and elaborating on the changes in their ideas through the process they went through (Akkuş, Günel & Hand, 2007). Accordingly, SWH template for teacher and student (Choi, Hand & Greenbowe, 2013; Hand, Wallace & Yang, 2004; Nam, Choi & Hand, 2011) and researchers' ABI application experiences were utilized while constructing the items of the ABI rating scale.

Argumentation is the process of constructing data and claims, and their justifications, by making experimental and theoretical connections (Erduran & Jimenez-Aleixandre, 2007). Osborne (2005) defined argumentation as the way of predicting, evaluating, and proving evidences and operating mechanisms of reasoning on the opposite/contradictory arguments in the process of knowledge construction. Argument, on the other hand, is a form of discourse that needs to be taught explicitly through appropriate teaching activities, support, and modeling (Simon, Erduran & Osborne, 2006). As stated by Toulmin, an argument consists of basic components of claims, data, warrants, qualifiers, backings, and rebuttals (Toulmin, 1958). With the help of the utilization of these components, an argument includes the ability to put forward reasons for an event or situation and to test the causes of the event/situation with appropriate evidences from different viewpoints (Driver et al., 2000).

As an instructional approach that is designed to support students' science learning, ABI applications aim to foster science discourses among students (Hand & Norten-Meier, 2011) and supports creation of sound arguments (especially in written forms) in a scientific inquiry (Cavagnetto, Hand & Norten-Meier, 2010; Choi et al., 2010). By this way, ABI helps students construct scientific knowledge through scientific inquiry (Cavagnetto et al., 2010; Hand & Keys, 1999). ABI approach also helps students to personally experience the argumentation processes that scientists go through while constructing a scientific theory or law (Burke, Greenbowe & Hand, 2006) and, thus, enables students to better understand scientific explanations and related theories and laws (Erduran, Simon & Osborne, 2004).

In ABI approach, where thinking and writing activities are at the forefront, students ask questions, test their evidences, make claims based on their findings, and make decisions after

comparing their claims with the already existing scientific knowledge (Hand, 2008; Hand, Wallace & Yang, 2004; Martin & Hand, 2007). In this process, students organize their own research questions, create strategies/methods (e.g., making observations, doing experiments etc.) to answer them, analyze and interpret their findings, and share their claims (together with their evidences) with others (Hand et al., 2004; Martin & Hand, 2007). Small group discussions made with group mates and classroom discussions made with all of the students in the classroom are among the important elements of the ABI approach. During these processes students have the chance of experiencing testing and meaning making of their own knowledge about the issues (Burke et al., 2006). At this point, teacher guidance plays a vital role in the realization of these processes, and thus, efficiency of the application of the ABI approach.

1.1.1. *Argumentation-Based Inquiry Approach in Science Education*

Inquiry based teaching strategies are adopted in many science curricula all around world (e.g., Australian Curriculum, Assessment and Reporting Authority [ACARA], 2012; Ministry of National Education, Turkey [MoNE], 2018; National Research Council [NRC], 2000; NGSS Lead States, 2013). Contemporary science education curriculum standards make explicit reference to "Science is based on empirical evidence" (Guilfoyle, Erduran & Park, 2021; National Science Teaching Association [NSTA], 2020). In these curricula, it is highlighted that the inquiry processes should include more than making experiments but should foster students' skills in making explanations and generating arguments about their findings as well as the processes they went through while conducting their experiments (MoNE, 2018; NGSS, 2013). Relationships between argumentation and scientific literacy are also highlighted by Simon, Erduran, and Osborne (2006) who propose the ability to understand and follow scientific arguments as essential aspects of scientific literacy.

In addition to promoting scientific literacy, using argumentation in science education is reported to have many other benefits such as supporting cognitive development of students, creating opportunities for their critical thinking, and encouraging students for utilizing scientific language. These processes, in turn, are proposed to contribute to the development of students' social skills (e.g., communication skills), enable them to acquire a sense of culture of science, and develop more sophisticated epistemological beliefs (Jimenez-Aleixandre & Erduran, 2008). Moreover, argumentation approach has been found to improve students' conceptual understanding and play an important role in their science learning that is centered on thinking and reasoning processes (Chin & Osborne, 2010). In addition to promoting in-depth learning, argumentation processes make students curious and active, encourage them to create explanations, and provide opportunities for students and teachers to examine and solve errors that may be faced during learning of science (Kaya & Kılıç, 2008). Enabling students to approach events and issues from different perspectives and developing their creativity and imagination are also among the outcomes observed as a result of utilizing argumentation in educational settings (Aktamış & Atmaca, 2016; Gencel & Ilıman, 2019). Necessity of reflecting on evidences, identifying contradictory claims, imagining alternatives, and approaching issues and situations from different perspectives can be given as the main features of argumentation that result in the above-mentioned educational outcomes (Bean, 1996; Chen & She, 2012; King, 2000).

Based on the critical role that teachers play in the effectiveness of argumentation-based learning environments, many researchers emphasize the need for teachers who are well-equipped in this field (Sampson & Blanchard, 2012; Yıldırır & Nakiboğlu, 2014). The importance of teacher pedagogy for achieving desired learning outcomes has also been put forth in a number of research studies (Akkuş et al., 2007; Martin & Hand, 2007). More specifically, in order to efficiently utilize argumentation in science classes, teachers must have the necessary skills to perform evidence-based argumentation activities and be prepared for the difficulties they may

face during their implementation (Yıldırır & Nakiboğlu, 2014; Zohar, 2008). Teachers' level of knowledge about argumentation is also among the factors that are found to be influential on their classroom practices (Sampson & Blanchard, 2012; Simon et al., 2006). Therefore, it is important to improve teachers' pedagogical competencies and knowledge levels about argumentation strategies since teachers have vital roles in the implementation of educational reforms (Çepni & Çil, 2016).

Research shows that teachers do not have sufficient resources and pedagogical competencies for implementing argumentation in science classes (Sampson & Blanchard, 2012; Simon et al., 2006). Moreover, teachers frequently state that argumentation activities are time-consuming (Aktamış & Atmaca, 2016; Simon & Johnson, 2008; Torun & Şahin, 2016) and lesson hours are not sufficient for integrating argumentation in their teaching (Gencel & Ilıman, 2019; Namdar & Tuşkan, 2018). In addition to inexperience in using argumentation in their teaching, teachers' pedagogical insufficiencies and inabilities for making efficient planning for argumentation-based lessons may be regarded as the main reasons of these time-related concerns (Namdar & Tuskan, 2018). In this respect, teachers are suggested to use effective time management strategies and detailed planning in order to overcome many of the problems that may be faced during the implementation of argumentation in their lessons (Gencel & Ilıman, 2019). In line with these suggestions, in the present study it was aimed to develop a rating scale that can be used to guide teachers and teacher candidates in the preparation and implementation of argumentation-based lessons and evaluation of their efficiency in using argumentation strategies in their teaching, respectively.

Review of literature reveals that there is limited number of studies conducted on teaching of argumentation and most of the studies are focused on examining classroom practices of teachers after their participation in teacher training courses (Erduran, Ardac & Yakmacı-Güzel, 2006; Namdar & Tuşkan, 2018; Simon et al., 2006). Some of the studies are about the relationships between patterns of questioning and argumentation (Günel, Kıngır & Geban, 2012), efficiency of argumentation strategy for improving science teachers' self-efficacy perceptions toward technological pedagogical content knowledge (Çoban et al., 2016), and views of science teachers with different teaching experiences about scientific argumentation (Namdar & Tuşkan, 2018). As a common conclusion, researchers state that there is need for improving teachers' and teacher candidates' perceptions about and skills in using argumentation in their teaching (Aydeniz & Özdilek, 2016; Namdar & Tuşkan, 2018). Teachers should provide their students with appropriate discussion environments so that students can form valid arguments and support their arguments with variety of evidences (Cirit Gül, Apaydın & Çobanoğlu, 2021). In order to be able to integrate argumentation process into their teaching it is important for teachers to understand what they need to know in this process (McNeill et al., 2017). Therefore, it is necessary for the teacher to understand what argumentation is and how to carry out this argumentation process (Chan, Fancourt & Guilfoyle, 2020). In the literature, studies on teachers' learning and teaching of argumentation generally focus on science education (Chan & Erduran, 2022). In one of these research, İsbir and Yıldız (2021) examined limitations and difficulties faced by teachers during implementation of argumentation. The researchers grouped these limitations as limitations arising from (i) teacher, (ii) student, (iii) working with the group, (iv) educational environment, (v) method and the curriculum.

1.2. Purpose and Significance of the Study

In the present study it was mainly aimed to develop a reliable and valid rating scale for the use of the assessment and evaluation of lesson plans and subsequent teaching practices that are based on argumentation-based inquiry (ABI). The significance of this rating scale development study was (i) evaluating teachers'/teacher candidates' ABI lesson plans and subsequent teaching practices with a validated instrument, (ii) providing detailed feedback aligned to

certain criteria to teachers/teacher candidates regarding every stage of their ABI lesson plans and/or subsequent teaching practices, (iii) providing guidance on teaching of the ABI instructional model and supporting teachers'/teacher candidates' skills in designing ABI lessons in pre-service and in-service teacher training programs, and (iv) enabling teachers/teacher candidates to self-evaluate their ABI teaching with a validated instrument.

2. METHOD

2.1. Research Design

This research is a rating scale development study. Research design of the study is exploratory design. Exploratory design is a type of mixed-methods research that is especially useful in developing and testing instruments (Fraenkel, Wallen, & Hyun, 2012). In this study, as typically realized in exploratory design, application of the qualitative phase of the study was followed by quantitative analyses, which were used to validate quantitative findings. Preparation of the rating scale's draft form and taking expert opinions for confirming its validity constituted the qualitative dimension of the study; whereas, determination of the harmony among expert opinions and statistical analyses applied for testing reliability and validity of the rating scale required quantitative methods (McGartland et al., 2003).

2.2. Participants

Participants of the study were 72 pre-service science teachers (PSTs) who were enrolled in a public university located in East Black Sea region of Turkey. Criterion sampling method was used for sample selection. This allowed for making in-depth analyses with a group of participants who meet certain criteria of interest (Büyüköztürk et al., 2020). Experience with a phenomenon of interest is an important criterion for selecting participants with this method (Cresswell & Plano Clark, 2011). In accordance, PSTs who participated in the present study were selected among the ones who were experienced with the ABI approach. That is, the participants had taken courses in the university which were designed through ABI approach offered by the researchers of the study who have sufficient theoretical and practical expertise in ABI. 40 PSTs participated in the first year of the study for piloting the ABI rating scale. Data collected from these participants was not used in data analyses.

2.3. Context of the Study and Development of the Rating Scale

Rubrics are defined as criterion-based scoring tools which are developed by following theoretical processes and opinions of small samples of experts (Yurdagül, 2005). Accordingly, findings of previous research and expertise of researchers (including researchers of the present study) were utilized for the development of the ABI rating scale. In line with Goodrich Andrade (1997, 2001), Mertler (2001), and Kan's (2007) suggestions, the following stages were followed for the development of the ABI rating scale:

1) Review of the rating scale development and ABI literature
2) Determination of the criteria, definitions and scoring level to be used in the rating scale
3) Preparation of the draft version of the rating scale (see First Year of the Study section for detailed information)
4) Taking expert opinions (see Validity section for detailed information)
5) Application of the draft version of the rating scale (see Second Year of the Study for detailed information)
6) Determination of the reliability and validity values of the rating scale (see Reliability and Validity sections for detailed information)
7) Finalizing the rating scale

Development process of the ABI rating scale took place in "Science Teaching and Laboratory Applications" course (four hours a week) which was offered for 3rd grade PSTs. The PSTs who took the course were expected to realize laboratory experiments and activities on physics, chemistry, and biology subjects through Argumentation-Based Inquiry teaching approach. The study included two academic years (four academic semesters). The first year (Fall and Spring semesters) and the second year (Fall and Spring semesters) (see Figure 1).

Figure 1. *Procedures followed in the first and second year of the study.*

*Note: PSTs (Pre-service science teachers) who participated in the second year of the study were different from the ones who participated in the first year.

2.3.1. *First year of the study*

Before the beginning of the first semester, three researchers determined the sections and contents to be included in the ABI lesson plan template. In addition to previous research (Choi, Hand, & Greenbowe, 2013; Hand, Wallace, & Yang, 2004), personal experiences of the researchers in planning and applying ABI lessons were utilized for this phase. Researchers of the study are experienced in planning and implementing ABI lessons in primary school, secondary school, and university level science classes. Moreover, they have conducted teacher training programs for the development of teachers' skills in implementing argumentation-based science lessons.

The sections and contents expected to be given place in each section of the ABI lesson plans were submitted to two experts (one university professor and one science teacher) who implement ABI lessons in their courses. The first draft of the lesson plan was prepared in the light of the received feedback from these experts. Then, in line with this lesson plan draft, rating scale to be used for the assessment and evaluation of lesson plans were prepared.

Sections of the lesson plan (Appendix 3 and 4) are in the following: (i) Pre-lesson preparation: constructing concept map of the unit, determination of the big idea and the sub-ideas, (ii) Discussion on the research question to be investigated (planning of the introductory activity), (iii) Procedures followed during experiments/observations/research by students (investigation of research questions; formation of claims and evidences) (iv) Procedures followed during argumentation of students' claims and evidences, (v) Procedures followed during comparison of students' findings with the literature, (vi) Providing opportunities to reflect on the change of the ideas, (vii) Linking the lesson with Nature of Science and Nature of Scientific Inquiry throughout the lesson, (viii)Linking the lesson with the subsequent lesson, (ix)Additional lesson plan components.

In the first semester of the "Science Teaching and Laboratory Applications" course researchers planned and implemented 20 ABI science lessons (two implementations for each of the 10 weeks). During these 10 weeks the PSTs had student roles and worked in groups of 4-5 to follow the directions given by the instructors (researchers of the study). By this way, PSTs had the opportunity to learn and experience ABI approach and its implementation in science lessons. At the beginning of the second semester of the course, the researchers presented one of the lesson plans they implemented in the previous semester (as an example) to the same PSTs to explain how they prepared ABI lesson plans and what they paid attention to while preparing and implementing the lesson plans. Questions of the PSTs about the lesson plans and their implementation were answered and necessary explanations were given in detail. Then, the PSTs were asked to form groups of two (a total of 20 groups was formed). For the rest of the semester (Fall semester of the first year), the PSTs in these groups were asked to prepare and implement two ABI lesson plans for two science subjects they selected. The PSTs who implemented their lesson plans had the roles of teachers and the rest of the class (including the researchers) had the roles of students during this process. The main purpose of this process was to develop PSTs' skills in preparation of ABI lesson plans and implementing the lesson plans in classroom environment in accordance with their plan.

Giving feedback was a very crucial element of this process (preparation of lesson plans and implementing them in the classroom environment). In order give feedback in the fastest and the most efficient way, an e-mail address was created for the course. PSTs sent their lesson plans one week prior to their implementation and took feedback by all of the three researchers before their classroom implementations. The researchers utilized Google Drive in order to be able to give joint feedback to the lesson plans. In addition, before each course day the researchers and the PSTs who would implement their lesson plans met face to face to discuss details of the lesson plan applications.

In addition to its use as a tool for evaluating the performance of the PSTs who implement their lesson plans (data collected from these participants was not used in data analysis), ABI rating scale and the item statements in it were subjected to a continuous evaluation in terms of their clarity, usability, measurability etc. After each classroom session the researchers discussed their evaluations in terms of the PSTs' performance and ABI rating scale's ability to evaluate those performance.

ABI rating scale was also used as a self-evaluation tool by the PSTs to evaluate their performance in planning and implementing ABI lessons. PSTs individually completed ABI rating scale and submitted it to the researchers after their ABI lesson plan implementations. In addition, classroom discussions were done after each lesson plan implementation where PSTs and the researchers discussed their ideas about the PSTs' performance and the rating scale (necessity of use during the process, its shortcomings etc.). Notes taken during these discussions and after-class discussions made among the researchers were utilized in the revision of the ABI lesson plan and ABI rating scale after two academic semesters. The first year of the research especially includes the determination and clarification of the criteria in the rating scale.

2.3.2. *Second year of the study*

This phase includes application of the ABI rating scale and processes realized for testing its reliability and validity. Issues related to the rating scale's validity (taking expert opinions, revisions done based on the taken expert opinions, calculation of Content Validity Ratio (CVR) and Content Validity Index (CVI) etc.) and details of the rating scale's reliability analysis findings are presented under Findings section. Data used for the reliability analyses were collected from 72 PSTs other than the ones who participated in the first year. The opinions of 10 experts were taken for validity before applications.

Procedures followed in the second year of the study were similar to the ones followed in the first year. That is, in the first semester (Fall semester) of the "Science Teaching and Laboratory Applications" course the researchers planned and implemented 20 ABI lessons on various physics, biology, and chemistry topics. PSTs participated in the courses in groups of 4-5 and followed instructions given by the researchers. In these classroom sessions, the PSTs learned and gained experience in lesson plan implementations realized through ABI approach. At the beginning of the second semester (Spring semester) the PSTs were presented a sample lesson plan that they experienced in the previous semester in order to give details about the preparation of ABI lesson plans and applications in classroom environment. After clarifying PSTs' questions about the ABI approach and related issues (preparation of the lesson plans, issues that should be paid attention during implementation of the lesson plans, etc.) PSTs were asked to form groups (two PSTs in each group) that they will work together until the end of the semester. Each week groups acted as teachers and implemented their ABI lesson plans in the classroom environment. Rest of the class (including the researchers) had student roles in these implementation sessions. Similar to the first year of the study, joint feedback was given to the lesson plans of the PSTs by the three researchers (via e-mail and Google Drive application) one week prior to the classroom implementations. Moreover, face to face discussions were made among the researchers and the PSTs who would be implementing their lesson plans. Each group of PSTs planned and implemented two ABI lessons in total. These lesson plans and implementations were evaluated by the three researchers (during the ABI lesson plan implementations) and the PSTs (as self-evaluation realized after the ABI lesson plan implementations) by use of the ABI rating scale. Researchers' evaluations were used for reliability analyses. See Appendix 1 and 2 for the Turkish and English versions of the rating scale.

2.4. Reliability of the Rating Scale

Consistency of scores obtained by the use of a rating scale by different raters and in different occasions refers to the reliability for that rating scale (Kutlu, Doğan & Karakaya, 2010; Moskal & Leyden, 2000). In order to achieve reliability of the ABI rating scale researchers paid attention to some important facets suggested by colleagues with regard to the development and design of rating scale such as writing criteria to be assessed by the rating scale in a clear and understandable way, limiting content of each criteria assessed by the rating scale in a way that they were intensely focused on the purpose of the criteria, and writing descriptive explanations of the level (degree) definitions in a way that they correctly reflected the levels of the scoring used in the rating scale(Jonsson & Svingby, 2007; Moskal & Leydens, 2000). Finally, as suggested by Kutlu et al. (2010), in order to obtain a more reliable scoring, levels used in the rating scale was designed based on a 5-point scale (0 = *not acceptable*; 1 = *poor*; 2 = *average*; 3 = *good*; 4 = *very good*).

The reliability of the rating scale is expressed as the scoring does not change from one rater to another (Kutlu et al., 2009). Rater reliability is examined in two ways: intra-rater reliability and inter-rater reliability. Cronbach's Alpha coefficient is generally used to calculate intra-rater reliability (consistency of scores given by the same individual) (Jonsson & Svingby, 2007). Cohen's Kappa is often used to determine inter-rater reliability (concordance between scores of more than one rater) (Cohen, 1960). Cohen's Kappa was used to calculate the inter-rater reliability of the scores (consistency of the scores given by independent raters) obtained by the use of the ABI rating scale because there was more than one rater in this study. Cohen's Kappa values range from 0 to 1 where grater values correspond to higher levels of consistency (Kutlu et al., 2010). Cohen's Kappa values calculated for the ABI rating scale indicated that the rating scale has a good inter-rater reliability (see Table 2 under findings section of the article). In addition, Pearson Correlation Coefficient was calculated to determine inter-rater reliability among the two researchers' total scores.

2.5. Validity of the Rating Scale

In the present study, the researchers consulted expert opinion while developing the ABI rating scale and for analyzing its content validity. That is, at the beginning of the second year of the study (see Figure 1) three experts in Measurement and Evaluation in Education departments of three different universities provided their expertise while revising the ABI rating scale that was used in the first year of the study. Based on taken expert opinions, item statements in the rating scale were written in a clearer way and some items were divided into two so that each item statement measured only one aspect of the ABI lesson plan and its implementation. In addition, explanations in the brackets were removed from the item statements so that the rating scale became simpler and easier to follow by its users.

"Modified Lawshe Technique" (Ayre & Scally, 2014; Wilson et al., 2012), which is a revised version of Lawshe's (1975) (critical CVR) content validity measure, was used to ensure the rating scale's content validity. This technique includes (i) establishment of experts group, (ii) preparation of the draft version of the rating scale, (iii) taking expert opinions, (iv) calculation of content validity ratios (CVR=Content Validity Ratio) of the item statements, (v) Calculation of content validity index (CVI= Content Validity Index) of the rubric, (vi) Development of the final version of the rating scale based on CVR and CVI values.

The quality and number of experts are of great importance in obtaining objective results from the analyses carried out for determining content validity. According to Ayre and Scally (2014) and Lawshe (1975), this number should be between 5 and 40. Correspondingly, opinions of 10 experts were used for the content validity analyses of the study. Three of the experts were university professors in the Measurement and Evaluation in Education department and four of

the experts were university professors who had numerous studies in the subjects of argumentation and Nature of Science. The remaining three experts were science teachers (with at least a master's degree) who implement argumentation-based science activities in their classrooms. The experts were asked to rate each item statement in the rating scale based on a three-point scale (1 = Should be removed (item does not measure the targeted structure); 2= Should be revised (item is related to the targeted structure but it is unnecessary); 3= Proper (item measures the targeted structure).

2.6. Data Analysis

Content validity is a professional subjective judgment of experts about the degree of relevant construct in an assessment instrument (Yaghmaie, 2003). The judgments of experts (N=10) were taken to test the content validity of the rating scale. Ayre and Scally (2014) stated that critical value for the CVR should be 0.80 for 10 experts at α =.05 significance level. This means that items with a CVR value below .80 should be excluded from the rating scale. In addition, when the CVI value is greater than the CVR value, the content validity of the remaining items in the rating scale is considered statistically significant (Lawshe 1975; Öngöz, 2011; Yeşilyurt & Çapraz, 2018).

Kolmogorov-Smirnov Normality Test ($p>.05$) showed that total scores given by the raters were normally distributed. Since collected data (i.e., scores given by the researchers) had a normal distribution, Pearson Correlation Coefficient was calculated in addition to Cohen's Kappa value to test inter-rater reliability of the rating scale. SPSS 21 program was used in the analysis.

3. FINDINGS

3.1. Reliability Findings

Cohen's Kappa values (κ) were calculated to determine inter-rater reliability of the scores (consistency of the scores given by independent raters) obtained by the use of the ABI rating scale. Cohen's Kappa values (κ) range from 0 to 1 where grater values correspond to higher levels of consistency (Kılıç, 2015; Kutlu et al., 2010). According to Landis and Koch (1977), Cohen's Kappa values (κ) between .61 and .80 indicate good agreement and Cohen's Kappa values (κ) between .81 and 1.00 indicate very good agreement between raters. Therefore, as tabulated in Table 1, Cohen's Kappa values (κ) calculated for the ABI rating scale might be considered to be good or very good in all criteria. All of the values were statistically significant between .60 and .91 ($p<.01$).

Consistency among raters can also be determined by looking at the level of compliance on the total scores obtained from rating scale (Kutlu et al., 2010). Accordingly, as a second analysis conducted for testing reliability of the ABI rating scale, inter-rater reliability among researchers' total scores were calculated. Results showed that minimum inter-rater consistency value was r =.96 ($p<.05$), which provided additional evidence for the reliability of the ABI rating scale.

Table 1. *Cohen's Kappa values for the item statements in the ABI Rating scale.*

Criteria	Item	κ	p
Pre-lesson preparation	1	.91	.00
	2	.87	.00
	3	.61	.00
	4	.83	.00
Discussion on the research question to be investigated	1	.78	.00
	2	.70	.00
	3	.60	.00
	4	.65	.00
	5	.64	.00
	6	.62	.00
	7	.81	.00
Testing/investigation of research questions	1	.76	.00
	2	.76	.00
	3	.60	.00
Claims and evidences	1	.84	.00
	2	.75	.00
	3	.69	.00
Discussion on the claims and evidences	1	.83	.00
	2	.60	.00
	3	.60	.00
	4	.68	.00
Comparison of the findings/observations with the existing literature	1	.84	.00
	2	.80	.00
Providing opportunities to reflect on the change of the ideas	1	.67	.00
	2	.72	.00
	3	.81	.00
Linking the lesson with Nature of Science and Nature of Scientific Inquiry	1	.81	.00
	2	.77	.00
Linking the lesson with the subsequent lesson	1	.76	.00
Additional lesson plan components	1	.80	.00
	2	.77	.00
Overall Evaluation	1	.70	.00
	2	.82	.00

Note. κ: Cohen's Kappa, N=72

3.2. Validity of the Rating Scale

Content Validity Ratio (CVR) and Content Validity Index (CVI) values were calculated as measures of the content validity of the rating scale. As seen in Table 2, CVR values of each item in the rating scale are above .80 as suggested by Ayre and Scally (2014). In addition, the CVI value belonging to the whole rating scale was determined as.94 (CVI values belonging to the sub-dimensions of the rating scale are also presented in Table 2). Since the CVI value (.94) is greater than the CVR (.80) value (i.e., CVI > CVR), content validity of the remaining items in the rating scale is accepted to be statistically significant.

Table 2. *CVR and CVI values for the item statement in the ABI Rating scale.*

Criteria	Item Number	Necessary	Unnecessary/ Unsatisfactory	Should be removed from the rating scale	CVR	CVI
Pre-lesson preparation	1	10	0	0	1.00	.95
	2	10	0	0	1.00	
	3	10	0	0	1.00	
	4	9	1	0	.80	
Discussion on the research question to be investigated	1	10	0	0	1.00	.94
	2	9	1		.80	
	3	10	0	0	1.00	
	4	10	0	0	1.00	
	5	10	0	0	1.00	
	6	10	0	0	1.00	
	7	9	0	1	.80	
Testing/investigating research questions	1	10	0	0	1.00	.93
	2	9	1	0	.80	
	3	10	0	0	1.00	
Claims and evidences	1	10	0	0	1.00	1
	2	10	0	0	1.00	
	3	10	0	0	1.00	
Argumentation on the claims and evidences	1	10	0	0	1.00	1
	2	10	0	0	1.00	
	3	10	0	0	1.00	
	4	10	0	0	1.00	
Comparison of the findings/observations with the existing literature	1	9	1	0	.80	.90
	2	10	0	0	1.00	
Providing opportunities to reflect on the change of the ideas	1	10	0	0	1.00	.93
	2	10	0	0	1.00	
	3	9	0	0	.80	
Linking the lesson with Nature of Science and Nature of Scientific Inquiry	1	10	0	0	1.00	1
	2	10	0	0	1.00	
Linking the lesson with the subsequent lesson	1	10	0	0	1.00	1
Additional lesson plan components	1	9	1	0	.80	90
	2	10	0	0	1.00	
Overall Evaluation	1	10	0	0	1.00	.90
	2	9	1	0	.80	
Strengths and weaknesses of the ABI implementation	1	10	0	0	1.00	.1
Total score	Total	9	0	1	.80	.80

Note. Number of experts = 10. Content Validity Ratio (CVR) =.80; Content Validity Index (CVI) =.94.

4. DISCUSSION and CONCLUSION

In this study, it was mainly aimed to develop a reliable and valid rating scale for the use of the assessment and evaluation of lesson plans and subsequent teaching practices that are based on argumentation-based inquiry. Rating scales have some benefits in guiding teachers and students in the teaching and learning processes. For example, rating scales show students the learning goals of the lessons in a clear way, guide them in getting prepared for their studies and provide them with feedback through self-assessment and peer assessment (Frazel, 2010; Wolf & Steven, 2007). In addition, rating scales guide teachers in the assessment and evaluation processes and serve for making assessment and evaluation of the learning outcomes more accurate and fairer. Therefore, the ABI rating scale developed throughout the present study is not planned just a scoring tool but as a guide for teachers, teacher candidates and teacher educators who want to practice argumentation in their teaching.

CVR values of each item in the developed rating scale was calculated to be significant and larger than .80 (there was only one item with a CVR value below .80 and this item was removed from the rating scale with the consensus of expert opinions). Threshold CVR value was determined to be .80 since opinions of 10 experts were used for the validity analyses (Ayre & Scally, 2014). CVI of the rating scale was large (.94) and greater than the CVR value, indicating significance of the content validity of the rating scale (Lawshe, 1975; Öngöz, 2011; Yeşilyurt & Çapraz, 2018).

Kutlu et al., (2010) states that rating scale is reveal the differences among the graded/scored individuals and result in more reliable if grading is realized on a 4 to 7-point scale. Based on this, the ABI rating scale developed throughout the present study was designed on a 5-point scale (0 = *not acceptable*; 1 = *poor*; 2 =*average*; 3 = *good*; 4 = *very good*). Findings of the reliability analyses calculated for each item of the rating scale ($\kappa_{min.}$ = .60) indicated that consistency among the raters ranged from "good" to "very good" (Landis & Koch, 1977; Şencan, 2005). Moreover, total scores given by the raters by use of the ABI rating scale were found to be highly correlated providing additional evidence for the reliability of instrument.

ABI rating scale consists of two parts. The first part includes 33 items which allows raters to make quantitative evaluations regarding the appropriateness of the lesson plans and lesson plan implementations for argumentation-based inquiry teaching (ABI) approach. These 33 items are grouped into 11 sections (e.g., pre-lesson preparation, discussion on the research question to be investigated, testing/investigation of research questions, etc.; see Table 1 and Table 2 for a full list of 11 sections and their validity and reliability values). At the beginning of the rating scale, raters are presented with criteria for scores (scoring criteria section) that will be used for evaluating ABI lesson plans and lesson plan implementations. The second part of the rating scale includes a general evaluation where raters can write their views about the strengths and weaknesses of the enacted ABI lessons.

Each section of the ABI rating scale corresponds to an important step for the argumentation process. For instance, pre-lesson preparation section includes the processes that are critical for the teacher to do before practicing of the planned lesson, such as determining the objectives targeted in the application, creating a concept map of the unit to be taught, and determining the big idea and sub-ideas of the unit. At this point, creating his/her own concept map about the unit will make it easier for the teacher/teacher candidate to be able to evaluate the sufficiency of his/her knowledge about the subject area, focus on the purpose of the subject to be taught together with connections of the subject related concepts with each other, and determine the big idea and sub-ideas of the lesson. The big idea can be described as the point that we want our students to reach in accordance with the objectives of the unit plan, and sub-ideas are the main themes of each argumentation activity implemented throughout the lesson and act as paths to reach the big idea of the unit (Yeşildağ-Hasancebi & Akbay, 2017). Accordingly, determination

of the big idea and the sub-ideas of the lesson provides a basis for the subsequent steps of the lesson and ensures that argumentation processes are carried out in a better way (e.g., keeping the focus of the argumentation on the related subject).

As another example, claims and evidences and discussion on the claims and evidences sections of the ABI rating scale are essential steps for constructing reasoning components of the argumentation process. Reasoning components of the argumentation basically include students' justifications about how their evidences support their claims (Berland & McNeill, 2010; Sandoval & Millwood, 2005). Moreover, argumentation includes reasoning about whether information at hand is scientific or not (Arık & Akçay, 2017). Findings of research show that students generally have difficulties in presenting skills in the reasoning components of the argumentation process such as developing qualified arguments in their argumentation-based lessons (Aydeniz & Bilican, 2016; Bell & Linn, 2000; McNeill et al., 2006). Therefore, providing guidance in structuring claims and evidences based on the data gathered for research questions, establishing question-claim-evidence relationships, forming convincing arguments, and creating supporting or counter arguments in response to a presented argument is very crucial for the sake of the desired outputs (e.g., skills in developing qualified arguments) of the argumentation process.

In order for teachers to integrate the argumentation process into their own teaching, it is necessary to understand what they need in this process (McNeill et al., 2017). In addition, teachers' own science learning experiences are mostly limited to textbooks or curricula determined by exams and they do not know how to argue due to lack of experience in engaging and maintaining scientific discussion (Sampson & Blanchard, 2012; Zembal-Saul & Vaishampayan, 2019; Zohar, 2007). Therefore, keeping in mind that argumentation requires knowledge and experience (Türkmenoğlu & Çopur, 2021), teachers may need a guide to create and continue argumentation processes in the classroom environments.

ABI rating scale developed in the present study includes all these essential steps and thus might be used as an effective tool for guiding teachers, teacher candidates, and students in the implementation of argumentation in their lessons. The rating scale provides a roadmap that its users may use as a base for their ABI lessons by focusing on what is expected in an ABI lesson and what they should focus on during the planning and implementation of their ABI lessons. The rating scale might also be used as an evaluation tool for evaluation of the ABI lessons. Moreover, teachers and teacher candidates can benefit from the ABI rating scale to self-evaluate themselves and develop their skills in the planning and implementation of ABI lessons.

4.1. Suggestions for Further Research

ABI rating scale developed throughout the present study was shown to be a reliable and valid instrument to be used in the evaluation of ABI science lesson plans and subsequent implementations. Nonetheless, findings of further research carried out with diverse samples will add to improving its reliability and validity. Use of the ABI rating scale with science teachers will provide additional data for testing the efficiency of its use in ABI science lesson plans and implementations. Similarly, literature will benefit from further research that utilize the developed ABI rating scale in other disciplines such as social studies courses which can benefit from argumentation approach in their implementations in schools (Torun & Şahin, 2016). Findings of research carried out with diverse samples (i.e., teachers and teacher candidates from different school disciplines) will provide evidences regarding the generalizability of the present study's findings and efficiency of the use of the ABI rating scale in scholarships other than science education. Finally, more detailed information about the efficiency of the use of the ABI rating scale and its potential contributions for the teachers and teacher candidates can be gathered though the use of qualitative research methods. For instance, interviews can be conducted with teachers/teacher candidates who use the rating scale in their

lessons in order to collect data on their views about the efficiency of the use of the ABI rating scale and suggestions for its further development.

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**: Giresun University/ Scientific Research and Publication Ethics Committee, 2018-18/2.

Authorship Contribution Statement

Authors are expected to present author contributions statement to their manuscript such as; **Funda Yesildag Hasancebi**: Investigation, resources, determination of the methodology of the study, data collection, data analyses, writing-original draft of the manuscript, revision of the manuscript. **Busra Tuncay Yuksel:** Determination of the methodology of the study, data collection, writing original draft of the manuscript, revision of the manuscript. **Gunkut Mesci**: Determination of the methodology of the study, data collection, supervision and control of the original draft of the manuscript, revision of the manuscript.

Orcid

Funda Yesildag Hasancebi https://orcid.org/0000-0001-9365-940X
Busra Tuncay Yuksel https://orcid.org/0000-0002-4129-7256
Gunkut Mesci http://orcid.org/0000-0003-0319-5993

REFERENCES

Akkus, R., Günel, M., & Hand, B. (2007). Comparing an inquiry-based approach known as the science writing heuristic to traditional science teaching practices: Are there differences? *International Journal of Science Education, 14*(1), 1745-1765. https://doi.org/10.1080/09500690601075629

Aktamış, H., & Atmaca, A.C. (2016). Fen bilgisi öğretmen adaylarının argümantasyon tabanlı öğrenme yaklaşımına yönelik görüşleri [View's of pre service science teachers about argumentation based learning approach]. *Electronic Journal of Social Sciences*, *15*(58), 936–947. http://doi.org/10.17755/esosder.48760

Arık, M., & Akçay, B. (2017). Argümantasyon tabanlı öğrenme [Argumantation based learning]. In B. Akcay (Ed.), *Fen bilimleri eğitimi alanındaki öğrenme ve öğretme yaklaşımları [Learning and teaching approaches in science education] (pp.177 192). Pegem A Yayıncılık.*

Australian Curriculum, Assessment and Reporting Authority. (2012). *The Australian curriculum: Science* (Version 3.0). Commonwealth of Australia, NSW.

Aydeniz, M., & Ozdilek, Z. (2016). Assessing and enhancing pre service science teachers' self efficacy to teach science through argumentation: Challenges and possible solutions. *International Journal of Science and Mathematics Education*, *14*(7), 1255-1273. http://doi.org/10.1007/s10763-015-9649-y

Ayre, C., & Scally, A.J. (2014). Critical values for Lawshe's content validity ratio: Revisiting the original methods of calculation. *Measurement and Evaluation in Counseling and Development*, *47*(1), 79-86. https://doi.org/10.1177/0748175613513808

Bean, J.C. (1996). *Engaging ideas: The professor's guide to integrating writing, critical thinking, and active learning in the classroom*. Jossey-Bass Press.

Bell, P., & Linn, M.C. (2000). Scientific arguments as learning artifacts: Designing for learning from the web with KIE. *International Journal of Science Education*, *22*(8), 797 817. https://doi.org/10.1080/095006900412284

Berland, L.K., & McNeill, K.L. (2010). A learning progression for scientific argumentation: Understanding student work and designing supportive instructional contexts. *Science Education, 94*(5), 765 793. https://doi.org/10.1002/sce.20402

Burke, K.A., Greenbowe, T.J., & Hand, B.M. (2006). Implementing the science writing heuristic in the chemistry laboratory. *Journal of Chemical Education, 83*(7), 1032 1038. https://doi.org/10.1021/ed083p1032

Buyukozturk, Ş., Kılıc-Cakmak, E., Akgun, Ö.E., Karadeniz, S. & Demirel, F. (2020). *Bilimsel araştırma yöntemleri [Scientific research methods].* (29th ed.). Pegem Yayıncılık.

Cavagnetto, A.R., Hand, B., & Norten Meier, L. (2010). Negotiating the inquiry question: A comparison of whole class and small group strategies in grade five science classrooms. *Research in Science Education, 41*(2), 193-209. https://doi.org/10.1007/s11165-009-9152-y

Chan, J., Fancourt, N., & Guilfoyle, L. (2021). Argumentation in religious education in England: An analysis of locally agreed syllabuses. *British Journal of Religious Education, 43*(4), 458-471. https://doi.org/10.1080/01416200.2020.1734916

Chan, J., & Erduran, S. (2022). The impact of collaboration between science and religious education teachers on their understanding and views of argumentation. *Research in Science Education*, 1-17. https://doi.org/10.1007/s11165-022-10041-1

Chen, C.H., & She, C. (2012). The impact of recurrent online synchronous scientific argumen tation on students' argumentation and conceptual change. *Educational Technology & Society, 15*(1), 197-210. https://www.jstor.org/stable/jeductechsoci.15.1.197

Chin, C., & Osborne, J. (2010). Students' questions and discursive ınteraction: their ımpact on argumentation during collaborative group discussions in science. *Journal of Research in Science Teaching*, 47(7), 883-908. https://doi.org/10.1002/tea.20385

Choi, A., Notebaert, A., Diaz, J., & Hand, B. (2010). Examining arguments generated by year 5, 7, and 10 students in science classrooms. *Research in Science Education, 40*(2), 149-169. https://doi.org/10.1007/s11165-008-9105-x

Cirit Gül, A., Apaydın, Z., & Çobanoğlu, E.O., (2021). Türkiye'de Argümantasyon ile ilgili yapılan lisansüstü tezlerin incelenmesi [Investigation of graduate thesis about argumentation in Turkey: Thematic Content Analysis]. *Ondokuz Mayıs University Journal of Faculty of Education, 40*(2), 591-628. https://doi.org/10.7822/omuefd.863712

Creswell, J.W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). SAGE Publications

Cresswell, J.W., & Plano Clark, V.L. (2011). *Designing and conducting mixed method research* (2nd ed.). Thousand Oaks.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Çepni, S., & Çil, E. (2016). *Fen bilimleri dersi öğretim programı (Tanıma, Planlama, Uygulama ve TEOG ile ilişkilendirme). İlkokul ve ortaokul öğretmen el kitabı* [*Science lesson instruction program (Familiarity, Planning, Application and Relating with TEOG). Primary and middle school teacher handbook*]. Pegem.

Çoban, G.Ü., Akpınar, E., Baran, B., Sağlam, M.K., Özcan, E., & Kahyaoğlu, Y. (2016). The evaluation of "technological pedagogical content knowledge based argumentation practices" training for science teachers. *Education & Science/Egitim ve Bilim, 41*(188), 1-33. https://doi.org/10.15390/EB.2016.6615

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentati on in classrooms. Science Education, 84(3), 287-312. https://doi.org/10.1002/(SICI)1098-237X(200005)84:3<287::AID-SCE1>3.0.CO;2-A

Ecevit, T., & Kaptan, F. (2021). Describing the argument based inquiry teaching model designed for gaining the 21st century skills. *Hacettepe University Journal of Education, 36*(2), 470-488. https://doi.org/10.16986/HUJE.2019056328

Erduran, S., Ardaç, D., & Yakmacı-Guzel, B. (2006). Promoting argumentation in preservice teacher education in science. *Eurasia Journal of Mathematics, Science and Technology Education, 2*(2), 1-14.

Erduran, S., & Jimenez-Aleixandre, P. (2007). *Argumentation in science education: Perspecti ves from classroom-based research*. Springer Science.

Erduran, S., Simon, S., & Osborne, J. (2004). Tapping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science Education, 88*(6), 915-933. https://doi.org/10.1002/sce.20012

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

Frazel, M. (2010). *Digital storytelling guide for educators*. International Society for Technolo gy in Education (ISTE).

Gencel, İ.E., & Ilıman, M. (2019). Argümantasyona dayalı öğretime ilişkin bir durum çalışması [A case study on argumentation based teaching]. *International Journal of Curriculum and Instructional Studies, 9*(1), 53-72. https://doi.org/10.31704/ijocis.2019.003

Guilfoyle, L., Erduran, S. & Park, W. (2021) An investigation into secondary teachers' views of argumentation in science and religious education, *Journal of Beliefs & Values, 42*(2), 190-204. https://doi.org/10.1080/13617672.2020.1805925

Gunel, M., Kingir, S., & Geban, Ö. (2012). Analyses of argumentation and questioning patterns in argument-based inquiry classrooms. *Egitim ve Bilim, 37*(164), 316-330.

Hand, B. (2008). Introducing the science writing heuristic approach. In B. Hand (Ed.), *Science inquiry, argument and language: A case for the science writing heuristic.* Sense Publish ers.

Hand, B., & Keys, C. (1999). Inquiry investigation: A new approach to laboratory reports. *The Science Teacher, 66*(1), 27-29.

Hand, B., & Norton-Meier, L. (Eds.) (2011). *Voices from the classroom: Elementary teachers ' experience with argument-based inquiry*. Sense Publishers.

Hand, B., Wallace, C., & Yang, E. (2004). Using the science writing heuristic to enhance learning outcomes from laboratory activities in seventh grade science: Quantitative and qualitative aspects. *International Journal of Science Education, 26*(1), 131-149. https://doi.org/10.1080/0950069032000070252

İsbir, B., & Yıldız, A. (2021). Argümantasyon yönteminin uygulanması sürecinde karşılaşılan sınırlılıkların tartışılması [Discussing the limitations during the implementation of the argumentation method]. *Journal of Social Research and Behavioral Sciences, 7*(13), 23 6-258. https://doi.org/10.52096/jsrbs.6.1.7.13.13

Jimenez Aleixandre, M., & Erduran, S. (2008). *Argumentation in science education: An Overview*. In M Jimenez Aleixandre & S. Erduran (Eds.), *Argumentation in science education: Perspectives from classroom-based research* (pp. 3-27). Springer

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130-144. https://doi.org/ 10.1016/j.edurev.2007.05.002

Kabataş Memiş, E. (2017). Argümantasyon uygulamalarına katılan öğretmen adaylarının küçük grup tartışmalarına ilişkin görüşleri [Opinions of teacher candidate on small group discussions in argumentation applications]. *Kastamonu Education Journal, 25*(5), 2037-2056. Retrieved from https://dergipark.org.tr/en/pub/kefdergi/issue/31226/342940

Kaya, O.N, & Kılıç, Z. (2008). Etkin bir fen eğitim için tartışmacı söylev [Argumentative discourse for the effective teaching of science]. Ahi Evran University Journal of Kırşehir

Education Faculty, *9*(3), 89-100. https://dergipark.org.tr/en/pub/kefad/issue/59524/855999

King, P.M. (2000). Learning to make reflective judgments. In M.B. Baxter Magolda (Ed.), Linking student development, learning, and teaching: New directions for teaching and learning (pp. 15–26). Jossey-Bass.

Kutlu, Ö., Doğan, C.D., & Karakaya, İ. (2009). *Öğrenci başarısının belirlenmesi performansa ve portfolyoya dayalı durum belirleme [Determining student success, determining the situation based on performance and portfolio]* (2nd ed.). Pegem Akademi.

Kılıç, S. (2015). Kappa test. *Journal of Mood Disorders*, *5*(3), 142-144. https://doi.org/10.5455/jmood.20150920115439

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

Lawshe, C.H. (1975). A quantitative approach to content validity. *Personnel psychology*, *28*(4), 563-575.

Leou, M., Abder, P., Riordan, M., & Zoller, U. (2006). 'Using HOCS-centered learning' as a pathway to promote science teachers' metacognitive development. *Research in Science Education, 36*(1-2), 69-84. https://doi.org/10.1007/s11165-005-3916-9

Martin, A.M., & Hand, B. (2007). Factors affecting the implementation of argument in the elementary science classroom. A longitudinal case study. *Research in Science Education, 39*, 17-38. https://doi.org/10.1007/s11165-007-9072-7

McGartland, R.D., Berg Weger, M., Tebb, S., Lee, E.S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research, 27*(2), 94-104. https://doi.org/10.1093/swr/27.2.94

McNeill, K.L., Lizotte, D., Krajcik, J.S., & Marx, R.W. (2006). Fading scaffolds for argument ation and explanation. Journal of the Learning Sciences, 15(2), 153-191. https://doi.org/10.1207/s15327809jls1502_1

McNeill, K.L., Gonzalez Howard, M., Katsh Singer, R., & Loper, S. (2017). Moving beyond pseudoargumentation: Teachers' enactments of an educative science curriculum focused on argumentation. *Science Education, 101*(3), 426-457. https://doi.org/10.1002/sce.21274

Ministry of National Education (2018). *İlköğretim kurumları fen bilimleri dersi öğretim programı* [*Primary education institutions' science instruction program*]. Talim Terbiye Kurulu Başkanlığı. http://mufredat.meb.gov.tr/Dosyalar/201812312311937

Moskal, B.M., & Leydens, J.A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research, and Evaluation, 7*(1), 1-10. https://doi.org/10.7275/q7rm-gg74

Namdar, B., & Tuskan, İ.B. (2018). Fen bilgisi öğretmenlerinin argümantasyona yönelik görüşleri [Science teachers' views of scientific argumentation]. *Hacettepe University Journal of Education*, *33*(1), 1-22. https://doi.org/10.16986/HUJE.2017030137

Nam, J., Choi, A., & Hand, B. (2011). Implementation of the science writing heuristic (SWH) approach in 8th grade science classrooms. *International Journal of Science and Mathematics Education, 9*(5), 1111-1133. https://doi.org/10.1007/s10763-010-9250-3

National Research Council [NRC] (2000). *Inquiry and the national science education standards*. National Academic Press.

National Research Council. (2011). *Successful STEM education: A workshop summary*. National Academies Press.

National Science Teaching Association [NSTA] (2020). Nature of Science. https://www.nsta.org/nstas-official-positions/nature-science

NGSS Lead States (2013). *Next generation science standards: For states, by states*. National Academies. http://www.nextgenscience.org/next-generation-science-standards

Nam, J., Choi, A., & Hand, B. (2011). Implementation of the science writing heuristic (SWH) approach in 8th grade science classrooms. *International Journal of Science and Mathem atics Education, 9*(1), 1111-1133. https://doi.org/10.1007/s10763-010-9250-3

Osborne, J. (2005). The role of argument in science education. *Research and the Quality of Sc ience Education, 7*(1), 367-380. https://doi.org/10.1007/1-4020-3673-6_29

Öngöz, S. (2011). Elektronik ders kitabı değerlendirme formunun geliştirilmesi: Geçerlik ve güvenirlik çalışması [Development of electronic textbook evaluation form: Validity and reliability study*]. 11th International Educational Technology Conference, IETC*: Procee dings Book (Volume II), pp.1481-1485.

Sampson, V., & Blanchard, M.R. (2012). Science teachers and scientific argumentation: Trends in views and practice. *Journal of Research in Science Teaching, 49*(9), 1122 1148. http://doi.org/10.1002/tea.21037

Sandoval, W.A., & Millwood, K. (2005). The quality of students' use of evidence in written s cientific explanations. *Cognition and Instruction, 23*(1), 23-55. https://doi.org/10.1207/s 1532690xci2301_2

Sevgi, Y., & Şahin, F. (2017). The effects of discussion the socio-scientific subject in the new spaper based on argumentation 7th grades students' critical thinking. *Journal of Human Sciences, 14*(1), 156-170. https://j-humansciences.com/ojs/index.php/IJHS/article/view/ 4289

Sampson, V., & Blanchard, M. (2012). Science teachers and scientific argumentation: Trends in views and practice. *Journal of Research in Science Teaching, 49*(9), 1122-1148. https ://doi.org/10.1002/tea.21037

Simon, S., Erduran, S., & Osborne, J. (2006). Learning to teach argumentation: Research and development in the science classroom. *International Journal of Science Education, 28*(23), 235-260. http://doi.org/10.1080/09500690500336957

Simon, S., & Johnson, S. (2008). Professional learning portfolios for argumentation in school science. *International Journal of Science Education, 30*(5), 669-688. https://doi.org/10. 1080/09500690701854873

Şahin, A., Ayar, M.C., & Adiguzel, T. (2014). STEM related after school program activities and associated outcomes on student learning. *Educational Sciences: Theory and Practice, 14*(1), 309-322. https://doi.org/10.12738/estp.2014.1.1876

Şencan, H. (2005). *Sosyal ve davranışsal ölçmelerde güvenirlik ve geçerlilik [Reliability and validity in social and behavioral measures]*. SeckinYayıncılık.

Stohlmann, M., Moore, T.J., & Roehrig, G.H. (2012). Considerations for teaching integrated STEM education. *Journal of Pre-college Engineering Education Research (J-PEER), 2*(1), 28-34. https://doi.org/10.5703/1288284314653

Toulmin, S. (1958). *The uses of argument.* Cambridge University Press.

Torun, F., & Şahin, S. (2016). Determination of students' argument levels in argumentation-based social studies course. *Education and Science, 41*(186), 233-251. https://doi.org/10.15390/EB.2016.6322

Tunkham, P., Donpudsa, S., & Dornbundit, P. (2016). Development of STEM activities in chemistry on "protein" to enhance 21st century learning skills for senior high school students. *Silpakorn University Journal of Social Sciences, Humanities, and Arts, 16*(3), 217-234. https://doi.org/10.14456/sujsha.2016.17

Turkish Industry and Business Association [TÜSİAD] (2017). Faaliyet raporu [Activity report]. https://tusiad.org/tr/faaliyet-raporlari/item/9911-tusiad-faaliyet-raporu-2017

Türkmenoğlu, M., & Çopur, E. (2021). Sınıf öğretmenlerinin argümantasyona ilişkin görüşlerinin ve argüman oluşturma düzeylerinin incelenmesi. *Uluslararası Temel Eğitim Çalışmaları Dergisi, 2*(1), 29-42. https://dergipark.org.tr/en/pub/ijpes/issue/60113/7776 04

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*(1), 197 210. https://doi.org/10.1177/0748175612440286

Wolf, K., & Stevens, E. (2007). The role of rubrics in advancing and assessing student learning. *The Journal of Effective Teaching*, *7*(1), 3-14.

Yesildag Hasancebi, F., & Günel, M. (2014). Delving into the effect of argumentation based inquiry approach on learning science from multiple perspectives. *Journal of Research in Education and Society, 1*(1), 23-44.

Yeşilyurt, S., & Çapraz, C. (2018). A road map for the content validity used in scale development studies. *Erzincan Üniversitesi Eğitim Fakültesi Dergisi, 20*(1), 251 264. https://doi.org/10.17556/erziefd.297741

Yaghmaie, F. (2003). Content validity and its estimation. *Journal of Medical Education, 3*(1), 25-27.

Yıldırır, H.E., & Nakiboğlu, C. (2014). Kimya öğretmen ve öğretmen adaylarının derslerinde kullandıkları argümantasyon süreçlerinin incelenmesi [Examination of chemistry teachers and preservice teachers'argumentation processes used in their courses]. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 14*(2), 124-154. https://hdl.handle.net/20.500.12462/4228

Yurdugül, H. (2005). *Ölçek geliştirme çalışmalarında kapsam geçerliği için kapsam geçerlik indekslerinin kullanılması* [*Using content validity indexes for content validity in scale development studies*]. XIV. Ulusal Eğitim Bilimleri Kongresi [XIV. National Educational Sciences Congress], 1, 771-774.

Zembal Saul, C., & Vaishampayan, A. (2019). *Science teachers' continuous professional development inargumentation*. In S. Erduran (Ed.), Argumentation in chemistry education: Research, policy and practice (pp. 142 172). Royal Society of Chemistry.

Zohar, A. (2007). *Science teacher education and professional development in argumentation*. In S. Erduran & M. Jiménez-Aleixandre (Eds.), Argumentation in science education (pp. 245–268). Springer.

APPENDIX

Appendix 1. Argumentation-based Inquiry Rating Scale (English)

Name-Surname: Date:

Scores	Criteria for scoring
Very good (4)	All of the elements that make up the items in each stage are available with rich details and fully, appropriately and accurately planned and implemented. Another teacher can use this plan as it is.
Good (3)	More than half of the elements that make up the items in each stage have been fully, appropriately and accurately planned and partially implemented with rich details. Another teacher can use this criterion of the plan with minor changes.
Average (2)	Approximately half of the elements that make up the items in each stage are available with some details and are fully, appropriately and correctly planned (but not implemented). Other teachers can use this criterion of the plan with changes.
Poor (1)	Less than half of the elements that make up the items in each stage are available with some details and are fully, appropriately and correctly planned (but not implemented). Other teachers should re-plan this criterion of the lesson.
Not acceptable (0)	Basic elements of the lesson are not available (and are not implemented).

Criteria						**Explanations**
Pre-lesson preparation						
0	1	2	3	4	Concepts and/or skill to be covered in the lesson are comparable with the current science curriculum	
0	1	2	3	4	Lesson plan objective(s) are appropriate	
0	1	2	3	4	The big idea and the sub ideas are appropriate	
0	1	2	3	4	Concept map includes many concepts and relationships	
Discussion on the research question to be investigated (Planning of the introductory activity)						
0	1	2	3	4	Introductory activity reveals students' prior knowledge about the lesson objective(s)	
0	1	2	3	4	Introductory activity increases students' interests in learning	
0	1	2	3	4	Introductory activity provides opportunities for students to discuss and ask questions	
0	1	2	3	4	Introductory activity draws students' attention and leads them to questions they are curious about	
0	1	2	3	4	Introductory activity initiates and sustains discussion among students	
0	1	2	3	4	Research questions expected from the students are sufficiently specified in the lesson plan together with alternative strategies to be realized if students do not express expected research questions	
0	1	2	3	4	Necessary materials are completely specified and provided	
Testing/investigating research questions						
0	1	2	3	4	Students are guided to make experiments/research/observations appropriate with their research questions	
0	1	2	3	4	Activities planned for testing/investigating research questions are student-centered	

0	1	2	3	4	Important points to be considered during the testing/investigation of research questions are clearly specified with examples and applied accordingly	
Claims and evidences						
0	1	2	3	4	Planning and implementation of the lesson was clearly specified and sufficient enough to reveal how the teacher will enable students to construct claims and evidences based on data they obtained	
0	1	2	3	4	Planning and implementation of the lesson was clearly specified and sufficient enough to reveal how the teacher will enable students to establish the relationships among questions, claims, and evidences	
0	1	2	3	4	Planning and implementation of the lesson was clearly specified and sufficient enough to reveal how the teacher will enable students to develop persuasive arguments about their research questions	
Argumentation on the claims and evidences						
0	1	2	3	4	Sequence of the group presentations (about their claims and evidences) are appropriate for the subject matter and flow of the discussion	
0	1	2	3	4	Questions that will lead the argumentation on the claims and evidences are clearly specified in the lesson plan and asked accordingly during the lesson	
0	1	2	3	4	Questions and guidance that will encourage students to support/refute/develop counter arguments are clearly planned and sufficiently provided	
0	1	2	3	4	Procedures to be followed to enable students to come to a conclusion from the discussions are clearly planned and sufficiently enacted	
Comparison of the findings/observations with the literature						
0	1	2	3	4	Guidelines to relate students' findings with the literature are clearly planned and sufficiently enacted	
0	1	2	3	4	Students are directed to appropriate and reliable resources	
Providing opportunities to reflect on the change of the ideas						
0	1	2	3	4	Opportunities are provided to students to realize changes in their ideas about the subject matter	
0	1	2	3	4	Assessment and evaluation procedures of the lesson are clearly planned and sufficiently enacted	
0	1	2	3	4	Assessment and evaluation procedures of the lesson are appropriate for the subject matter	
Linking the lesson with Nature of Science and Nature of Scientific Inquiry						
0	1	2	3	4	At least one of the Nature of Science and Nature of Scientific Inquiry themes are explicitly covered	
0	1	2	3	4	Details of linking the lesson with Nature of Science and Nature of Scientific Inquiry (opportunities to be provided to students in each phase of the lesson) are clearly planned and sufficiently enacted	
Linking the lesson with the subsequent lesson						
0	1	2	3	4	Linking the lesson with the subsequent lesson is appropriate	
Additional lesson plan components						
0	1	2	3	4	Security measures are clearly planned and sufficiently enacted	
0	1	2	3	4	Time planned for each stage of the lesson are appropriate and time management is properly enacted during the lesson	

General evaluation						
0	1	2	3	4	Subject matter knowledge of the teacher/teacher candidate is sufficient	
0	1	2	3	4	Classroom management skills of the teacher/teacher candidate are sufficient	

***Answers to the items in this section are open-ended.**		
General Evaluation	**Strengths**	**Weaknesses**
Implementation of the argumentation-based inquiry procedures		
Use of the Nature of Science and Nature of Scientific Inquiry Themes		
Total score		

Appendix 2. Argumentation-based Inquiry Rating Scale (Turkish)

Adı soyadı: Tarih:

Puan	Puanlama Kriterleri
Çok iyi (4)	İlgili maddeyi oluşturan unsurların tamamı zengin ayrıntılar ile birlikte mevcut, tam, uygun ve doğru bir şekilde planlanmış ve uygulanmıştır. Başka bir öğretmen bu planın ilgili maddesini değiştirmeden olduğu gibi kullanabilir.
İyi (3)	İlgili maddeyi oluşturan unsurların yarısından fazlası zengin ayrıntılar ile birlikte tam, uygun ve doğru bir şekilde planlanmış ve kısmen uygulanmıştır. Başka bir öğretmen bu planının ilgili maddesini küçük değişikliklerle kullanabilir.
Orta (2)	İlgili maddeyi oluşturan unsurların yaklaşık yarısı bazı ayrıntılar ile birlikte mevcut tam, uygun ve doğru bir şekilde planlanmış ancak uygulanamamıştır. Başka bir öğretmen bu planının ilgili maddesini değişiklikler yaparak kullanabilir.
Zayıf (1)	İlgili maddeyi oluşturan unsurların yarısından azı küçük detaylar ile birlikte mevcut, tam, uygun ve doğrudur. Başka bir öğretmenler bu planının ilgili maddesini yeniden planlamadır.
Uygun değil / Kabul edilemez (0)	İlgili maddenin temel unsurları mevcut değil. Açıklamalar uygun değil.

Kriterler						Açıklamalar
Ders Öncesi Hazırlık						
0	1	2	3	4	Ders için seçilen kavramlar ve /veya beceriler MEB güncel Fen Bilimleri Dersi programına uygundur.	
0	1	2	3	4	Ders planı uygun kazanım/lar içermektedir.	
0	1	2	3	4	Planlanan ders için hazırlanan büyük düşünce ve alt düşünceler uygundur.	
0	1	2	3	4	Oluşturulan kavram haritası konu ile ilgili birçok kavramı ve kavramlar arasındaki ilişkiyi içermektedir.	
Araştırılacak Soru Üzerinde Uzlaşma						
0	1	2	3	4	Giriş etkinliği öğrencilerin kazanım/lara yönelik önbilgilerini ortaya çıkarır bir şekilde planlanmış ve uygulanmıştır.	
0	1	2	3	4	Giriş etkinliği öğrencilerin öğrenmeye olan ilgilerini artıracak şekilde planlanmış ve uygulanmıştır.	
0	1	2	3	4	Giriş etkinliği öğrencilerin tartışmaları ve soru sormaları için fırsat/lar sunacak şekilde planlanmış ve uygulanmıştır.	
0	1	2	3	4	Giriş etkinliği dikkat çekicidir ve öğrencileri merak ettikleri sorulara götürecek şekilde planlanmış ve uygulanmıştır	
0	1	2	3	4	Giriş etkinliği tartışma başlatacak ve devam ettirecek sorular içerecek seklide planlanmış ve uygulanmıştır.	
0	1	2	3	4	Öğrencilerden beklenen araştırma soruları ders planında yeterince belirtilmiş ve beklenilen araştırma sorularının öğrencilerden gelmemesi durumunda yapılabilecekler planlanmıştır.	
0	1	2	3	4	Öğrencilerin ihtiyaç duyabileceği malzemeler eksiksiz olarak belirtilmiş ve temin edilmiştir.	
Öğrencilerin araştırma sorularını test etmesi/ araştırması/deney (etkinlik) yapması						
0	1	2	3	4	Öğrenciler araştırma sorularına uygun deneyler/araştırmalar/gözlemler yapmaları için yönlendirilmiştir	
0	1	2	3	4	Öğrencilerin araştırma sorularını test etmesi için yapılması planlanan öğrenme aktiviteleri öğrenci merkezlidir.	
0	1	2	3	4	Deneyler/araştırmalar/gözlemler esnasında nelere dikkat edilmesi gerektiği açıkça örnek/ler ile belirtilmiştir ve uygulanmıştır.	
İddia ve delil üretme						
0	1	2	3	4	Öğrencilerin elde ettikleri verilerden yola çıkarak deliller ve iddialar oluşturmalarının nasıl sağlanacağı örnek/ler ile ders planında belirtilmiş ve uygulanmıştır.	
0	1	2	3	4	Öğrencilerin soru-iddia-delil arasındaki ilişkiyi kurmalarının nasıl sağlanacağı örnek/ler ile ders planında belirtilmiş ve ders uygulamasında sağlanmıştır.	

0	1	2	3	4	Öğrencilerin araştırma sorularına yönelik ikna edici bir argüman oluşturmalarının nasıl sağlanacağı planda belirtilmiş ve uygulanmıştır.	
Argümanların savunulması ve uzlaşma süreci (İddia ve delillerin savunulduğu tartışma)						
0	1	2	3	4	Argümanların savunulduğu tartışma sürecinde, öğrenci gruplarının konuya ve tartışmanın akışına uygun sıraya göre iddia ve delillerini sunması hem planlanmış hem de uygulanmıştır.	
0	1	2	3	4	Tartışmayı yönlendiren öğretmen soruları açıkça planda belirtilmiş ve uygulamada sorulmuştur.	
0	1	2	3	4	Öğrencileri, sunulan argümana karşı destekleme/çürütme/karşı argüman oluşturma konusunda teşvik edecek sorular ve yönlendirmeler planlanmış ve uygulanmıştır.	
0	1	2	3	4	Bu aşamada yapılan tartışmalardan öğrencilerin bir sonuca varmasının nasıl sağlanacağı planda belirtilmiş ve uygulanmıştır.	
Bulduklarının okudukları ile karşılaştırılması						
0	1	2	3	4	Öğrencilerin buldukları sonuçlar ile alanyazındaki bulguları ilişkilendirebilmeleri için yönlendirmeler planlanmış ve uygulanmıştır.	
0	1	2	3	4	Öğrenciler konu ile uygun güvenilir kaynaklara yönlendirilmiştir	
Fikirlerin nasıl değiştiğini yansıtmak için fırsatlar sağlama						
0	1	2	3	4	Öğrencilerin araştırma boyunca dersin konusuna dair düşüncelerindeki değişim fark ettirilmiştir	
0	1	2	3	4	Öğrencilerin dersi anlayıp anlamadıklarının nasıl değerlendirileceği açık bir şekilde ders planında belirtilmiş ve derste uygulanmıştır.	
0	1	2	3	4	Yapılan ölçme ve değerlendirme etkinliği konuya uygundur.	
Bilimin/bilimsel sorgulamanın doğası ile ilişki kurma						
0	1	2	3	4	Ders boyunca bilimin/bilimsel sorgulamanın doğası temalarından en az birine açık bir şekilde planda yer verilmiş ve derste uygulanmıştır.	
0	1	2	3	4	Bilimin/ bilimsel sorgulamanın doğası ile ilişki kurma ve nasıl vurgu yapılabileceği adına dersin hangi aşamasında öğrenciye ne tür fırsatlar sunulacağı örneklerle planda belirtilmiş ve derste uygulanmıştır.	
Bir sonraki derse geçiş						
0	1	2	3	4	Bir sonraki konuya geçiş uygun bir şekilde planda belirtilmiş ve derste uygulanmıştır.	
İlave Ders Planı Bileşenleri						
0	1	2	3	4	Gerekli güvenlik önlemleri planda belirtilmiş ve derste uygulanmıştır.	
0	1	2	3	4	Ders planı aşamalarının her biri için belirlenen süre uygun bir şekilde planlanmış ve uygulamada zaman yönetimi sağlanmıştır	
Genel değerlendirme						
0	1	2	3	4	Öğretmen/öğretmen adayı yeterli konu alan bilgisine sahiptir, bunu dersi planına ve uygulamaya yansıtmaktadır.	
0	1	2	3	4	Öğretmen/öğretmen adayı sınıf yönetimi açısından öğrencileri ve süreci yönetebilmektedir	

***Bu bölümdeki maddelere verilecek cevaplar açık uçludur.**		
Genel Ders Değerlendirmesi	**Güçlü yönleri**	**Zayıf yönleri**
Öğretmen adayının argümantasyon tabanlı bilim öğrenme sürecini uygulaması ve yönetmesi ile ilgili genel değerlendirme		
Öğretmen adayının planladığı dersi uygularken bilimin/bilimsel sorgulamanın doğası temalarını kullanımını ile ilgili genel değerlendirme		
Değerlendirme Sonucu Alınan Toplam Puan		

Appendix 3. Lesson Plan Template for Argumentation-based Inquiry (English)

<table>
<tr><td>Group members</td><td colspan="2"></td></tr>
<tr><td>Name of the group member who implement the lesson</td><td colspan="2">Date:</td></tr>
<tr><td>Name of the unit:</td><td colspan="2"></td></tr>
<tr><td>Grade level</td><td colspan="2"></td></tr>
<tr><td>Duration</td><td colspan="2"></td></tr>
<tr><td>Subject</td><td colspan="2"></td></tr>
<tr><td rowspan="2">Objectives (science):
Please consult current Science Curriculum for determining objectives
Objectives (Nature of science/ Nature of scientific inquiry):
Please write objectives related to Nature of Science/Nature of Scientific Inquiry themes.</td><td colspan="2"></td></tr>
<tr><td colspan="2"></td></tr>
<tr><td>The big idea and sub-ideas of the unit
Write the sub-idea of the unit that will guide you in this lesson in bold*</td><td colspan="2">The big idea of the unit:

Sub-ideas of the unit:</td></tr>
<tr><td>Concepts:</td><td colspan="2"></td></tr>
<tr><td>Skills (e.g., Science Process Skills, Life Skills, Engineering and Design Skills, etc.)**</td><td colspan="2"></td></tr>
<tr><td>Teaching methods and techniques
Note: This course will be planned based on Argumentation Based Inquiry Approach. Please indicate the teaching methods and techniques you will utilize during the lesson.</td><td colspan="2"></td></tr>
<tr><td>Nature of Science/Nature of Scientific Inquiry themes that will be addressed during the lesson:

(<u>You need to address at least one of Nature of Science/Nature of Scientific Inquiry themes</u>)</td><td>☐ Tentativeness of scientific knowledge
☐ Science is empirical based
☐ Subjectivity and theory-laden of scientific knowledge
☐ Creativity and imagination
☐ Socio-cultural embeddedness
☐ Science is based on observation and inferences
☐ Scientific theories and Laws</td><td>☐ Scientific investigations all begin with a question and do not necessarily test a hypothesis;
☐ There is no single set or sequence of steps followed in all investigations;
☐ Inquiry procedures are guided by the question asked;
☐ All scientists performing the same procedures may not get the same results;
☐ Inquiry procedures can influence results;
☐ Research conclusions must be consistent with the data collected;
☐ Scientific data are not the same as scientific evidence;
☐ Explanations are developed from a combination of collected data and what is already known</td></tr>
<tr><td>Safety precautions:</td><td colspan="2"></td></tr>
</table>

Pre-lesson preparation: (Constructing the concept map and determination of the big idea and the sub-ideas) *** *** *Please attach the concept map to your lesson plan as Appendix 1* Please explain the way you followed for determining the big idea and the sub-ideas	
1. **Discussion on the research question to be investigated (Planning of the introductory activity)**	**Duration:** *Please indicate how much time you plan to spend for this section of the lesson plan
What can I do to prepare the learning environment and get students' attention?	
What are the questions that will start and continue the introductory discussion?	
What are the research questions expected from students?	
What can I do if I do not receive the research questions I expect from students?	
What are the materials students might need to answer their research questions?	
2. Testing/investigating research questions	**Duration:** *Please indicate how much time you plan to spend for this section of the lesson plan
How can I guide students to make experiments/research/observations appropriate with their research questions?	
What should I pay attention to while students test/investigate their research questions?	
3. Claims and evidences	**Duration:** *Please indicate how much time you plan to spend for this section of the lesson plan
How can I get students to create evidence and claims based on the data they have obtained? How can I direct students to establish the relationship between question-claim-evidence?	
4. Argumentation on the claims and evidences	**Duration:** *Please indicate how much time you plan to spend for this section of the lesson plan
How should I lead the discussion? (e.g., What can I ask during the discussion? How should I end the discussion? etc.)	
What are the topics (concepts, relationships between concepts, events,	

phenomena etc.) that should theoretically addressed in this course? (*Please explain them as you plan to address in the lesson)	
5.Comparison of the findings/observations with the literature How can I get students to compare their results with findings in the literature? What are the resources that I especially expect students to read? How can I direct students on this issue? * Please clearly specify the reference/links of the resources.	
6.Providing opportunities to reflect on the change of the ideas How can I direct students to realize changes in their ideas about the subject matter?	
7. Assessment & Evaluation How can I assess and evaluate students for this lesson? Which measuring tools can I use? What might my questions in these measurement tools be? *Please pay attention to use alternative assessment and evaluation tools such as concept map, fish bone, etc.	
8. Linking the lesson with Nature of Science and Nature of Scientific Inquiry Please clearly specify the stages that you will link the lesson with Nature of Science and Nature of Scientific Inquiry. Please clearly explain how you plan to link the lesson with Nature of Science and Nature of Scientific Inquiry.	
9. Linking the lesson with the subsequent lesson How can I link the lesson with the subsequent lesson? * You can leave this section blank if a new unit starts after this lesson.	

***a. Science Process Skills:** Include skills such as observing, measuring, classifying, recording data, making hypotheses, using and modeling data, changing and controlling variables, and conducting experiments etc. that scientists use during their studies.

***b. Life Skills:** Include skills such as analytical thinking, decision-making, creativity, entrepreneurship, communication and teamwork, etc. that are used for accessing and using scientific knowledge.

* **c. * Engineering and Design Skills:** Include innovative thinking skills.

* **Big idea and sub-ideas:** Big idea is the basic idea that forms and reflects the roof of the unit and subject. The process / activities that will take place throughout the unit are planned around the big idea. It should cover the whole unit and reflect the goal we want to achieve at the end of the unit.

Sub-idea is the basic idea of each activity (the lesson you plan for 2 lesson hours) that we will do to reach the big idea. Determine how many lesson activities are needed to reach the big idea. Each lesson activity should target a sub-idea/sub-ideas. The sub idea(s) that you have identified should lead us to the big idea at the end of the unit.

Features of big idea:

- It should cover the whole topic/unit and emphasize the main point.
- It should be clear, understandable, meaningful and express a judgment that consists of a few words
- Should reflect the goal we want to achieve at the end of the unit

Features of sub-idea:

- Should be determined for each activity to be held throughout the unit
- Should be basically linked to the big idea of the unit but more specific when compared to the big idea
- Should be clear, understandable, meaningful and express a judgment that consists of a few words
- Should guide the teacher in planning their activities.

Example:

Unit: Force and Motion

Big idea: Matters move under the effect of force.

Sub ideas: 1- **If the object has a bigger density than a liquid, it floats; if it is not, it sinks**

2- Gases and liquids exert buoyancy**.**

3- Force causes pressure.

Note: See Yesildag-Hasançebi and Akbay (2017) for further details.[†]

[†]Yesildag-Hasancebi, F., & Akbay, Y. (2017). The role of big ıdea in argumentation based science ınquiry classrooms. In Hand, B., Norton-Meier, L., Jang, Jy. (eds), *More voices from the classroom* (pp. 35-44). SensePublishers. https://doi.org/10.1007/978-94-6351-095-0_3

Appendix 4. Lesson Plan Template for Argumentation-based Inquiry (Turkish)

Grup Elemanlarının Adı Soyadı		
Dersi Uygulayan Grup Elemanı	**Tarih:**	
Ünitenin Adı:		
Dersin Sınıf Seviyesi		
Dersin Süresi		
Konu:		
Kazanımlar: Fen kazanımı için fen öğretim programından yararlanınız. **Bilimin/bilimsel sorgulamanın doğası kazanımı:** Planladığınız derste yer alacak bilimin /bilimsel sorgulamanın temasına yönelik kazanım yazınız		
Dersin büyük düşüncesi ve alt düşünceleri Yazdığınız alt düşüncelerden bu ders ile ilgili olan (sizi yönlendirecek olan) alt düşünceyi koyu renk yaparak belirtiniz.*	**Büyük düşünce:** **Alt düşünceler:** **1.** **2.**	
Kavramlar:		
Beceriler (BSB -Yaşam becerileri) Bu ders içerisinde öğrencilerin kazanabileceği Bilimsel Süreç Becerileri ve Yaşam Becerileri nelerdir?*		
Yöntem ve Teknikler Bu ders Argümantasyon Tabanlı Bilim Öğrenme yaklaşımı esas alınarak planlanacaktır. Süreçte kullanmak istediğiniz teknikler varsa belirtiniz. Ayrıcaders planınızın Bilimin Doğası temalarını içinde barındırmasına dikkat ettiniz.		
Derste Değinilebilecek Bilimin/Bilimsel Sorgulamanın Doğası Temaları: Bu derste bilimin ve bilimsel sorgulamanın doğası temalarından hangisi/hangilerine dikkat çekebilirim?	□ Bilimsel bilginin değişebilirliği □ Bilimsel bilginin deneysel yapısı □ Bilimsel bilginin öznel yapısı □ Bilimsel bilginin bilim insanının yaratıcılığını ve hayal gücünü içermesi	□ Bütün bilimsel araştırmalar bir soru ile başlar, ancak mutlaka bir hipotez ile test edilmesi gerekmez. □Tek bir bilimsel yöntem yoktur. □Sorgulama sürecine, sorulan sorular yön verir. □ Bilim insanları aynı prosedürleri uygulasalar bile aynı sonuçlara ulaşamayabilirler.

(En az bir tane bilimin/bilimsel sorgulamanın doğası temasını dersinize dahil etmelisiniz)	□ Bilimsel bilginin sosyal ve kültürel yapısı □ Bilimsel bilginin gözlem ve çıkarımlara dayanması □ Teoriler ve kanunlar arasındaki farklar	□Sorgulama süreçleri elde edilen sonuçları etkileyebilir. □Araştırma sonuçları toplanan veriler ile tutarlı olmalıdır. □ Bilimsel veriler ile bilimsel deliller birbirinden farklıdır. □Açıklamalar, toplanan veriler ve var olan bilgiler (ön bilgiler) ışığında geliştirilir.
Güvenlik önlemleri: **(**Deneyler esnasında ne tür güvenlik önlemleri almalıyız?)		
Ders öncesi hazırlık: (Kavram haritası yapılması ve büyük düşüncenin belirlenmesi) Büyük ve alt düşünce belirlemede izlediğiniz yolu aktarınız. * Kavram haritanızı EK-1 olarak ekleyiniz.		
1. **Araştırılacak Soru Üzerinde Uzlaşma**	**Süre:** *Bu bölümü kaç dakikada gerçekleştirmeyi planladığınızı yazınız.	
Ortamı hazırlama ve dikkat çekme için ne yapabilirim?		
Giriş tartışmasını başlatacak ve devam ettirecek sorular neler olabilir? Bu süreçte öğrencilere sormayı planladığınız soruları yazınız.		
Öğrencilerden beklenen araştırma soruları nelerdir?		
Beklediğim araştırma soruları öğrencilerden gelmezse ne yapabilirim?		
Öğrencilerin araştırma sorularına cevap bulmak için ihtiyaç duyabileceği malzemeler nelerdir?		
2. Araştırma Sorularını Test Etme/Araştırma/Deney Yapma	**Süre:** *Bu bölümü kaç dakikada gerçekleştirmeyi planladığınızı yazınız.	
Soruları test ettirebilmek için ne yapabilirim? Öğrencileri araştırma sorularına uygun deneylere nasıl yönlendirebilirim?		
Deneyler/gözlemler/araştırmalar esnasında nelere dikkat etmeliyim?		
3. İddia ve Delil Üretme	**Süre:** *Bu bölümü kaç dakikada gerçekleştirmeyi planladığınızı yazınız.	

Öğrencilerin elde ettikleri verilerden yola çıkarak deliller ve iddialar oluşturmalarını nasıl sağlarım? Öğrencilerin soru-iddia- delil arasındaki ilişkiyi kurmalarını sağlamak için onları nasıl yönlendirebilirim?	
4. Argümanların Savunulması ve Uzlaşma Süreci (İddia ve Delillerin Savunulduğu Tartışma)	**Süre:** *Bu bölümü kaç dakikada gerçekleştirmeyi planladığınızı yazınız
Tartışmayı nasıl yönlendirmeliyim? Hangi soruları sorabilirim? Tartışmayı nasıl sonlandırırım?	
Teorik olarak bu derste değinilmesi gereken konular (kavramlar, kavramlar arası ilişkiler, olaylar, olgular vb.) neler olmalı? (Konu ile ilgili teorik bilgiyi ders planının bu bölümünde yazabilirsiniz)	
5.Bulduklarımın Okuduklarım ile Karşılaştırılması (Uzmanların konu hakkında ne söylediğini belirleme) Öğrencilerin buldukları sonuçlar ile bilimsel sonuçları karşılaştırmalarını nasıl sağlarım? Özellikle öğrencilerin okumasını beklediğimiz metinler neler olabilir? Bu konuda öğrencileri nasıl yönlendirmeliyim? * Okuma örneklerine ait referans/link açık bir şekilde belirtilmelidir.	
6. Öğrencilerin Fikirlerinin Nasıl Değiştiğini Yansıtmak İçin Fırsatlar Sağlama Öğrencilerin araştırma boyunca dersin konusuna dair düşüncelerindeki değişimi onlara nasıl fark ettiririm?	
7. Ölçme-Değerlendirme Öğrencilerin dersi anlayıp anlamadıklarını nasıl değerlendiririm? Hangi ölçme araçlarını kullanabilirim? Bu ölçme araçlarındaki sorularım neler olabilir? *Özellikle alternatif ölçme değerlendirme araçlarını (kavram haritası, anlam çözümleme tablosu, balık kılçığı vb.) kullanmaya özen gösteriniz	

8. Bilimin/Bilimsel Sorgulamanın Doğası ile İlişki Kurma Bilimin/bilimsel sorgulamanın doğası ile ilişki kurma adına dersin **hangi aşamasında** ne tür fırsatlar olabilir? Derste Bilimin/bilimsel sorgulamanın doğası temalarından hangisine/ hangilerine **nasıl** vurgu yapabilirim	
9. Bir Sonraki Derse Geçiş Bir sonraki konuya/derse geçişi nasıl sağlarım? Öğrencileri nasıl yönlendiririm? *Planladığınız dersten sonra yeni bir ünite başlıyorsa bu bölümü boş bırakabilirsiniz	

***a. Bilimsel Süreç Becerileri:** Bu alan; gözlem yapma, ölçme, sınıflama, verileri kaydetme, hipotez kurma, verileri kullanma ve model oluşturma, değişkenleri değiştirme ve kontrol etme, deney yapma gibi bilim insanlarının çalışmaları sırasında kullandıkları becerileri kapsamaktadır.

***b. Yaşam Becerileri:** Bu alan; bilimsel bilgiye ulaşılması ve bilimsel bilginin kullanılmasına ilişkin analitik düşünme, karar verme, yaratıcılık, girişimcilik, iletişim ve takım çalışması gibi temel yaşam becerilerini kapsamaktadır.

***Mühendislik ve Tasarım Becerileri**: Bu alan yenilikçi (İnovatif) düşünme becerisini kapsamaktadır.

***Büyük düşünce ve alt düşünceler:** Büyük düşünce ünite ve konunun çatısını oluşturan ve onu yansıtan temel düşüncedir. Ünite boyunca gerçekleşecek süreç/etkinlikler büyük düşünce etrafında planlanır. Tüm üniteyi kapsamalı ve ünite sonunda ulaşmak istediğimiz hedefi yansıtmalıdır. Alt düşünce ise büyük düşünceye ulaşmamız için yapacağımız her bir etkinliğin (2 ders saati için planladığınız dersin) temel düşüncesidir. Büyük düşünceye ulaşmak için kaç ders etkinliği gerekiyorsa her biri için bir düşünce belirleyiniz (Yani bir ünite kaç aşamada işlenecekse her bir aşamanın hedeflediği bir düşünce olmalıdır). Belirlediğiniz bu alt düşünceler ünite sonunda bizi büyük düşünceye ulaştırmalıdır. (Yeşildağ-Hasançebi & Akbay, 2017) Aşağıdaki örneği inceleyiniz.

Not: Hazırladığınız ders planı ünite bazında belirlenen alt düşüncelerden hangisi ile ilgili ise onu koyu renk yaparak belirtiniz. Diğer alt düşünceleri planlamak zorunda değilsiniz.

Büyük düşüncenin özellikleri

- Tüm konuyu/üniteyi kapsamalı ve temel noktaya vurgu yapmalıdır.
- Açık, anlaşılır, anlamlı olmalı ve birkaç kelimeden oluşan bir yargı bildirmelidir.
- Ünite sonunda ulaşmak istediğimiz hedefi yansıtmalıdır.

Alt düşüncenin özellikleri

- Ünite boyunca yapılacak her etkinlik için belirlenir.
- Temelde büyük düşünceye bağlıdır ama daha özeldir.
- Açık, anlaşılır, anlamlı olmalı ve birkaç kelimeden oluşan bir yargı bildirmelidir.
- Öğretmenin etkinliklerini planlamada ona yol gösterir.

Büyük düşünce ve alt düşüncenin özellikleri ve bir fizik ünitesi için örnek aşağıda sunulmuştur (Yeşildağ-Hasançebi & Akbay, 2017)

Örnek: **Fizik ünitesi:** Kuvvet ve Hareket Ünitesi

Büyük düşünce: Maddeler kuvvetin etkisiyle hareket eder.

Alt Düşünceler: 1) **Cisim; sıvı içinde yoğunsa batar değilse yüzer**

2) Gazlar ve sıvılar kaldırma kuvveti uygular.

3) Kuvvet basınca neden olur

Yesildag-Hasancebi, F., & Akbay, Y. (2017). The role of big idea in argumentation-based science ınquiry classrooms. Ed. Hand, B., Norton-Meier, L., Jang, Jy. (eds), *More voices from the classroom* (pp. 35-44). Sense Publishers. https://doi.org/10.1007/978-94-6351-095-0_3

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 4, 998–1012
https://doi.org/10.21449/ijate.1058300
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Using Rasch analysis to examine raters' expertise Turkish teacher candidates' competency levels in writing different types of test items

Ayfer Sayin[1,*], **Mehmet Sata**[2]

[1]Gazi University, Faculty of Education, Department of Educational Sciences, 06500, Ankara, Türkiye
[2]Ağrı İbrahim Çeçen University, Faculty of Education, Department of Educational Sciences, Kars, Türkiye

ARTICLE HISTORY

Received: Jan. 15, 2022
Revised: Oct. 12, 2022
Accepted: Nov. 29, 2022

Keywords:
Test item,
Raters' expertise,
Many Facet Rasch,
Validity,
Reliability.

Abstract: The aim of the present study was to examine Turkish teacher candidates' competency levels in writing different types of test items by utilizing Rasch analysis. In addition, the effect of the expertise of the raters scoring the items written by the teacher candidates was examined within the scope of the study. 84 Turkish teacher candidates participated in the present study, which was conducted using the relational survey model, one of the quantitative research methods. Three experts participated in the rating process: an expert in Turkish education, an expert in measurement and evaluation, and an expert in both Turkish education and measurement and evaluation. The teacher candidates wrote true-false, short response, multiple choice and open-ended types of items in accordance with the Test Item Development Form, and the raters scored each item type by designating a score between 1 and 5 based on the item evaluation scoring rubric prepared for each item type. The study revealed that Turkish teacher candidates had the highest level of competency in writing true-false items, while they had the lowest competency in writing multiple-choice items. Moreover, it was revealed that raters' expertise had an effect on teacher candidates' competencies in writing different types of items. Finally, it was found that the rater who was an expert in both Turkish education and measurement and evaluation had the highest level of scoring reliability, while the rater who solely had expertise in measurement and evaluation had the relatively lowest level of scoring reliability.

1. INTRODUCTION

Language is the most effective means by which human beings convey their feelings and opinions. Language education is a developmental process which starts at birth – even before birth – and continues a lifetime. Thus, Turkish education programs that also constitute the basis of other disciplines are based on four fundamental skills, namely reading, writing, listening and speaking. The Ministry of National Education (MoNE) reports that "The Turkish Education Program is regarded as the development of language skills and competencies and a prerequisite to learning, personal and social development and acquisition of vocational skills" (2019). This statement indicates that language skills essentially form the basis of other disciplines. It is

*CONTACT: Ayfer SAYIN ayfersayin@yahoo.com Gazi University, Faculty of Education, Department of Educational Sciences, Assessment and Evaluation in Education, 06500 Beşevler, Ankara, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

known that teacher quality has an important role in students' reaching the learning outcomes in education programs. It is important to utilize valid and reliable tools not only to identify the extent to which students reach the learning outcomes in the program and to make decisions about students, but also to provide students with effective feedback. Thus, in the present study, the aim was to examine Turkish teacher candidates' competencies in writing different types of items to measure reading comprehension skills. With respect to the reading comprehension skill in Turkish education programs, the aim is for students to read fluently and to accurately comprehend the texts they encounter in their daily life by using the right methods, to critically interpret and evaluate what they read, and to adopt the habit of reading (MoNE, 2019). Reading comprehension skills are observed to have an important place in the Turkish language test section of exams administered within the school transitional system in Turkey. Furthermore, the importance of developing students' reading comprehension skills is also highlighted in such international test administrations as PIRLS and PISA. As in all skills and competencies, it is essential not only to equip students with reading comprehension skills but also to measure these skills in a valid and reliable way. In parallel to the changes in the expertise expected of an individual in the 21st century, the changes in teaching and learning environments should be reflected in the measurement tools as well. In other words, in an education system where the development of students' higher order skills is aimed at, measurement tools are also expected to have the quality of measuring higher order skills (Sayın & Kahraman, 2020).

During pre-service trainings, teachers receive training in writing items in accordance with item writing principles and writing items that can measure not only lower-level skills but also higher order skills. Test development includes the processes of individuals' use of knowledge, abilities, talents, areas of interest, attitudes and other characteristic expertise to develop items and transform them into a test format within the framework of a plan. It also includes the procedures of identifying the appropriate test administration conditions, how the scoring of the test performance is to be done and how the scores are to be announced to the test takers (Crocker & Algina, 2008). Even though details regarding test development, which includes numerous steps and a long process, vary in different sources (Linn & Gronlund, 2000; Walsh & Betz, 1995), test development is comprised of the following steps: identifying the purpose of the test, defining the constructs to be measured via the test, writing the items, revising the items based on expert opinion, preparing the pilot form, conducting a pilot study, scoring, item analysis, selection of items, and finalizing the test (Baykul, 2000). However, such institutions as the Higher Education Council (HEC) and MoNE in Turkey, which administer high scaled tests, are unable to conduct their pilot studies during the test development process owing to issues of confidentiality. In-class tests are also developed generally without a pilot study, based solely on expert opinion, because of the small number of participants and other reasons. In other words, the test development process is completed at the stage when items are evaluated based on expert opinion. Thus, expertise of the experts to evaluate the test items formed during test development comes forward. It is imperative that items measuring the target learning outcome be developed in accordance with measurement and evaluation principles. Even if it has a correct response, an item that is not well-structured may not serve its purpose. For this reason, it was ensured that the raters participating in the present study to evaluate the test items had diverse expertise.

Since the study aimed to determine the effect of rater qualifications in evaluating the different item-type writing skills of pre-service teachers, the multi-faceted Rasch model was used. It gives individual and group-level statistics on a single comparable scale (logit scale) (Linacre, 1993). In addition, the multi-faceted Rasch model contributes to the reliability and validity of the measurements in determining the expected effects of the variability within the scope of the research (e.g., the mutual interactions between the rater and the item type). When a multi-faceted Rasch bias analysis is performed, the researcher looks for evidence in the rater's scoring

pattern (Myford & Wolfe, 2003). The effects of rater biases, beliefs, or personal characteristics on scoring behavior can be studied using the multi-faceted Rasch measurement model approach. Similarly, the effects of the rater's past experiences on the scoring behavior can be examined. The multi-surface Rasch approach was preferred in this context in the related research.

When the rater effect is mentioned, it was examined whether the raters were experienced (Barkaoui, 2010; Davis, 2016; Erman Aslanoğlu & Şata, 2021; Kim, 2020) or the scoring rigidity within themselves (Anthony, Styck, Volpe, & Robert, 2022; Jones & Bergin, 2019; Kamış & Doğan, 2017; Primi, Silvia, Jauk, & Benedek, 2019). In this research, the effect of the field expertise of the raters was examined, which is quite significant in terms of both the examination and the result. Since it is essential that the people who will work in the test development process give information about their expertise; similarly, it is expected to contribute to the field by giving feedback on item types and seeing which item types the pre-service teachers are better.

Just as the in-class learning outcomes to be measured and their levels vary, the item types to be included in a test also vary because true-false and short response items that are appropriate for measuring all kinds of learning outcomes at lower levels may not be conducive to measuring higher order level skills (Özçelik, 2010b). Hence, including different types of items in a test to form evidence for content validity is also important. Gorin (2007) and Sireci (2007) state that for any condition of assessment, there generally needs to be more than one test and item type.

1.1. Research Questions

1. Do raters’ expertise influence the process of evaluating teacher candidates' competency levels when developing test items?
2. Do Turkish teacher candidates' competencies differ when writing different test items?
3. What kind of interaction exists between raters’ expertise and teacher candidates' competency levels in writing different test items?

2. METHOD

2.1. Research Model

In the present study, the relational survey design, one of the quantitative research methods, was employed. The aim in a relational survey model is to examine the existence and degree of a relationship between two or more variables without any intervention (Büyüköztürk et al., 2018; Karasar, 2018).

2.2. Study Group

The study group of the present study was comprised of 84 Turkish teacher candidates whose %71 (n=60) is female, and 29% (n=24) is male. They are at the 6th term of thecurriculum , and the teacher candidates started to write the items ten weeks after attending their measurement and evaluation course. The test items developed by the teacher candidates were scored by three raters with different expertise. One of the raters was an expert in measurement and evaluation (Rater 3), one was an expert in Turkish language education (Rater 2), and the final rater was an expert in both Turkish education and measurement and evaluation (Rater 1).

2.3. Data Collection Tools

The data collection process was performed in two stages. First, the Turkish teacher candidates were required to develop a test consisting of different types of items. Subsequently, the items produced were evaluated.

2.3.1. *Item writing*

After the 12 hours of face-to-face education that teacher candidates received during the test development unit in the measurement and evaluation course, they formed specification tables based on the learning outcomes regarding reading comprehension skill in the Turkish education program. As the curriculum is spiral in nature, there are similarities between the prescribed learning outcomes for different grade levels. After the preparation of the specifications table, the teacher candidates were asked to write the learning outcomes planned to be measured by means of true-false, short response, multiple choice and open-ended items. After matching the learning outcomes with the appropriate item type, the teacher candidates passed onto the stage of selecting texts. By its very nature, the reading comprehension skill is shaped based on the type of text used. Such expertise as length of text, style of expression and statements have a direct impact on the type and level of the item to be developed (Sayın & Takıl, 2017). The items based on the related learning outcomes that were written based on the selected or written texts in accordance with the points to be considered in text selection were written on the item writing form. The form consisted of five sections: the related learning outcome(s), text, instruction, items, and answer key. In addition, at the beginning of the form was included a section on the item writing principles to be considered for each item type. The teacher candidates wrote a total of 14 items: 5 true-false, 5 short response, 5 multiple choice and 1 open-ended. As the teacher candidates initially organized their texts, and then wrote items based on these texts, the probability of copying their items from elsewhere was minimized. Moreover, the items written by the teacher candidates were checked for originality via a software before the rating stage began.

2.3.2. *The Scoring of the items*

The test consisting of different item types and developed by the teacher candidates within the scope of this study was scored with the use of a holistic rubric developed for each test item by the researchers. Taking into consideration the qualities that test items need to possess, the researchers based the rubric on a five-point measurement scale. Each item type was scored within its own category. During the scoring stage, three experts were asked to assign a score for each item. With the aim of identifying the impact of raters' expertise on scoring, the raters' areas of expertise showed variation. The first rater (Rater 1) was an expert in both Turkish education and measurement and evaluation. The second rater (Rater 2) was an expert in Turkish education but did not have direct expertise in measurement and evaluation. The third rater (Rater 3) was an expert in measurement and evaluation but did not have direct expertise in Turkish education. Using the holistic rubric, the raters independently rated all the item types written by all the teacher candidates.

After the holistic rubric was prepared and used, data was collected for the validity and reliability of the measurements (Appendix 1). Factor analysis was utilized for the validity of the measurements, and the McDonald (1999) ω coefficient was employed for reliability purposes. Since the factor loading of each criterion is different (since the congeneric measurement is in item), the omega coefficient, which makes a more consistent estimation, was used (Osburn, 2020). Prior to an exploratory factor analysis (EFA) for validity, the underlying assumptions of this analysis need to be tested. Hence, the statistical analyses to test the assumptions revealed that the requiured minimum sample size (minimum five people per variable) was met, there was no outliers or loss of data in the data set, there was a linear relationship among the criteria of the measurement tool, and all the variables showed a normal distribution. Ater all the assumptions were found to be met, whether or not the data set could be factorized was examined, and it was revealed that it could be (for the related data set the Kaiser-Meyer-Olkin value was found to be .654, and the Bartlett's sphericity test was found to be statistically significant (χ^2(fd) = 37.411 (6), p = .000)). According to the EFA results, it was found that the

measurement tool represented a single factor structure (The variance explained was 46.67%, and the factor loadings of the criteria were 0.803, 0.653, 0.595, 0.665, respectively). After evidence for the validity of the measurements was obtained, the McDonald ω coefficient was used to assess the reliability of the measurement tool. As a result of the analysis run via the Mplus (version 8) package program, the McDonald ω coefficient was found to be .733. Based on these findings, it can be claimed that the measurements obtained from the holistic rubric used to assess the teacher candidates' competencies in writing different types of test items were valid and reliable.

2.4. Data Analysis

In the present study, which aimed to evaluate teacher candidates' competency levels in writing different types of test items, the many facet Rasch analysis (Linacre, 2012) was used as it was appropriate for the nature of the study. Since more than one variable source can be analyzed simultaneously in many facet Rasch analysis, it can be used in many different designs. In the present study, there are three dimensions (source of variability): raters, teacher candidates, and item type. All the variability sources in the study were taken into consideration, and a full factorial design, in which all the raters, all the teacher candidates, and all the item types were evaluated, was utilized. During data analysis, the guidelines defined by Myford & Wolfe (2003, 2004) were taken into consideration. In accordance with these guidelines, the statistics of the group, followed by those of the individuals, were presented. As many facet Rasch analysis is a member of the item response theory, it rests on certain assumptions that need to be met (Farrokhi, Esfandiari & Schaefer, 2012; Farrokhi, Esfandiari & Vaez Dalili, 2011). These assumptions are unidimensionality, local independence and model-data fitting. In terms of the first assumption – unidimensionality – as stated in the measurement tool section, it was identified that the holistic scoring rubric was based on a single factor; that is, it met (the) unidimensionality assumption. Since the unidimensionality of a measurement tool indicates local independence, it was accepted that the assumption of local independence was also met. Finally, the standardized residual values were examined for the model-data fitting. To meet the assumption of model-data fitting, the number of standardized residual values that do not fall within the ±2 interval must not be more than 5% of the total observation numbers. Also, it is reported that the standardized residual values that do not fall within the ±3 interval should not be more than 1% of the total number of data (Linacre, 2017). When the standardized residual values were examined, it was found that there were 51 (5.06%) values within the ±2 interval and 11 items (1.09%) within the ±3 interval, thus concluding that the model-data fitting was at an acceptable level (total number of observations 3x4x84 = 1 008).

3. FINDINGS

In the present study, which aimed to evaluate Turkish teacher candidates' competency levels in writing different types of items, initially the impact of raters' expertise on the evaluations was examined. Within this scope, the measurement reports for the rater dimension were obtained and presented in Table 1.

As can be observed in Table 1, the discrimination ratio for the group level statistics, discrimination index and discrimination index reliability values were low (<0.70). The reliability of the discrimination index is interpreted as Cronbach's alpha coefficient, and values below .70 indicate that the reliability of individuals in discrimination according to their performance is low (Marais & Andrich, 2008).

Table 1. *Measurement report for the rater dimension*

Rater	Logit	Standard error	Infit		Outfit		*t*-value	Rasch-Kappa
			MnSq	ZStd	MnSq	ZStd		
Rater 1	+0.12	0.08	0.95	-0.40	0.78	-2.10	1.50	0.44
Rater 2	+0.02	0.08	0.95	-0.40	1.09	0.80	0.25	0.34
Rater 3	-0.15	0.07	1.08	0.70	1.18	1.70	-2.14	0.31
Mean	0.00	0.08	1.00		1.01			
SD	0.14	0.00	0.07		0.21			

Model, Sample: RMSE = .08 Standard deviation = .08
Discrimination ratio=1.43 Discrimination index = 2.25
Discrimination index of reliability= 0.67
Model, Fixed (all same) chi square=6.20 *df* =2 *p*= .04
Model, Random (normal) chi square =1.50 *df* = 1 *p*= .22
Observed inter-rater agreement: 67.00%
Expected inter-rater agreement: 48.10%
Kappa inter-rater reliability statistics: 0.37

$t_{critical}$ (0.05, 2) = 4.30; $\chi^2_{critical}$ (0.05, 2) = 5.99

Thus, this indicates that the scores of the raters who evaluated the teacher candidates' competency levels in writing different item types showed slight variations. The p-value for the fixed effects chi-square value regarding the statistical variation was found to be 0.04. A chi-square value that is higher than the critical chi-square value indicates that the measurements show a statistically significant difference. In other words, it indicates that raters' expertise had an impact on the evaluations. When the t-value for each rater was examined, and since the critical t-value was observed to be small, it was revealed that the evaluations made by the raters in the study showed similarity in levels of strict versus lenient scoring.

Even though there was no statistically significant difference between the raters' lenient or strict scoring levels, the examination of each rater's Rasch-Kappa values showed that the first rater had a higher level of reliability when compared to that of the other two raters. Accordingly, it was deduced that raters' expertise had an effect on teacher candidates' competency levels in writing different types of test items. An examination of raters' expertise revealed that the rater who had expertise in both Turkish education and measurement and evaluation had the highest level of reliability in scoring. Then followed the rater with expertise in solely Turkish education. The lowest reliability in scoring among the three raters belonged to the rater who had expertise solely in measurement and evaluation.

In the process of writing different items of Turkish teacher candidates, the measurement report on the item type related to a statistical difference according to item type was examined. This measurement report by item type is presented in Table 2. As can be observed in Table 2, the discrimination ratio for item types, the discrimination index and the discrimination reliability values are very high (>0.70). Moreover, the chi square value was found to be statistically significant. Accordingly, a variation was revealed between the competency levels of the teacher candidates in writing different types of test items. In order to identify the source of this variation at the group level, the variables at the individual level were examined. Initially, the logit values were calculated for each item type; the highest and lowest logit values were found to be 0.89 and -0.82, respectively. A positive logit value indicates a high level of item writing competency, while a negative logit value indicates a low competency level. Accordingly, the Turkish teacher

candidates' competency levels in writing true-false type of items were found to be high, while their competency levels in writing multiple choice items was found to be low.

Table 2. *Measurement report for the dimension of item type*

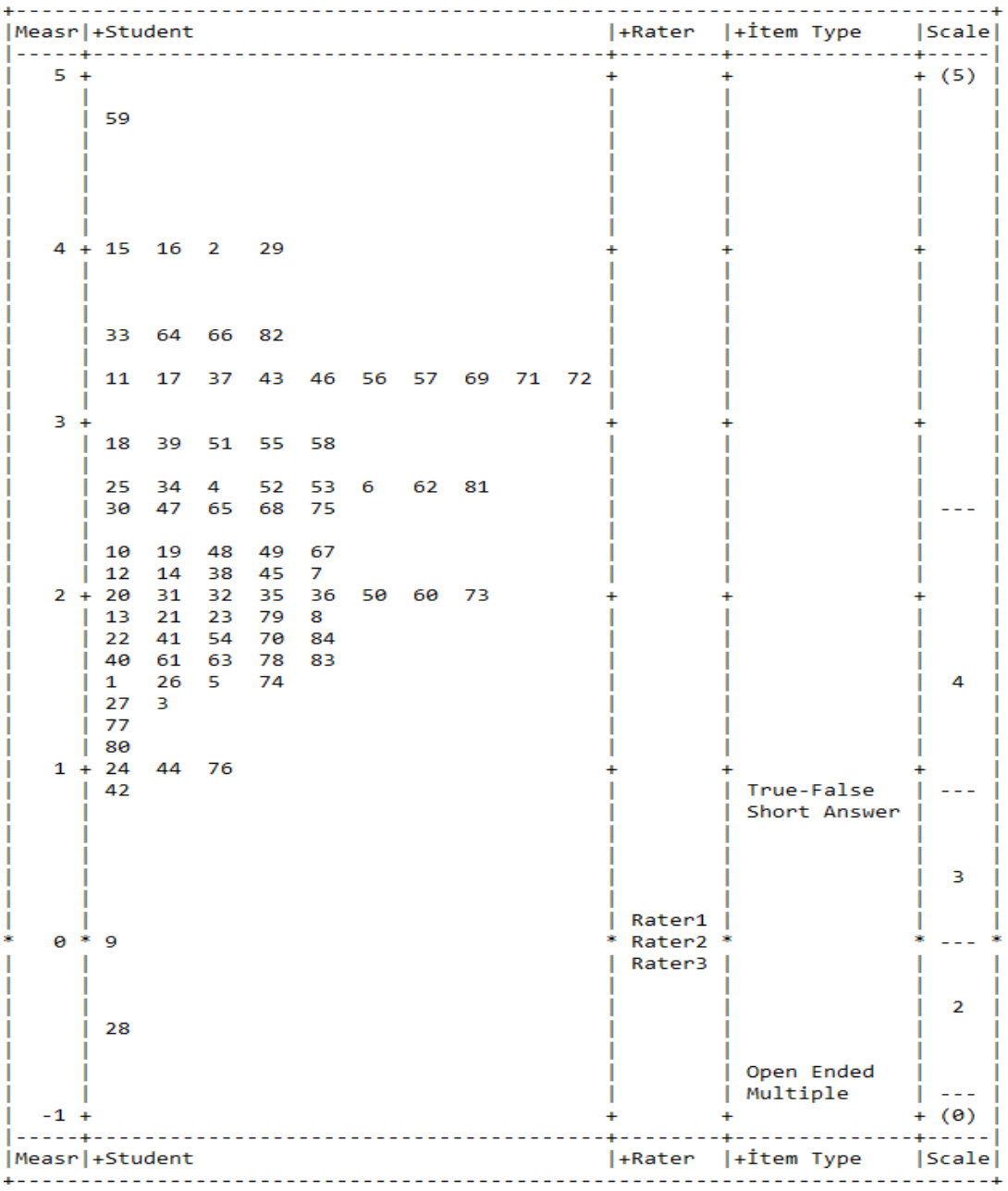
Item Type	Logit	Standard error	Infit		Outfit	
			MnSq	ZStd	MnSq	ZStd
True False	+0.89	0.12	0.87	-1.00	0.80	-1.40
Short response	+0.72	0.11	1.63	4.40	1.55	3.50
Open-ended	-0.79	0.08	0.71	-3.30	0.74	-2.90
Multiple choice	-0.82	0.07	1.06	0.60	0.97	-0.20
Mean	0.00	0.10	1.07		1.01	
Standard deviation	0.93	0.02	0.40		0.37	

Model, Sample: RMSE = .10 Standard deviation= .80
Discrimination ratio =9.44 Discrimination index =12.92
Discrimination index of reliability= .99
Model, Fixed (all same) chi square=269.10 *df*=3 *p*= .00
Model, Random (normal) chi square=3.00 *df*=2 *p*= .22

The standardized forms of the residual values were examined in order to determine in which item type the most unexpected scores were given during the raters' evaluation of different item types. The analyses revealed that there were 51 outlier values: 11 of these (21.57%) belonged to the first rater, while 19 (37.25%) and 21 (41.18%) of them belonged to the second rater and the third rater, respectively. An examination of which item type outliers were more existent revealed that there were 6 (11.76%) outliers in the multiple choice items, 7 (13.73%) outliers in the open-ended items, 8 (15.69%) outliers in the true-false items and 30 (58.82%) outliers in the short response items. Accordingly, it can be claimed that raters showed the lowest agreement in their scorings of short response items where the highest ratio of outliers were observed. That is, short response items were the most affected by raters' expertise. Appendix 2 depicts the distribution of the outliers (standardized residual values) by item type. The common map obtained by converting each of the variable sources (each dimension) addressed within the scope of the study into logit values is displayed in Figure 1.

Figure 1 shows that teacher candidates, raters, and competency levels in relation to item types were converted to the same logit measure. This common measure allows for a comparability among all variability sources. It is depicted that the most successful teacher candidate was candidate number 59, while the least successful candidate was candidate number 28. Similarly, it can be observed that while rater 1 was the most lenient scorer, rater 3 was the strictest scorer. In addition, it can be observed that the competence level for preparing true-false items was found to be high, while the competence level for preparing multiple choice items was low.

Figure 1. *Logit map of the variables in the study*



4. DISCUSSION, CONCLUSION and SUGGESTIONS

The present study aimed to utilize the Rasch analysis to examine the competency levels of Turkish teacher candidates in writing test items. In addition, the effect of the expertise of the raters who scored the items developed by teacher candidates was examined within the scope of the study. There are studies in which the tasks of teachers and prospective teachers are evaluated with multi-faceted Rash analysis (Erguvan & Aksu Dünya, 2021; Goodwin, 2016; Li, 2022). Because Rash analysis, the multi-faceted Rasch model, contributes to the reliability and validity of the measurements in determining the expected effects of the variability within the scope of the research (e.g., the mutual interactions between the rater and the item type). When a multi-faceted Rasch bias analysis is performed, the researcher looks for evidence in the rater's scoring pattern (Myford & Wolfe, 2003).

The conclusions derived from the Rasch analysis run on the data obtained from the test items developed by 84 Turkish teacher candidates and the data obtained from the 3 raters are as follows:

One conclusion that was arrived at was that raters' expertise had an impact on teacher candidates' competency levels in writing different types of items. When the raters' expertise were examined, it was observed that the most reliable scoring belonged to the rater who was an expert in both Turkish education and measurement and evaluation. Then followed the rater who was an expert in Turkish education, who was also observed to score in a reliable way (though with a lower reliability score). The least reliable rater was found to be the rater with expertise solely in measurement and evaluation. Most of the studies in the literature are those where the effect of a higher number of raters is investigated (Atılgan & Tezbaşaran, 2005; Bıkmaz Bilgen & Doğan, 2017; Kamış & Doğan, 2017). In addition, in a study by Erman Aslanoğlu & Şata (2021), it was reported that raters with similar expertise were effective in scoring items, and in a study by Kara & Kelecioğlu (2015), it was revealed that raters' expertise were effective in scoring reliability such as determining the cut-off values. In the literature, it is seen that rater qualities are examined more in the process of evaluating language skills (Song et al., 2014). In the study by Leckie and Baird (2011), it was determined that inexperienced raters were more rigid than experienced raters in assessing students' language skills. Similarly, Meadows & Billington (2010) stated that experienced that raters make more consistent assessments than others. In the study conducted by Wiseman (2012), on the other hand, students had two types of compositions, narrative and persuasion, scored by eight raters. It was determined that the scorers' scores changed according to different composition types. This result indicates that rater qualifications effectively score and support the study's results. Institutions such as the Higher Education Council and the Ministry of National Education develop and administer numerous tests, primarily tests that serve as references for the school transitional system. Owing to issues of confidentiality, institutions are unable to administer pilot studies of the test they develop and, hence, solely base their test development process on expert opinions. The present study revealed that test items should be developed by raters that have expertise both in the related subject domain and in the area of measurement and evaluation. Alternatively, the findings of the study indicate that an expert on the subject domain and an expert on measurement evaluation should work together. As opposed to studies reporting that raters should have similar expertise, the present study revealed that raters with different areas of expertise score with higher reliability. The findings of the present study indicate that even though the rater who was an expert solely in the subject domain performed a higher level of reliable scoring than the rater who was an expert solely in measurement and evaluation, it is concluded that together they will produce results with a higher level of reliability. Hence, it is recommended that they do the scorings together. A person who completes measurement and evaluation graduate programs has expertise in this field. Although people who graduated from different undergraduate programs participate in graduate education because there is no undergraduate program, generally, those who graduated from the field of digital education do postgraduate education. The reason for this is the limited number of graduate programs in universities and the high placement scores of the applicants. For this reason, finding an assessment and evaluation specialist in all disciplines is difficult. The results of this research show how important the cooperation between the subject matter expert and the measurement and evaluation expert is, and it is necessary to work together in the test development and scoring process.

After the education which the Turkish teacher candidates received in relation to measurement and evaluation and the test development process, they developed a test consisting of different types of items. Subsequent to the analyses, it was revealed that the teacher candidates had the highest level of competence in true-false items and then followed short response, and open-ended items. The teacher candidates' lowest competence among the different types of items

was observed to be in writing multiple choice items. This finding is consistent with the literature in that writing multiple choice items is difficult. Among the different types of items, the True-False item type can be described as an item type where there is a single statement which needs to be identified as true or false. Open-ended items are more difficult than short response items because they are written to measure higher level skills. Although often still, multiple-choice tests form the backbone of most standardized and classroom tests for various reasons. The advantages of multiple-choice assessments over most free-response assessments include lower costs for scoring, higher reliability, broader sampling of content, and the ability to obtain a wide range of scores (Gierl, Bulut, & Zhang, 2017; Fuhrman, 2018). In this study, pre-service teachers formed multiple-choice items at understanding, application, and analysis levels. Similarly, open-ended items were prepared to measure high-level skills. In other words, it is seen that pre-service teachers have the most difficulty in formulating items to measure high-level skills. This result is consistent with the literature. Asim, Ekuri, & Eni (2013) also determined in their study that pre-service teachers struggled to write multiple-choice items to measure high-level skills. Haladayna, Downing, & Rodriguez (2002) drew attention to the difficulty of writing multiple-choice items for teachers and pre-service teachers in their study where they determined the principles of test development. Özçelik (2010) asserts that multiple choice items can only be written after a certain period of preparation and experience. According to Özçelik (2010a), one must first start by writing short response items and by doing so learn how to write multiple choice items. Preparing a test consisting of multiple choice items would require quite a long period of time because writing the items requires not only expertise in the subject domain but also certain knowledge and skills in measurement and evaluation (Tan, 2012). The findings obtained in the present study are consistent with those reported in the related literature. However, further studies are needed on teacher candidates' practice in writing particularly open-ended and multiple choice test items. Teachers state that they are not competition at the item writing. For this reason, pre-service teachers need to gain theoretical knowledge about measurement and evaluation processes and practice. The findings obtained in the present study are consistent with those reported in the related literature. However, further studies are needed on teacher candidates' practice in writing, particularly open-ended and multiple-choice test items. However, reducing the measurement and evaluation course to 2 hours per week in 2020 makes this situation difficult. For this reason, increasing the course hours or taking a separate course before the service for test development is recommended.

When the raters' expertise and the interaction between different types of items were examined, it was found that raters' expertise were mostly influential on scoring of short response items. In other words, variations among the raters' scores were mostly observed in the short response items. Short response items are those where students provide a number, word or a sentence as a response (Özçelik, 2010b), and since there are no options in the item and the student needs to provide his/her own response, subjectivity can be involved in scoring these items (Tekin, 2004). When the scoring criteria of short response items were examined, it could be observed that short response items had such expertise as having a single correct answer, being understood in the same way by different people, being clear and comprehensible, and matching the measured target learning outcome. While the rater with expertise in solely measurement and evaluation assigned a high score to a single response to an item developed, by for instance student no. 52, the rater with expertise in solely Turkish education assigned a low score. As previously mentioned, these findings indicate the importance of collobarative work in scoring by an expert on the subject domain and an expert on measurement and evaluation during the development of test items.

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**: Ağrı İbrahim Çeçen University, 01/12/2021, E-95531838-050.99-25942.

Authorship Contribution Statement

Ayfer Sayin: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing -original draft. **Mehmet Sata**: Methodology, Supervision, and Validation. Authors may edit this part based on their case.

Orcid

Ayfer Sayin https://orcid.org/0000-0003-1357-5674
Mehmet Sata https://orcid.org/0000-0003-2683-4997

REFERENCES

Anthony, C.J., Styck, K.M., Volpe, R.J., & Robert, C.R. (2022). Using many-facet rasch measurement and generalizability theory to explore rater effects for direct behavior rating–multi-item scales. *School Psychology. Advance online publication*. https://doi.org/10.1037/spq0000518

Asim, A.E., Ekuri, E.E., & Eni, E.I. (2013). A Diagnostic Study of Pre-Service Teachers' Competency in Multiple-Choice Item Development. *Research in Education, 89*(1), 13–22. https://doi.org/10.7227/RIE.89.1.2

Atılgan, H., & Tezbaşaran, A. (2005). Genellenebilirlik kuramı alternatif karar çalışmaları ile senaryolar ve gerçek durumlar için elde edilen g ve phi katsayılarının tutarlılığının incelenmesi. *Eğitim Araştırmaları, 18(1)*, 28-40.

Barkaoui, K. (2010). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly*, *44*(1), 31–57.

Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme*. ÖSYM Yayınları.

Bıkmaz Bilgen, Ö., & Doğan, N. (2017). Puanlayıcılar Arası Güvenirlik Belirleme Tekniklerinin Karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *8*(1), 63-78. https://doi.org/10.21031/epod.294847

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2018). *Eğitimde bilimsel araştırma yöntemleri*. Pegem Akademi. https://doi.org/10.14527/9789944919289

Crocker, L.M. & Algina, L. (2008). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, *33*(1), 117–135.

Erguvan, I.D. & Aksu Dünya, B. (2021). Gathering evidence on e-rubrics: Perspectives and many facet Rasch analysis of rating behavior. *International Journal of Assessment Tools in Education , 8*(2) , 454-474 . https://doi.org/10.21449/ijate.818151

Erman Aslanoğlu, A., & Şata, M. (2021). Examining the differential rater functioning in the process of assessing writing skills of middle school 7th grade students. *Participatory Educational Research (PER), 8*(4), 239-252. https://doi.org/10.17275/per.21.88.8.4

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. https://doi.org/10.37546/JALTJJ34.1-3

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. https://doi.org/10.4304/tpls.1.11.1531-1540

Fuhrman, M. (1996) Developing Good Multiple-Choice Tests and Test Items, *Journal of Geoscience Education, 44*(4), 379-384. https://doi.org/10.5408/1089-9995-44.4.379

Gierl, M.J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. *Review of Educational Research, 87*(6), 1082-1116. https://doi.org/10.3102/0034654317726529

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(1),* 21-31. https://doi.org/10.1016/j.asw.2016.07.004

Gorin, J.S. (2007). Reconsidering issues in validity theory. *Educational Researcher, 36*(8), 456-462. https://doi.org/10.3102/0013189X07311607

Haladyna, T.M., Downing, S.M., & Rodriguez, M.C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment, *Applied Measurement in Education, 15*(3), 309-333. https://doi.org/10.1207/S15324818AME1503_5

Jones, E., & Bergin, C. (2019) Evaluating Teacher Effectiveness Using Classroom Observations: A Rasch Analysis of the Rater Effects of Principals, *Educational Assessment, 24*(2), 91-118. https://doi.org/10.1080/10627197.2018.1564272

Kamış, Ö. & Doğan, C.D. (2017). How consistent are decision studies in G theory?. *Gazi University Journal of Gazi Educational Faculty, 37*(2), 591-610.

Kara, Y., & Kelecioğlu, H. (2015). Puanlayıcı Niteliklerinin Kesme Puanlarının Belirlenmesine Etkisinin Genellenebilirlik Kuramı'yla İncelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 6*(1), 58-71. https://doi.org/10.21031/epod.47997

Karasar, N. (2018). *Bilimsel araştırma yöntemi* (33th ed.). Ankara: Nobel Yayıncılık.

Kim, H. (2020). Kim, H. Effects of rating criteria order on the halo effect in L2 writing assessment: a many-facet Rasch measurement analysis. *Lang Test Asia 10*(16), 1-23, https://doi.org/10.1186/s40468-020-00115-0

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

Li, W. (2022). Scoring rubric reliability and internal validity in rater-mediated EFL writing assessment: Insights from many-facet Rasch measurement. *Read Writ*. https://doi.org/10.1007/s11145-022-10279-1

Linacre, J.M. (1993). Rasch-based generalizability theory. *Rasch Measurement Transaction*, 7(1), 283-284.

Linacre, J.M. (2012). *FACETS* (Version 3.70.1) [Computer Software]. MESA Press.

Linacre, J.M. (2017). *FACETS* (Version 3.80.0) [Computer Software]. MESA Press.

Linn, R.L., & Grolund, N.E. (2000). *Measurement and assessment in teaching* (8th ed.). Merrill/Prentice Hall.

Marais, I., & Andrich, D. (2008). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *J Appl Meas*, *9*(3), 200-215.

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

Meadows, M., & Billington, L. (2010). *The effect of marker background and training on the quality of marking in GCSE English*. AQA Education.

Milli Eğitim Bakanlığı (2019). *Türkçe Dersi Öğretim Programı (İlkokul ve Ortaokul 1, 2, 3, 4, 5, 6, 7 ve 8. Sınıflar)*. MEB Yayınları.

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.

Osburn, H.G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological methods*, *5*(3), 343-355.

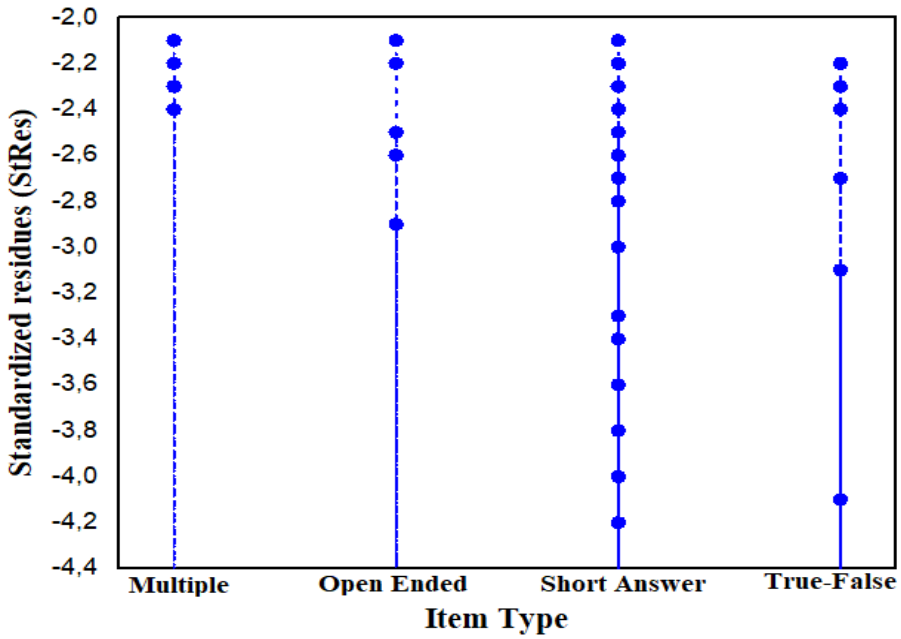
Özçelik, D.A. (2010a). *Ölçme ve değerlendirme.* Pegem Akademi.
Özçelik, D.A. (2010b). *Test geliştirme kılavuzu*. Pegem Akademi.
Primi, R., Silvia, P.J., Jauk, E., & Benedek, M. (2019). Applying many-facet Rasch modeling in the assessment of creativity. *Psychology of Aesthetics, Creativity, and the Arts, 13*(2), 176–186. https://doi.org/10.1037/aca0000230
Sayın, A., & Kahraman, N. (2020). A measurement tool for repeated measurement of assessment of university students' writing skill: development and evaluation. *Journal of Measurement and Evaluation in Education and Psychology*, *11*(2), 113-130. https://doi.org/10.21031/epod.639148
Sayın, A., & Takıl, N.B. (2017). Opinions of the Turkish teacher candidates for change in the reading skills of the students in the 15 year old group. *International Journal of Language Academy*, *5*(2), 266-284. http://dx.doi.org/10.18033/ijla.3561
Sireci, S.G. (2007). On validity theory and test validition. *Educational Researcher*, *36*(8), 477-481. https://doi.org/10.3102/0013189X07311609
Song, T., Wolfe, E.W., Hahn, L., Less-Petersen, M., Sanders, R., & Vickers, D. (2014). *Relationship between rater background and rater performance*. Pearson.
Tan, Ş. (2012). *Öğretimde ölçme ve değerlendirme KPSS el kitabı*. Ankara: Pegem Akademi.
Tekin, H. (2004). *Eğitimde ölçme ve değerlendirme*. Yargı Yayınevi.
Walsh, W.B., & Betz, N.E. (1995). *Tests and assessment*. Prentice-Hall, Inc.
Wiseman, C.S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing*, *17*(3), 150-173. https://doi.org/10.1016/j.asw.2011.12.001

APPENDIX

Appendix 1. Rubric

Item Type	Criteria	Score
True/False	• Text selection (originality, suitability for student level, language, expression, etc.) • Compliance with the principles of item writing (not containing only absolutely, etc. expressions, having only one correct answer, not giving clues, not being used one-to-one in the text, etc.)	5
Multiple-choice	• Text selection (originality, suitability for student level, language, expression, etc.) • Compliance with the principles of item writing (having only one correct line, the structure of the options, appropriateness of the item root, etc.)	5
Short-answered	• Text selection (originality, suitability for student level, language, expression, etc.) • Compliance with the principles of item writing (having only one correct answer, not giving clues, not being one-to-one in the text, limited response, etc.)	5
Open-ended	• Text selection (originality, suitability for student level, language, expression, etc.) • Compliance with the principles of item writing (suitability for measuring high-level mental skills, the correctness of the answer key, etc.)	5

Appendix 2. *The distribution of standardized residual values by item type*



International Journal of Assessment Tools in Education
2022, Vol. 9, No. 4, 1013–1029
https://doi.org/10.21449/ijate.1164590
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Investigation of the effect of parameter estimation and classification accuracy in mixture IRT models under different conditions

Fatima Munevver Saatcioglu[1,*], **Hakan Yavuz Atar**[2]

[1]Ankara Yildirim Beyazit University, Rectorate, Ankara, Turkiye
[2]Gazi University, Faculty of Education, Department of Educational Sciences, Ankara, Turkiye

ARTICLE HISTORY
Received: Aug. 19, 2022
Revised: Dec. 07, 2022
Accepted: Dec. 18, 2022

Keywords:
Mixture Item Response Theory Models,
Maximum Likelihood Estimation,
Item Parameter Recovery,
Classification Accuracy,
Missing Data,
Latent Class.

Abstract: This study aims to examine the effects of mixture item response theory (IRT) models on item parameter estimation and classification accuracy under different conditions. The manipulated variables of the simulation study are set as mixture IRT models (Rasch, 2PL, 3PL); sample size (600, 1000); the number of items (10, 30); the number of latent classes (2, 3); missing data type (complete, missing at random (MAR) and missing not at random (MNAR)), and the percentage of missing data (10%, 20%). Data were generated for each of the three mixture IRT models using the code written in R program. *MplusAutomation* package, which provides the automation of R and Mplus program, was used to analyze the data. The mean RMSE values for item difficulty, item discrimination, and guessing parameter estimation were determined. The mean RMSE values as to the Mixture Rasch model were found to be lower than those of the Mixture 2PL and Mixture 3PL models. Percentages of classification accuracy were also computed. It was noted that the Mixture Rasch model with 30 items, 2 classes, 1000 sample size, and complete data conditions had the highest classification accuracy percentage. Additionally, a factorial ANOVA was used to evaluate each factor's main effects and interaction effects.

1. INTRODUCTION

Tests are widely used in different contexts such as education, psychology, industry, and health. In educational and psychological fields, test results are preferred for various purposes such as selecting individuals, following their development, or evaluating the efficiency of education systems. A growing awareness of the importance and the impact of testing has led to designing better tests and developing statistical methods used for the analysis of test scores. Item Response Theory (IRT) models are among the most commonly used models in various testing settings. Although IRT models have many advantages, they have strict assumptions such as unidimensionality, homogeneity population, local independence, and the invariance of item parameters (Embretson & Reise, 2000; Hambleton et al., 1991). The advantages of IRT models depend on the validity of the model whose assumptions are to be met. Traditional IRT models assume that data are drawn from a single homogeneous population. However, it may not always be possible because population may include two or more subpopulations that consist of different

*CONTACT: F. Munevver Saatcioglu fmyigiter@gmail.com , Ankara Yildirim Beyazit University, Rectorate, Ankara, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

latent classes. Mixture IRT models assume that the overall population includes multiple latent classes that can be identified based on the item response patterns (Rost, 1990). In this case, the mixture IRT modeling approach is used. In social science research, there have been many studies that use mixture IRT models (Alexeev et al., 2011; Cohen et al., 2005; De Ayala & Santiago, 2017; Finch & French, 2012; Maij-de Meij et al., 2008; Lee, 2012; Oliveri et al., 2014; Sen, 2016; Zhang et al., 2015). The three-parameter Mixture IRT model including item parameters and the guessing parameter for each class is shown as the following equation:

$$P(x_{ij}=1|\theta_j)=P_{ij}=\sum_{g=1}^{G}\pi_g\left(\Upsilon_{ig}+(1-\Upsilon_{ig})\frac{exp[\alpha_{ig}(\theta_{jg}-\beta_{ig})]}{1+exp[\alpha_{ig}(\theta_{jg}-\beta_{ig})]}\right) \quad (1)$$

In equation (1), $g=(1,2,..,G)$ indicates latent class membership, (β_{ig}), (α_{ig}), and (Υ_{ig}) represent the difficulty, discrimination, and guessing parameters, respectively for item *i*, (θ_{jg}) denotes the ability parameter for individual *j* in class g , and π_g indicates the mixing proportion of individuals in a class. The probability that each individual belongs to one latent class and the mixing proportion of individuals in each class is estimated with the (π_g), $\sum_{i=1}\pi_g=1$ and 0 $\leq\pi_g\leq1$ restriction (Rost, 1990). When the guessing parameter is equal to zero, the two-parameter mixture IRT model; with the assumption that the guessing parameter is equal to zero and the item discrimination parameter is equal to 1, the Mixture Rasch model can be obtained.

Much of the current research has focused on the Rasch and 2PL version of mixture IRT models, while there is a relatively small body of literature on the Mixture 3PL model (Cho, Cohen & Kim, 2013; Choi et al., 2020; Li et al., 2009).

When the Mixture IRT literature is examined, the sample size, the number of items, and the number of latent classes appear to affect parameter estimates of the Mixture IRT models. For example, Preinerstorfer and Formann (2012) indicated that increasing the sample size (500, 1000, 2500) and the number of items (10, 15, 25, 40) leads to higher accuracy in estimating the parameters of the mixture Rasch model. Moreover, Li et al. (2009) found that recovery of item parameters in mixture models such as the one-parameter logistic (1PL), the two-parameter logistic (2PL), and the three-parameter logistic (3PL) differed based on the sample sizes (600, 1200); the number of latent classes (1, 2, 3, 4); and the number of items (6, 15, 30). When the number of latent classes increased, the mean root mean square error (RMSE) values increased for item difficulty and disrimination parameters. Also, according to the study of Li et. al (2009), the mean RMSE values decreased as the sample size and the number of items increased. The classification accuracy increased with an increasing number of items. Different sets of sample size, number of items, and number of classes that have been used in the mixture IRT models in previous studies can be seen in the review study by Sen and Cohen (2019). The present study focuses specifically on examining the effects of factors on the estimation of item parameters and classification accuracy for mixture IRT models including 1PL, 2PL and 3PL.

Also, it is suggested that the data set should be examined in terms of missing data so that the latent variables which the tests aim to measure can be obtained (Little & Rubin, 1987). Missing data in the response patterns cause negative situations such as bias, higher standard errors in parameter estimations, and lower power of a test (De Ayala et al., 2001; Finch, 2008; Hohensinn & Kubinger, 2011; Pohl et al., 2014). At this point, it would be beneficial to determine the percentage of missing data and the mechanism of the missing data type before analyzing the data. Also, there is no study with missing data and 3PL mixture IRT models in the literature. In the context of the findings to be obtained from this study, it is therefore thought that the research is important in terms of making extensive and detailed comments on the error values and

classification accuracy obtained as a result of the mixture 3PL model and the missing data type and missing data percentage factors.

Another significance of the research is examining the RMSE and bias values of the parameter estimations obtained from the mixture IRT models, which is important in terms of evaluating the performance of the mixture IRT models in different conditions and determining which model has less errors in the determined conditions. The findings to be obtained in this direction are considered important in terms of providing information and guiding the practitioners in terms of which model would be appropriate to choose according to their own conditions in their studies.

In line with these purposes, this study tries to answer the following questions:

1) How do the mean RMSE values obtained through parameter estimations change based on the sample size, the number of items, the estimation model, the number of classes, the percentage of missing data and the missing data type factors?

2) How does the interaction effect of the variables considered change according to the mean RMSE values obtained as a result of parameter estimations?

3) How does the classification accuracy obtained from the combination of the factors change?

2. METHOD

In this study, the factors for simulation conditions were designed to investigate the effects of the model, number of latent classes, number of items, sample size, model missing data type and missing data rate on the estimates of mixture IRT model parameters and classification accuracy. The simulation conditions for this study are as follows: three Mixture IRT models (Rasch, 2PL, 3PL); number of latent classes (2, 3); number of items (10, 30); sample sizes (600, 1000); missing data mechanisms ((complete data, missing at random (MAR), missing at not random (MNAR), and missing data percentages (10%, 20%). Overall, 144 conditions were simulated in this study. One hundred replications were generated for each condition. All data sets were analyzed for each of the mixture IRT models with the computer program Mplus version 8.5 (Muthe´n & Muthe´n, 1998-2020).

2.1. Simulation Conditions

2.1.1. *Number of classes*

The examinees have different response patterns on items and according to these different patterns, they are assigned to different latent classes. This situation enables estimating group-specific parameters for latent classes in mixture IRT models. According to the study conducted by Sen and Cohen (2019), the number of latent classes used in the studies ranges from one to ten. However, according to the results of the model-data fit studies, it is stated that the data generally fit the mixture IRT model with two or three latent classes (Finch & French, 2012; Park et al., 2016). Therefore, in this study, the conditions for the number of classes were determined as two and three to identify poor, average, and good performing individuals (Li et al., 2009).

2.1.2. *Number of items*

The number of items has been one of the manipulated variables in various simulation studies in the existing literature. The study conducted by Sen and Cohen (2019) shows that the number of items used in previous studies varies between 4 and 470 (Cho et al., 2012; Jilke et al., 2015). In this research, item numbers were taken as 10 and 30 as reported in Lee (2012) to generate different profile of latent classes (poor, average and good performing) according to item parameter values.

2.1.3. *Distribution of item and ability parameters*

Data were generated for each mixture IRT model (i.e., Rasch, 2PL, 3PL) using R program (R Core Team, 2020). The distributions of ability and item parameters were generated to be the same for each model. Then, class-specific item parameters were generated for each model and item parameter values for the classes were obtained (see Table 1). Item difficulty parameter values ranged from -2.7 to +2.7 for the 10-item condition, and for the 30-item condition, they were randomly generated based on a uniform distribution in the range of -3 to +3. Guess parameters were generated for the 0.25, 0.2, and 0.1 corresponding to easy items, medium difficulty items, and difficult items, respectively (Li et al., 2009).

Item difficulty parameter values were written in the Mplus input file as the first threshold and guessing parameter values as the second threshold (Muthén & Muthén, 1998-2021). Similar to the study of Li et al. (2009), item discrimination parameters were set as 1 for the poor and average performing classs and 2 for the good performing class. Ability parameters were obtained from the standard normal distribution (N(0,1)) and randomly generated with the runif function. In Table 1, the item parameter values generated for 10 items in the Mixture IRT models are given.

Table 1. *Item parameter values generated for the 10 items in Mixture IRT models.*

	Class1			Class2			Class1			Class2			Class3		
Item	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c
1	2	-2.7	0.10	1	2.7	0.25	2	-2.7	0.10	1	-0.5	0.20	1	2.7	0.25
2	2	-2.1	0.10	1	2.1	0.25	2	-2.1	0.10	1	-0.4	0.20	1	2.1	0.25
3	2	-1.5	0.10	1	1.5	0.25	2	-1.5	0.10	1	-0.3	0.20	1	1.5	0.25
4	2	-0.9	0.10	1	0.9	0.25	2	-0.9	0.10	1	-0.2	0.20	1	0.9	0.25
5	2	-0.3	0.20	1	0.3	0.20	2	-0.3	0.20	1	-0.1	0.20	1	0.3	0.20
6	1	0.3	0.20	2	-0.3	0.20	1	0.3	0.20	1	0.1	0.20	2	-0.3	0.20
7	1	0.9	0.25	2	-0.9	0.10	1	0.9	0.25	1	0.2	0.20	2	-0.9	0.10
8	1	1.5	0.25	2	-1.5	0.10	1	1.5	0.25	1	0.3	0.20	2	-1.5	0.10
9	1	2.1	0.25	2	-2.1	0.10	1	2.1	0.25	1	0.4	0.20	2	-2.1	0.10
10	1	2.7	0.25	2	-2.7	0.10	1	2.7	0.25	1	0.5	0.20	2	-2.7	0.10

In Table 1, item parameter values generated according to all class numbers are presented for cases where the number of latent classes is two and three. For the two-class case, arranging the item difficulty parameters from easy to difficult in Class 1 means that the individuals in Class 1 produced a poorer performance when answering the items correctly, whereas arranging the item difficulty parameters from difficult to easy in Class 2 means that the individuals in Class 2 performed better when answering the items correctly. In both classes, item discrimination and guessing parameters were found to be compatible with item difficulty values. For the three-class case the item difficulty parameters in Class 2 are of medium difficulty, which means that the individuals in Class 2 produced an average performance in answering the items correctly. In all three classes, item discrimination and guessing parameters were found to be compatible with item difficulty values.

2.1.4. *Sample Size*

In previous simulation studies, sample sizes larger than 500 were selected (Lee et al., 2021; Li et al., 2009) for mixture models in simulation studies. More specifically, Li et al. (2009) reported that a sample size of 600 would be appropriate when the number of items is between 15 and 30 for the Mixture Rasch models; they also suggested that a sample size of 600 would be sufficient for a model with 1 to 4 classes for both Mixture 2PL and Mixture 3PL models for a 15-item test. Cho et al. (2013) suggested that a sample size larger than 360 can be used for the Mixture

Rasch model. Cohen and Bolt (2005) successfully applied the Mixture 3PL model with a sample size of 1000. Considering these, the sample size of the study was determined as 600 and 1000.

2.1.5. *Missing data*

Rubin (1976) classified missing data as completely at random (MCAR) and missing at random (MAR) and these missing data mechanisms have no systematic cause if they are ignored; that is, the missing data is a simple random sample of the observed data. However, if the missing pattern is missing not at random (MNAR), in this case ignoring nonignorable missing responses leads to biased parameter estimates (Little & Rubin, 1987).

In the scope of this study, MAR and MNAR data generation was based on the study of Finch (2008): for a 10-item data set, 3 most difficult items were set as target items. A total score was calculated for the remaining 7 items. Based on the total scores excluding the target items, the simulations were divided into four fractiles (0-1, 2-3, 4-5, 6-7) for each class. Four fractiles were created with four different values of the missing response probabilities on the target items. The mean of these probabilities for the fractiles was designed to be equal to the total percentages of missing responses, namely 10% and 20%. Generating missing data through this way, response patterns were formed for poor, average, and good performing simulatives based on the total scores of the items excluding the target items.

2.2. Estimation

Parameters for mixture IRT models can be estimated by Bayesian estimation with Markov chain Monte Carlo (MCMC) algorithms or maximum likelihood estimation (MLE) techniques. There are some differences in the way these two techniques are implemented. Edwards and Finch (2018) stated that the Full Information Maximum Likelihood (FIML) method produced better results in their study where they examined the parameter estimations for MAR and MNAR cases by considering the 2PL and 3PL IRT models. As the name suggests, FIML method estimates model parameters using a maximum likelihood fitting function with all the data available. Thus, individuals with missing data are included in the parameter estimation process with all the information related to them, and these are ignored for variables with missing values. In addition, FIML does not involve the assignment of missing values, thus making the use of this method less cumbersome than some of the other proposed approaches, especially those that rely on data assignment. Finally, FIML is available in most statistical software, which, in practical terms, makes it very easy to use.

2.3. Analysis

The data were analyzed with the *MplusAutomation* package, which can integrate between the Mplus program and the R program (Hallquist & Wiley, 2018). Input files to be used for 100 replications and output files obtained were also produced with the *MplusAutomation* package and analyzed. In this simulation study, the performance of Mixture IRT models was evaluated on the basis of two criteria: Item parameter recovery and classification accuracy.

2.3.1. *Item parameter recovery*

In this study, root mean square error (RMSE) values were used to assess the accuracy of item parameter estimates, calculated with the help of the following equation by using the item number, the number of classes, and the number of replications for the estimated item difficulty parameter values:

$$RMSE(\beta_i) = \sqrt{\frac{\sum_{r=1}^{R}\sum_{i=1}^{I}\sum_{g=1}^{C}\left(\hat{\beta}_{igr} - \beta_{ig}\right)^2}{RIC}}$$

In this equation, $\hat{\beta}_{igr}$ represents the estimated item difficulty parameter obtained from R replication for item i in class g, β_{ig} represents the true value of item parameter for item i in class g, R denotes the number of replications, I indicates the number of items, and C denotes the number of classes. Equation 1 was also used for the assessment of item discrimination and item guessing parameter estimates. Before calculating the RMSE for a given replication, parameter estimates were first transformed to the scale of the generating values with mean equating (Kolen & Brennan, 2004). The parameter estimates are exactly the same as the true value when RMSE equals zero. Lower values (e.g., <0.10) indicate better fit.

2.3.2. *Effect size*

The effect size is defined as the variance ratio describing each main effect, relationship, and error in the ANOVA design and takes a value between 0.00 and 1.00 (Cohen, 1988). Eta-square, which does not require the assumption of linearity between the variables, shows how effective the independent variable is on the dependent variable. According to Cohen (1988), 0.01 for the small effect size value; 0.06 for the medium effect size value; and 0.14 for high effect size value are recommended as lower limit values. In the presence of more than one estimator, partial eta-squared measures the proportion of the total variance explained by a given estimator, after keeping the variance explained by other estimators constant. It is recommended to use partial eta-square to determine interaction effects in multi-way or factorial ANOVA designs (Richardson, 2011; Norouzian & Plonsky, 2018). In this study, the mean RMSE values obtained from the estimated item parameters were taken as the dependent variable and the factors were also taken as independent variable. Main and interaction effects were interpreted with eta-squared values in line with the values suggested by Cohen (1988).

2.3.3. *Classification accuracy*

Within the data sets produced for classification accuracy, there is a posterior probability for each person in each latent class based on person's response pattern. Each person in the latent class was assigned to a latent class according to their highest posterior probability values, saved in the Mplus output and these values were extracted with the MplusAutomation package. For a data set with 1000 examinees, classification accuracy value was calculated as 0.92, which means there is a matched assignment for 920 of the 1000 cases.

2.3.4. *Label switching*

Since there is no information about the number and nature of estimated classes in mixture IRT models, sometimes the parameters estimated for Class 1 can be labeled as Class 2. In such cases, the problem of label switching can be overcome by taking the estimated item parameter values as starting values in Mplus syntax (Kutscher et al. 2019).

3. RESULTS

3.1. Item Parameter Recovery Results

3.1.1. *Item difficulty parameter*

The mean RMSE values of the estimated item difficulty parameters for the mixture models are presented in Table 2. The codes in this table for simulation conditions are designed to represent the combination of factors for a given situation. To specify the simulation conditions, codes with 10-13 digits were created. The first two characters of the codes denote class number (2C, 3C); the following three characters refer to missing data percentage (10P, 20P); the next grouping indicates sample size (600, 1000), and the last two characters represent the number of items (10,30). For example, in the 2C10P60010 codes the number of classes is denoted by 2C, the percentage of missing data by 10P, the sample size by 600, and the number of items by 10.

Table 2. *The mean RMSE values of the estimated item difficulty parameters for the Mixture models.*

	Mixture Rasch			Mixture 2PLM			Mixture 3PLM		
Conditions	COMP	MAR	MNAR	COMP	MAR	MNAR	COMP	MAR	MNAR
2C10P60010	0.045	0.052	0.075	0.041	0.065	0.089	0.367	0.560	0.373
2C10P60030	0.025	0.026	0.054	0.025	0.039	0.077	0.112	0.225	0.303
2C10P100010	0.035	0.043	0.067	0.032	0.038	0.070	0.219	0.265	0.263
2C10P100030	0.021	0.022	0.050	0.021	0.022	0.052	0.090	0.168	0.352
2C20P60010	0.046	0.083	0.082	0.057	0.121	0.241	0.348	0.654	0.592
2C20P60030	0.025	0.028	0.251	0.037	0.061	0.070	0.130	0.213	0.222
2C20P100010	0.035	0.071	0.540	0.034	0.233	0.177	0.219	0.586	0.520
2C20P100030	0.023	0.024	0.188	0.023	0.024	0.919	0.078	0.124	0.217
3C10P60010	0.139	1.160	0.236	1.551	1.775	1.950	1.386	1.898	2.831
3C10P60030	0.105	0.435	0.170	0.070	0.032	0.064	0.210	0.221	0.257
3C10P100010	0.102	0.206	0.125	0.973	1.256	1.778	1.075	1.704	2.636
3C10P100030	0.082	0.069	0.036	0.051	0.096	0.075	0.160	0.196	0.195
3C20P60010	0.148	1.350	1.690	1.761	1.994	2.570	1.160	4.471	2.357
3C20P60030	0.081	0.397	0.191	0.279	1.436	1.709	0.246	0.292	0.283
3C20P100010	0.090	0.884	0.383	1.641	1.832	2.237	2.652	2.673	2.341
3C20P100030	0.024	0.058	0.088	0.226	0.439	0.113	0.141	0.253	0.212

Table 2 shows that the mean RMSE values of the item difficulty parameters obtained for the Mixture Rasch model decreased as the number of items and the number of classes increased. As can be seen in Table 2, in the complete data, the mean RMSE values decreased as the number of items and sample size increased, and the mean RMSE values increased as the number of classes and the percentage of missing data increased. In MAR and MNAR data conditions, the mean RMSE values generally decreased as the number of items and sample size increased, and the mean RMSE values generally increased as the number of classes and the percentage of missing data increased. It can also be seen that item difficulty parameter values had the highest mean RMSE values in complete, MAR, and MNAR data with 3 class, 20% missing data percentage, 600 sample size, and 10 item (3C20P60010) condition. The lowest mean RMSE value was observed in complete data, 2 class, 10% missing data, 1000 sample size, and 30 item (2C10P100030) condition.

In Table 2, it can be seen that the mean RMSE values of the item difficulty parameters obtained for the Mixture 2PL model were higher than the mean RMSE values of the item difficulty parameters obtained for the Mixture Rasch model. Also, the mean RMSE values decreased as the number of items and sample size increased, and the mean RMSE values increased as the number of classes and the percentage of missing data increased. When the mean RMSE values were examined according to the missing data types, higher RMSE values were obtained for the MNAR condition. The item difficulty parameter values for the mixture 2PL model were obtained with the highest RMSE values, while the MAR and MNAR data with 3 class, 20% missing data percentage, 600 sample size, and 10 item (3C20P60010) condition. The lowest mean RMSE value was observed in the complete data with 2 class, 10% missing data, 1000 sample size, and 30 item (2C10P100030) condition.

As shown in Table 2, the mean RMSE values of the item difficulty parameters obtained for the Mixture 3PL model were higher than those for the Mixture 2PL model. Also, the mean RMSE values increased as the complexity of the model increased (i.e from Rasch to 3PL model). The mean RMSE values decreased as the number of items and sample size increased, and the mean RMSE values increased as the number of classes and the percentage of missing data increased.

When the mean RMSE values were examined according to the missing data types, higher RMSE values were obtained for the MAR data condition. Table 2 shows that the highest mean RMSE value of the item difficulty parameter values for the mixture 3PL model was MAR data, with 3 class, 20% missing data percentage, 10 item, and a sample size of 600 (3C20P60010) condition and also the lowest RMSE value was seen with complete data, 2 class, 10% missing data, 30 item, and a sample size of 1000 (2C10P100030) condition.

3.1.2. *Item discrimination parameter*

The mean RMSE values of item discrimination parameter values for the Mixture 2PL and 3PL model are given in Table 3:

Table 3. *The mean RMSE values of the estimated item discrimination parameters for the Mixture 2PL and 3PL models.*

	Mixture 2PLM			Mixture 3PLM		
Conditions	COMP	MAR	MNAR	COMP	MAR	MNAR
2C10P60010	0.165	0.212	0.497	0.331	0.645	0.634
2C10P60030	0.179	0.310	0.277	0.161	0.234	0.171
2C10P100010	0.057	0.061	0.063	0.264	0.320	0.216
2C10P100030	0.024	0.038	0.052	0.094	0.115	0.120
2C20P60010	0.234	0.261	0.563	0.371	0.654	0.662
2C20P60030	0.259	0.256	0.523	0.192	0.229	0.262
2C20P100010	0.183	0.197	0.208	0.337	0.405	0.417
2C20P100030	0.032	0.041	0.055	0.101	0.151	0.176
3C10P60010	0.843	0.937	1.222	1.140	1.293	1.337
3C10P60030	0.725	0.873	0.916	0.776	0.821	0.857
3C10P100010	0.866	0.910	1.469	0.988	1.113	1.228
3C10P100030	0.675	0.784	0.833	0.581	0.696	0.705
3C20P60010	0.975	1.277	1.366	1.262	1.463	1.472
3C20P60030	0.837	0.927	0.982	0.920	0.943	0.952
3C20P100010	0.922	0.981	1.032	1.023	1.242	1.281
3C20P100030	0.786	0.854	0.967	0.723	0.817	0.832

As shown in Table 3, the mean RMSE values of the item discrimination parameter estimations for the complete data condition were lower, slightly higher for the MAR condition, and at the highest for the MNAR condition. The lowest RMSE values were obtained for complete, MAR and MNAR data with for 2 class, 10% missing data, 1000 sample size, and 30 item (2C10P100030) condition, while the highest RMSE value was obtained for the MNAR data with 3 class, 10% missing data percentage, 1000 sample size, and 10 item (3C10P100010) condition. It seems to be consistent with the conditions where the highest RMSE values were obtained for item discrimination parameter estimations and the highest mean RMSE values for item difficulty parameter estimations. For the mixture 3PL model, the mean RMSE values were lower for the complete data case of item discrimination parameter estimations, but higher for the MAR and MNAR conditions. The lowest RMSE values were obtained for complete, MAR, and MNAR data with 2 class, 10% missing data, 1000 sample size, and 30 item (2C10P100030) condition, while the highest RMSE value was obtained for MAR and MNAR data with data with 3 class, 20% missing data percentage, 600 sample size, and 10 item (3C10P100010) condition. It can be stated that these results and the conditions in which the highest RMSE values were obtained for item discrimination and item difficulty parameter estimation values in the Mixture 2PL model were similar.

3.1.3. *Guessing parameter*

Table 4 provides the mean RMSE values obtained for the guessing parameter values for mixture 3PL model.

Table 4. *The Mean RMSE values of the estimated guessing parameters.*

Conditions	COMP	MAR	MNAR
2C10P60010	0.076	0.076	0.076
2C10P60030	0.046	0.046	0.046
2C10P100010	0.077	0.077	0.078
2C10P100030	0.046	0.046	0.046
2C20P60010	0.076	0.076	0.077
2C20P60030	0.046	0.046	0.046
2C20P100010	0.078	0.078	0.079
2C20P100030	0.046	0.046	0.046
3C10P60010	0.059	0.059	0.060
3C10P60030	0.038	0.038	0.039
3C10P100010	0.061	0.058	0.061
3C10P100030	0.038	0.039	0.039
3C20P60010	0.059	0.059	0.061
3C20P60030	0.039	0.038	0.039
3C20P100010	0.061	0.061	0.061
3C20P100030	0.039	0.039	0.039

It can be seen in Table 4 that when the number of items for the guessing parameter increased, the mean RMSE values decreased as well. Also, the mean RMSE values for guessing parameters had lower values than the mean RMSE values obtained for item difficulty and discrimination parameters. The reason for this could be that when the guessing parameter values are between zero and one, item discrimination and difficulty parameter values can take larger absolute values.

3.2. A Linear Model Analysis of Simulation Results

Effects of each condition were evaluated using a factorial ANOVA for the RMSE values. The results related to partial eta-squared, degree of freedom (*df*), sum of squares (SS), mean square (MS), and F-values from the factorial ANOVA are presented in the following sections.

3.2.1. *ANOVA Results for item difficulty parameter*

In Table 5, main effects, two-way and three-way interactions for each factor are shown for item difficulty parameter. As can be seen in Table 5, all factors had a significant effect on item parameter estimation. According to partial eta-squared values, number of items (i), number of classes (C), and model (M) were the most influential factors on RMSE for item difficulty parameter. Missing data type and missing data percentage had also a large effect on the results. The least influential factor was the sample size (N).

The interaction effects between factors shown in Table 5 indicate that type and class (txC), type and percentage (txP), item and class (ixC), item and model (ixM), sample and class (NxC), class and model (CxM), and percentage and model (PxC) affected the RMSE values. Based on partial eta-squared values, it can be seen that two-way interactions had a large effect on the results. Also, significant three-way interactions are given in Table 5 and it can be seen that type, item and class (txixC), type, class and model (txCXM), item, class and model (ixCxM), and

sample, class and model (NxCxM) had significant interaction effects. These results suggest that interactions of factors may affect model parameter estimates.

Table 5. *ANOVA results for main effects and interaction effects of simulation conditions for item difficulty parameter.*

Factor	η_p^2	*df*	Sum of Squares	Mean Square	F
t	0.832	2	3.881	1.940	54.318*
i	0.922	1	9.313	9.313	260.685*
N	0.540	1	0.922	0.922	25.798*
C	0.951	1	15.340	15.340	429.413*
P	0.702	1	1.854	1.854	51.902*
M	0.921	2	9.126	4.563	127.724*
txi	0.085	2	0.073	0.037	1.024
txN	0.039	2	0.032	0.016	0.444
txC	0.663	2	1.543	0.771	21.593*
txP	0.426	2	0.582	0.291	8.151*
txM	0.324	4	0.377	0.094	2.635
ixN	0.109	1	0.096	0.096	2.690
ixC	0.767	1	2.581	2.581	72.259*
ixP	0.005	1	0.004	0.004	0.106
ixM	0.489	2	0.752	0.376	10.531*
NxC	0.173	1	0.164	0.164	4.591*
NXP	0.016	1	0.013	0.013	0.360
NXM	0.077	2	0.066	0.033	0.920
CXP	0.007	1	0.006	0.006	0.165
CXM	0.687	2	1.723	0.862	24.119*
PXM	0.459	2	0.667	0.333	9.334*
txixC	0.440	2	0.617	0.309	8.637*
txCxM	0.548	4	0.954	0.239	6.676*
ixCXM	0.439	2	0.615	0.307	8.605*
NXCXM	0.268	2	0.288	0.144	4.025*
Error		22	0.786	0.036	

Note. *t* = missing data type, i = number of items, N = sample, C =number of classes, P=missing data percentage, M = model.
*$p<.05$.

3.2.2. *ANOVA Results for item discrimination parameter*

In Table 6, main effects, two-way and three-way interactions for each factor are shown for item disrimination parameter. As can be seen in Table 6, according to partial eta-squared values, all factors had a large effect size values on the results. Mean RMSE values for item discrimination parameter were also significantly affected by two-way interactions including type and item (txi), type and class (txC), item and class (ixC), type and model (txM), item and class (ixC), item and model (ixM), and sample and class (PxC). Based on partial eta-squared values, these interactions had a large effect on the results. Also, three-way inreactions type, sample and class (txNxC), and sample, class and model (NxCXM) affected mean RMSE values for item discrimination parameter. These results suggest that interactions of factors may affect model parameter estimates.

Table 6. *ANOVA results for main effects and interaction effects of simulation conditions for item discrimination parameter.*

Factor	η_p^2	*df*	Sum of Squares	Mean Square	F
t	0.905	2	0.445	0.223	57.014*
i	0.969	1	1.442	1.442	369.207*
N	0.923	1	0.565	0.565	144.609*
C	0.996	1	13.065	13.065	3345.500*
P	0.760	1	0.149	0.149	38.052*
M	0.782	1	0.168	0.168	43.128*
txi	0.593	2	0.068	0.034	8.756*
txN	0.379	2	0.029	0.014	3.658
txC	0.428	2	0.035	0.018	4.483*
txP	0.012	2	0.001	0.000	0.072
txM	0.558	2	0.059	0.030	7.570*
ixN	0.099	1	0.005	0.005	1.318
ixC	0.777	1	0.163	0.163	41.723*
ixP	0.007	1	0.000	0.000	0.084
ixM	0.856	1	0.279	0.279	71.379*
NxC	0.404	1	0.032	0.032	8.141*
NXP	0.123	1	0.007	0.007	1.686
NXM	0.021	1	0.001	0.001	0.258
CXP	0.103	1	0.005	0.005	1.379
CXM	0.185	1	0.011	0.011	2.716
PXM	0.022	1	0.001	0.001	0.268
txNxC	0.489	2	0.045	0.022	5.738*
NxCXM	0.495	1	0.046	0.046	11.774*
Error		12	0.047	0.004	

Note. t = missing data type, i = number of items, N = sample, C =number of classes, P=missing data percentage, M = model.
*p<.05.

3.2.3. *ANOVA Results for guessing parameter*

In Table 7 for each factor, main effect, two-way, and three-way interactions are shown for guessing parameter.

Table 7. *ANOVA results for main effects and interaction effects of simulation conditions for guessing parameter.*

Factor	η_p^2	*df*	Sum of Squares	Mean Square	F
t	0.951	2	0.000	0.000	19.316
i	1.000	1	0.008	0.008	42941.408*
N	0.64	1	0.000	0.000	54.039*
C	1.000	1	0.002	0.002	9336.320*
P	0.769	1	0.000	0.000	6.671
ixC	0.999	1	0.000	0.000	1547.526*
Error		2	0.000	0.000	

Note. *t* = missing data type, i = number of items, N = sample, C =number of classes, P=missing data percentage.
*p<.05.

When Table 7 is examined, according to partial eta-squared values, it can be seen that especially item and class factors had a large effect on the results, but main effects of missing data type and missing data percentage were found to have no significant effects. Mean RMSE values for item guessing parameter were also significantly affected by interaction between item and class factors.

3.3. Classification Accuracy Results

Table 8 shows the classification rates for mixture IRT models.

Table 8. *The Classification rates for the Mixture Models.*

	Mixture Rasch			Mixture 2PLM			Mixture 3PLM		
Conditions	COMP	MAR	MNAR	COMP	MAR	MNAR	COMP	MAR	MNAR
2C10P60010	98.62	85.03	81.06	98.36	86.85	72.54	98.51	83.71	77.93
2C10P60030	93.01	87.48	83.07	98.71	87.35	75.44	98.77	88.79	79.98
2C10P100010	98.62	85.26	84.58	98.34	86.71	72.79	96.90	83.87	77.92
2C10P100030	99.02	87.81	82.93	99.02	87.41	76.06	99.58	89.30	81.33
2C20P60010	98.60	79.98	72.03	98.26	72.18	65.02	95.66	81.20	68.47
2C20P60030	89.02	76.59	72.93	97.03	77.01	72.69	97.01	72.30	73.18
2C20P100010	98.65	75.61	75.70	98.33	72.97	64.90	96.22	67.25	68.26
2C20P100030	88.01	76.56	62.44	86.02	77.08	72.90	99.60	75.88	70.46
3C10P60010	85.26	64.49	64.87	83.70	75.34	68.44	84.37	61.45	61.54
3C10P60030	83.32	74.43	75.30	83.92	83.47	79.83	83.32	70.96	68.79
3C10P100010	88.58	67.72	71.61	86.51	76.82	72.01	86.76	62.20	65.25
3C10P100030	84.92	75.85	75.54	84.92	84.18	80.15	84.92	74.38	71.22
3C20P60010	77.66	61.75	59.25	83.33	71.42	67.10	80.21	62.64	66.75
3C20P60030	83.70	63.81	61.86	93.70	72.14	76.08	81.80	62.28	68.33
3C20P100010	78.60	65.20	60.83	91.66	74.86	68.05	84.65	66.06	69.67
3C20P100030	85.26	64.44	62.54	93.24	71.75	77.83	82.53	63.53	70.46

As can be seen in Table 8, higher classification accuracy percentages were obtained for the compelete data case in the Mixture Rasch model. In the complete data condition, the highest percentage of classification accuracy was achieved for 2 class with 10% of missing data, 30 item, and a sample size of 1000 (99.02), while the lowest percentage of classification accuracy was achieved for 3 class with 20% missing data, 10 item, and a sample size of 600 (77.67). According to the missing data type, lower classification accuracy percentages were obtained in MAR and MNAR pattern conditions. In the MAR pattern condition, the highest percentage of classification accuracy was achieved for 2 class, 10% missing data, 30 item, and a sample size of 1000 (87.81), while the lowest percentage of classification accuracy was obtained 3 class, 20% missing data, 10 item, and a sample size of 600 (61.75) condition. In the MNAR pattern condition, the highest percentage of classification accuracy was reached for 2 class, 10% missing data, 10 item, and a sample size of 1000 (84.58), while the lowest percentage of classification accuracy was found 3 classes, 20% missing data, 10 item, and a sample size of 600 (59.25) condition.

For the mixture 2PL model condition, higher percentages of classification accuracy were obtained for the complete data case. In the complete data condition, the highest percentage of classification accuracy was achieved in combinations of 2 class, 10% of missing data, 30 item, and a sample size of 1000 item condition (99.02), while the lowest percentage of classification accuracy was found for 3 class, 20% missing data, 10 item, and a sample size of 600 (83.32). According to the missing data type, lower classification accuracy percentages were obtained in

the MNAR pattern condition. In the MAR pattern condition, the highest percentage of classification accuracy was achieved for 2 class, 10% of missing data, 30 item, and a sample size of 1000 (87.41) while the lowest percentage of classification accuracy was achieved for 3 class, 20% missing data, 10 item, and a sample size of 600 (71.42) condition. In the MNAR pattern condition, the highest percentage of classification accuracy was achieved for 3 class, 10% of missing data, 30 item, and a sample size of 1000 (80.15), while the lowest percentage of classification accuracy was found for 3 class, 20% missing data, 10 item, and a sample size of 600 (64.90) condition.

For the mixture 3PL model condition, higher percentages of classification accuracy were obtained for the complete data case as well. In the complete data condition, the highest percentage of classification accuracy was achieved for the combinations of 10% (99.58) and 20% (99.60) missing data percentages of 2 class with 30 item and a sample size of 1000 condition. The lowest percentage of classification accuracy was achieved for 3 class, 20% missing data, 10 item, and a sample size of 600 (80.213) condition. According to the missing data type, lower classification accuracy percentages were obtained under MAR and MNAR missing data pattern conditions. The highest percentage of classification accuracy obtained for the MAR and MNAR missing data pattern was 2 class, 10% of missing data was in the condition of 30 item and a sample size of 1000, and the lowest percentage of classification accuracy was in 3 class, 10% missing data percentage, 10 item, and a sample size of 600 condition.

4. DISCUSSION and CONCLUSION

Although mixture IRT models have been found to be useful in the fields of pscyhology, education and medicine, little research has been reported on the effects of sample size, number of items, number of latent classes, missing data, factors on model parameter estimates, and classification accuracy. In this research, a simulation study was conducted to examine the effects of estimation model, the number of items, sample size, the number of latent classes, missing data type, the percentage of missing data conditions on item parameter recovery, and classification accuracy for three mixture IRT models. The mean RMSE values were examined for parameter recovery. Furthermore, the main effects and interaction effects of the factors were examined. In addition, classification accuracy percentages were obtained by comparing the estimated latent class memberships with the true class memberships.

The findings indicate that, in the estimation of item difficulty and discrimination parameters for mixture IRT models, lower mean RMSE values were obtained as the sample size and number of items increased; on the other hand, the mean RMSE value increased as the number of classes increased. In the estimation of the guessing parameter, it was seen that the mean RMSE value decreases as the sample size, number of items and classes increase. These results match the ones observed in other studies. Previous studies investigating the effect of sample size, number of items, number of classes on parameter recovery for Mixture IRT models on item difficulty, and item discrimination parameter estimation (Alexeev et al., 2011; Cho et al., 2013; Finch & French, 2012; Li et al., 2009; Preinerstorfer & Formann , 2012; Sen et al., 2016) found that the mean RMSE value decreased as the sample size and number of items increased, and the mean RMSE values increased as the number of classes increased. In the estimation of the guessing parameter, it was observed that the mean RMSE values decreased as the sample size, number of items, and number of classes increased (Finch & French, 2012; Sen et al., 2016). It can be said that the results of this study are consistent with those in the related literature. It has been suggested that when the number of classes increases, it is natural for the error values to increase due to the decrease in the number of individuals in the classes (Finch & French, 2012).

In the item difficulty and item discrimination parameter estimations for the mixture models, lower mean RMSE values were obtained for the complete data cases, and higher mean RMSE

values were obtained for the MAR and MNAR cases of the missing data type. In the cases where the percentage of missing data was 20%, higher mean RMSE values were achieved. Similar results were found in a study in the literature in which mixture Rasch and mixture 2PL model and missing data type and percentage conditions were discussed (Lee, 2012). Obtaining mean RMSE values close to each other for the guessing parameter according to the missing data types corroborate the findings of Finch (2008), where the mean RMSE values of MAR and MNAR conditions were found to be low and very close to each other in the estimation of guessing parameter for IRT models. Since the missing data generation mechanism was produced as in Finch (2008), and the missing data were analyzed by the FIML method without assigning missing data, it seems natural that the mean RMSE values for the guessing parameter are close to each other.

In the recovery of the item parameters, it was observed that the mean RMSE values obtained for the Mixture 3PL model were higher than the mean RMSE values obtained for the Mixture 2PL model, and that the mean RMSE values obtained for the Mixture 2PL model were higher than the mean RMSE values obtained for the Mixture Rasch model. In the estimation of item parameters, a pattern of RMSE values appears to increase as the complexity of the model increases. Therefore, it can be said that the item parameters obtained for the Mixture Rasch model have fewer errors than those of the Mixture 2PL and 3PL models. These results are in agreement with the studies that obtained lower RMSE values for the two-class Mixture Rasch model (Cho et al., 2013; Sen, 2014).

In addition, when the main effects and interaction effects of the factors were examined, significant and high effect size values were obtained for the main effects of all factors considered in the estimation of item difficulty and discrimination parameters; however, for guessing parameter, it was obtained only for item, class, and model factors. These results suggest that interactions of factors may affect model parameter estimates and factors with high effect size values are important factors.

When the classification accuracy percentages were examined, higher classification accuracy percentages were obtained for the complete data case in all the Mixture IRT models. For all the mixture IRT models, in the complete data and MAR data condition, the highest percentage of classification accuracy was obtained in the combinations of 2 class, 10% missing data, 30 item, and a sample size of 1000, while the lowest classification accuracy was reached for the 3 class, 20% missing data percentage, 10 item, and a sample size of 600 condition. It was observed that lower classification accuracy was obtained for all the models in MAR and MNAR conditions.

5. SUGGESTION AND LIMITATIONS

The values used in the generation of item difficulty, item discrimination, and guessing parameters in this specific study are limited to the values used in the study of Li et al. (2009). In further studies researchers can change the item parameter generating values using different distributions. In addition, in this research, it is assumed that the ability parameter is randomly obtained from the standard normal distribution; using different ability distributions, researchers can examine the accuracy of recovery of item paremeters. In our simulation study 100 replications were performed for each condition; researchers can interpret the analysis results by changing the number of replications. In this study, the analysis of missing data was carried out using FIML method without using missing data assignment methods; by using missing data assignment methods, researchers can examine the effects of these methods in Mixture IRT models. In addition, the MLR estimation method was used for the estimation of the parameters; researchers can use different methods such as Bayesian and these methods can be compared. The results of this study are based on dichotomously scored items; researchers can perform Mixture IRT models analyses with polytomous scored items.

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

Fatima Munevver Saatcioglu: Investigation, Resources, Visualization, Software, Analysis, and Writing-original draft. **Hakan Yavuz Atar**: Methodology, Supervision, and Critical Review.

Orcid

Fatima Munevver Saatcioglu https://orcid.org/0000-0003-4797-207X
Hakan Yavuz Atar https://orcid.org/0000-0001-5372-1926

REFERENCES

Alexeev, N., Templin, J., & Cohen, A.S. (2011). Spurious latent classes in the mixture Rasch model. *Journal of Educational Measurement*, *48*, 313–332.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Academic.

Cohen, A.S., & Bolt, D.M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*,133–148.

Cho, S.-J., Cohen, A.S., & Kim, S.-H. (2013). Markov chain Monte Carlo estimation of a mixture item response theory model. *Journal of Statistical Computation and Simulation*, *83*, 278–306. https://doi.org/10.1080/00949655.2011.603090

Cho, H.J., Lee, J., & Kingston, N. (2012). Examining the effectiveness of test accommodation using DIF and a mixture IRT model. *Applied Measurement in Education*, *25*(4), 281–304. https://doi.org/10.1080/08957347.2012.714682

Cho, S.-J., Cohen, A.S., & Kim, S.-H. (2013). Markov chain Monte Carlo estimation of a mixture Rasch model. *Journal of Statistical Computation and Simulation*, *83*, 278–306. https://doi.org/10.1080/00949655.2011.603090

Choi Y.J., & Cohen, A.S. (2020). Comparison of scale identification methods in Mixture IRT models. *Journal of Modern Applied Statistical Methods*, *18*(1), eP2971. https://doi.org/10.22237/jmasm/1556669700

Collins, L.M., & Lanza, S.T. (2010). *Latent class and latent transition analysis*. John Wiley & Sons.

De Ayala, R.J., Plake, B.S. & Impara, J.C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement,38*, 213–234. https://doi.org/10.1111/j.1745-3984.2001.tb01124.x

De Ayala, R.J. & Santiago, S.Y. (2017). An introduction to mixture item response theory models. *Journal of School Psychology*, *60*, 25-40. https://doi.org/10.1016/j.jsp.2016.01.002

Edwards, J.M., & Finch, W.H. (2018). Recursive partitioning methods for data imputation in the context of item response theory: *A Monte Carlo simulation. Psicológica, 39(1)*, 88-117. https://doi.org/10.2478/psicolj-2018-0005

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, *45*(3), 225-245. https://doi.org/10.1111/j.1745-3984.2008.00062.x

Finch, W.H., & French, B.F. (2012). Parameter estimation with mixture item response theory models: A monte carlo comparison of maximum likelihood and bayesian methods. *Journal of Modern Applied Statistical Methods*, *11*(1), 167-178.

Hallquist, M.N., & Wiley, J.F. (2018). MplusAutomation: An R package for facilitating large scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal, 25(4)*, 621-638. https://doi.org/10.1080/10705511.2017.1402334

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of Item Response Theory.* Sage.

Hohensinn, C., & Kubinger, K.D. (2011). Applying item response theory methods to examine the impact of different response formats. *Educational and Psychological Measurement, 71*, 732-746. https://doi.org/10.1177/0013164410390032

Jilke, S., Meuleman, B., & Van de Walle, S. (2015). We need to compare, but how? Measurement equivalence in comparative public administration. *Public Administration Review*, *75*(1), 36–48. https://doi.org/10.1111/puar.12318

Kolen, M.J., & Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices*. Springer.

Kutscher, T., Eid, M., & Crayen, C. (2019). Sample size requirements for applying mixed polytomous item response models: Results of a Monte Carlo simulation study. *Frontiers in Psychology*, *10*, 2494. https://doi.org/10.3389/fpsyg.2019.02494

Lee, S. (2012). *The Impact of Missıng Data on The Dichotomous Mixture IRT Models* [Unpublished Doctoral Dissertation]. The University of Georgia.

Lee, S., Han, S., & Choi, S.W. (2021). DIF detection with zero-inflation under the factor mixture modeling framework. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131644211028995

Li, F., Cohen, A.S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for mixture dichotomous MTK models. *Applied Psychological Measurement*, *33*, 353-373. https://doi.org/10.1177/0146621608326422

Little, R.J.A., & Rubin, D.B. (1987). *Statistical analysis with missing data*. Wiley.

Maij-de Meij, A.M., Kelderman, H., & van der Flier, H. (2008). Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement, 32*, 611-631. https://doi.org/10.1177/0146621607312613

Muthén, L.K. & Muthén, B.O. (1998-2017). Mplus User's Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén.

Norouzian, R., & Plonsky, L. (2018). Eta- and partial eta-squared in L2 research: A cautionary review and guide to more appropriate usage. *Second Language Research*, *34*, 257–271. https://doi.org/10.1177/0267658316684904

Oliveri, M.E., Ercikan, K., Zumbo, B.D., & Lawless, R. (2014). Uncovering substantive patterns in student responses in international large-scale assessments-Comparing a latent class to a manifest DIF approach. *International Journal of Testing*, *14*(3), 265-287. https://doi.org/10.1080/15305058.2014.891223

Park, Y.S., Lee, Y.-S., & Xing, K. (2016). Investigating the impact of item parameter drift for item response theory models with mixture distributions. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2016.00255

Pohl, S., Grafe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement, 74(3)*, 423–452. https://doi.org/10.1177/0013164413504926

Preinerstorfer, D., & Formann, A.K. (2012). Parameter recovery and model selection in mixed Rasch models. *British Journal of Mathematical and Statistical Psychology*, *65*, 251–262. https://doi.org/10.1111/j.2044-8317.2011.02020.x

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

Richardson, J.T. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review 6*, 135-47. https://doi.org/10.1016/j.edurev.2010.12.001

Rost, J. (1990). Rasch Models in Latent Classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271–282.

Rost, J., Carstensen, C., & von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In J. Rost, & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 324-332). Waxmann.

Sen, S. (2014). *Robustness of mixture IRT models to violations of latent normality*. [Unpublished Doctoral Dissertation]. The University of Georgia.

Sen, S., Choen, A.S., & Kim, S.H. (2016). The impact of non-normality on extraction of spurious latent classes in mixture IRT models. *Applied Psychological Measurement*, *40*(2), 98-113. https://doi.org/10.1177/0146621615605080

Sen, S., & Cohen, A.S. (2019). Applications of mixture IRT models: A literature review, *Measurement: Interdisciplinary Research and Perspectives*, *17*(4), 177-191. https://doi.org/10.1080/15366367.2019.1583506

Zhang, D., Orrill, C., & Campbell, T. (2015). Using the mixture Rasch model to explore knowledge resources students invoke in mathematics and science assessments. *School Science and Mathematics*, *115*(7), 356-365. https://doi.org/10.1111/ssm.12135

International Journal of Assessment Tools in Education
2022, Vol. 9, No. 4, 1030–1061
https://doi.org/10.21449/ijate.1058145
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

A Comperative Adaptation of the Crick Learning for Resilient Agency (CLARA) with Classical Test Theory and Item Response Theory

Hasan Fehmi Ozdemir[1*], **Omer Kutlu**[2], **Shaofu Huang**[3], **Ruth Crick**[4]

[1]Şırnak University, School of Physical Education and Sports, Department of Physical Education and Sports, Türkiye
[2]Ankara University, Faculty of Education, Department of Educational Sciences, Türkiye
[3]WILD Learning Sciences CIC, United Kingdom
[4]University of Technology Sydney, Connected Intelligence Centre, Australia

ARTICLE HISTORY
Received: Jan. 15, 2022
Revised: Aug. 01, 2022
Accepted: Sep. 18, 2022

Keywords:
Learning power, Resilient agency, Classical Test Theory, Item Response Theory, Intercultural test adaptation.

Abstract: The aim of this study is to adapt the Crick Learning for Resilient Agency (CLARA) to Turkish culture, and to examine the psychometric features of the Inventory according to both Classical Test Theory (CTT) and Item Response Theory (IRT). In this respect, it is a descriptive level survey design research. Two different study groups were formed in accordance with the purpose of the study. Lingual equivalence applications were performed on two separate groups, one of which consisted of English Language and Literature Department students and the other consisted of English Language instructors. 1054 students participated in the validity and reliability studies from 101 different undergraduate programs at Ankara University. Before testing the research questions, it was examined whether the assumptions of CTT and IRT were met. With the application data; the predicted item discrimination indices, ability levels, students' scores forming their learning power profiles, and reliability coefficient values were found to be similar in both theories. It can be said that with CLARA-Tr, obtained by adapting CLARA, a valid and reliable tool has been provided to the Turkish literature to be used in future studies.

1. INTRODUCTION

Psychological tests are the subject of determining the cognitive, affective and dynamic characteristics of people and are used in scientific fields such as medicine, psychology and education. In general terms, tests provide information about the psychological characteristics of individuals and help to make decisions about individuals based on the results obtained from their application (Cohen & Swerdlik, 2010; Cronbach, 1960).

Wherever there are psychological activities, emphasis is placed on studies related to psychological tests. Studies on test or scale development and adaptation have an important place in Turkish literature. As different aspects and characteristics of human behavior are discovered, the need for different assessment tools to measure these characteristics is increasing. Instruments, measuring different psychological structures for different age groups are needed.

*CONTACT: Hasan Fehmi Ozdemir ✉ hfozdemir@sirnak.edu.tr Şırnak University, School of Physical Education and Sports, Department of Physical Education and Sports, Türkiye

e-ISSN: 2148-7456 /© IJATE 2021

This requirement can be met by the development of new measurement tools or by adapting suitable measurement tools, developed in different cultures, to Turkish culture. Both ways have either superior, and advantageous or inferior, and disadvantageous aspects. However, scale adaptation studies have benefits such as; the widespread use of technical knowledge, the establishment of international joint research relationships and the increase of information exchange, the localization of psychology, the initiation of cross-cultural comparative studies, the increase in the potential of collecting objective data on various subjects in the country, and contributing to the production of knowledge through its use in other research studies (Hambleton & Patsula, 1999; Hambleton et al., 2005; International Test Commission, 2018; Savaşır, 1994).

Undoubtedly, one of the most important steps in scale development or adaptation studies is to demonstrate the experimental reliability and validity proofs of the instrument being developed or adapted. Because the value and usability of the findings or results, obtained from psychological measurement tools, to make decisions about individuals is directly related to the psychometric properties of these tools at scale and item levels. One step further, no matter how strong the theoretical background of a scientific research is, if the tools used in the data collection process do not have the necessary psychometric qualities, there will be a trust problem in the interpretation of the findings of a research study, and it will be inevitable to make wrong decisions with the results obtained from this tool (Özdemir et al., 2019). Another important point is to use different theories and various methods and techniques developed based on these theories to determine the psychometric properties of measurement tools.

In the Turkish literature, there are studies in which Classical Test Theory (CTT) and Item Response Theory (IRT) are used in measurement tool development processes or in the prediction of item and test parameters of previously developed tools (Karakılıç, 2009; Kelecioğlu, 2001; Nartgün, 2002; Uysal, 2015). However, it is observed that there are many measurement tools adapted to Turkish culture in order to measure psychological characteristics, and almost all of these instruments' adaptation processes are based on CTT, due to the ease of implementation. However, when both theories are compared, it is known that the CTT has some limitations compared to the IRT (Hambleton, Swaminathan, & Rogers, 1991). In the intercultural adaptation studies of measurement tools, it is important to determine the psychometric properties not only according to CTT, but also to IRT, which has stronger assumptions. As a result of examining the harmony of the qualities determined by the methods and techniques based on different test theories, the usability of the scores to be obtained by the application of the said measurement tools will also increase.

Another important issue that has been frequently criticized in recent years is the proliferation of the test-oriented teaching and learning practices. The widespread use of large-scale tests and evaluations based on their results, force many tutors around the world to teach learners only multiple-choice test taking tips and the strategies to deal with them. This causes many learners to fail, by preventing them from gaining knowledge about participation in learning processes and self-learning (Deakin Crick et al., 2004). The way that will lead individuals to a solution is to encourage them to learn in a willing and relevant way in the face of new needs and opportunities. For this reason, in order to raise individuals with the mentioned qualities, education and measurement policies should be structured differently, and educational institutions at all levels should be structured to serve this.

If the capacity and willingness to learn and continue learning throughout life is accepted as the central point in the concept of "learning", it is of great importance to use tools that measure the capacities and desires of individuals and their constantly evolving and changing qualities. "What makes the individual participate in the learning process, continue his/her learning, and want to learn effectively and efficiently while doing this?" The answer to this question has been

an important starting point for the development of different measurement tools. In this context, one of the measurement tools that we come across in the literature is the Crick Learning for Resilient Agency (CLARA), which defines and measures the "learning power" of an individual (Deakin Crick et al., 2015). In the Turkish literature reviews, no measurement tool was found to measure learning power. Therefore, it is thought that adapting the CLARA, which is widely used in the international literature and has appropriate psychometric properties, to Turkish culture will contribute to the Turkish society and the field of measurement and evaluation.

1.1. Aim of the Study and Research Questions

The main aim of this study is to adapt Crick Learning for Resilient Agency (CLARA) Inventory to Turkish culture as CLARA-Tr, and to analyze and compare the psychometric properties of the Inventory in the adaptation process according to the methods and techniques of both Classical Test Theory (CTT) and Item Response Theory (IRT). In line with this main aim, answers to the following research questions were sought:

1. Is there a relationship between the scores obtained from the English and Turkish forms of CLARA?
2. Is CLARA's original factor structure confirmed in Turkish culture?
3. What is the relationship between the values of the psychometric properties of Turkish form of CLARA estimated according to the CTT and IRT?
   - 3.1. Is there a relationship between the item discrimination indices (item score - corrected total score correlation and a_i parameter) of CLARA-Tr's items according to the CTT and IRT?
   - 3.2. Is there a relationship between the levels of the features/traits measured by CLARA-Tr items (arithmetic mean and b parameter) estimated according to the CTT and IRT?
   - 3.3. Is there a relationship between learning power levels estimated from CLARA-Tr according to CTT and IRT?
   - 3.4. What is the reliability of CLARA-Tr according to the CTT and IRT?

1.2. Significance of the Study

One of the priorities included in the Lifelong Learning Strategy Document and Action Plan for the period 2014-2018 (Ministry of National Education [MoNE], 2014; 2018), which was prepared to increase the effectiveness and efficiency of the lifelong learning system in Türkiye, is "constructing the culture and raising the awareness of lifelong learning in society". In this context, it is planned to expand the studies for the adult population to acquire basic skills (such as communication in mother tongue and foreign language, digital competencies, learning to learn etc.). In order to achieve this, the individual must first recognize himself/herself, recognize his/her weaknesses and strengths as a learner, and see learning as a necessity. Within the scope of "development of lifelong learning monitoring and evaluation system", which is another priority in the said Strategy Document, creating statistics and researches is expected to be done from responsible institutions and organizations (Ministry of National Education and Higher Education Council, Universities in this context) that will help develop policies and strategies.

One of the important reasons for conducting such a study is that there is no measurement tool that measures learning power in the Turkish literature reviews and the need to do more research on metacognitive skills such as self-awareness, curiosity, creativity, readiness to learn, and resilience, which are among the basic life skills. Due to the requisite and important need for resilient agency at every stage of individuals' learning journeys from purpose to performance, it has been acted with the thought that it will make a significant contribution to the priorities and achievement of these priorities in Türkiye Lifelong Learning Strategy Document and

Action Plan 2014-2018. In this context, it was decided to adapt and introduce Crick Learning for Resilient Agency (CLARA) self-assessment tool into Turkish culture.

Another important reason for the selection of this Inventory is that CLARA has not only the ability to satisfy the requirements of the researches in which it is used as a data collection tool, but also to provide instant feedback and forward notifications based on monitoring (formative) to individuals as learners. Because, some studies conducted in Türkiye reveal the low lifelong learning disposition levels of undergraduate students (Diker Coşkun & Demirel, 2012; Tunca et al., 2015). Changing this negative perception and letting the university students to see the strengths and weaknesses of their own learning power as independent learners at the undergraduate level, is seen as an important investment in their learning journeys that continue from cradle to grave, and therefore to themselves and the society they live in.

As a result, both a new measurement tool has been added to the Turkish literature for the researchers who want to have knowledge about learning power and for their future studies, and also an example was provided for the comparison process of test theories in the intercultural test/scale adaptation process.

1.3. The Crick Learning for Resilient Agency (CLARA) Profile

The research programme which has led to the publication and various applications of CLARA began in 2000 at the University of Bristol, UK. Originally funded by the LifeLong Learning Foundation, and building on the work of Carr and Claxton (2002) it addressed the challenge of identifying personal qualities and characteristics which define a ‘good learner’- someone who is able to engage effectively and profitably with new learning opportunities across the lifespan. As well as identifying these qualities, the purpose of the research was to devise a learning analytics tool that could be used to assess where an individual was located on those qualities at any given time and in any given context and thus provide them with data that could be used formatively to enable them to develop their capacity to learn how to learn. Then the Assessment Reform Group (2010, December) in the UK had developed a significant programme of work, which aimed make ‘assessment for learning’ a focus for policy and practice Broadfoot (1998). There was, even then, substantial evidence of the negative impact of high stakes testing and summative assessment on students’ motivation for learning Harlen and Deakin Crick (2003a and 2003b) and this programme of research set out to develop alternative forms of assessment for learning that could be both formative for teachers and ipsative for learners in that it could provide a foundation for teacher supported but student-led, self-directed change in learning how to learn.

The original research (Deakin Crick et al., 2004; Deakin Crick and Wilson, 2005; Deakin Crick, 2005; Deakin Crick, 2007) was a factor analytic study which drew together items created to reflect what was known at the time about lifelong learning and ‘learning power’ a popular term coined first by (Claxton, 1999) to refer to a person’s capacity for learning how to learn. It drew on a susbtantive literature review and included items from socio-cultural learning theory and pedagogical studies. The factor analysis produced seven latent variables, which have remained constant over time through successive quantitative studies (Arthur et al., 2006; Deakin Crick and Yu, 2008; Deakin Crick et al., 2013; Deakin Crick et al., 2015). The original tool was called ELLI (The Effective Lifelong Learning Inventory).

From 2001 the data was used in practice as well as research, and returned to teachers in classrooms as a digital learning analytic in the form of a spider diagram. In keeping with the theoretical foundations of the study, this was designed so that teachers and learners were encouraged to explore patterns and interpretations, rather than a numerical score, or set of scores, which would inevitably lead to a more summative self-judgement (Deakin Crick, 2005; Deakin Crick, 2006; Deakin Crick and McCombs, 2006; Deakin Crick, 2009a and 2009b).

A programme, which attended to both practice, research and policy, was a challenge in a traditional University. After three failed attempts to commercialise the work, the University enabled a re-analysis of accumulated data and the publication of a revised version known as the Crick Learning for Resilient Agency profile (Deakin Crick et al., 2015). As a creative commons publication this opened up new opportunities for ongoing research and development. The study reported in this paper builds on this work.

1.4. Learning Power

The term learning power has come into popular usage to describe the capacity a person has to learn and to engage profitably with risk, uncertainty and challenge. In other words, they know how to go about finding out what to do when the solution to a challenge is not known in advance. The ELLI tool and subsequently the CLARA tool built on theoretical foundations which took seriously a holistic approach to learning. This included the role of (i) dispositions, awarenesses and skills (ii) identities – the beliefs, values and attitudes about self, learning and knowledge held by the learner, (iii) narratives – the socio-cultural formation of learners over time and (iv) the quality and substance of learning relationships (Deakin Crick, Broadfoot and Claxton, 2004). This led to a set of Scales, known as dimensions of learning power, which measured eight variables. Each of these included cognition, affect and volition and were presented to learners in real time as a reflection of their 'learning power' in a particular context at a particular point in time. On the basis of the underlying theory of agency and choice, the feedback was designed to stimulate learner ownership, awareness and responsibility for self-directed change. For this reason, the visual imager was important in assessment terms because it stimulates reflection on one's self-identity and story and offers opportunities for reflexive self-awareness and change in purposeful agency.

The Scales of CLARA are presented in greater depths elsewhere (Deakin Crick et al., 2015). A summary is presented in Figure 1 below.

Figure 1. *The Scales of CLARA.*

1.4.1. *Mindful agency scale (9 items)*

Mindful Agency is taking responsibility for your own learning. It's about how you manage your feelings, your time, your energy, your actions and the things you need to achieve your goals. It's knowing your purpose - then knowing how to go about achieving it; stepping out on the path towards your goals.

1.4.2. *Hope and optimism scale (3 items)*

Hope and Optimism is being confident that you can change and learn and get better over time. It is helped by having a positive learning story to reflect upon, that gives you a feeling of having 'come a long way' and of being able to 'go places' with your learning.

1.4.3. *Sense making scale (7 items)*

Sense Making is making connections between ideas, memories, facts - everything you know - linking them and seeing patterns and meaning. It's about how 'learning matters' to you, connecting with your own story and things that really matter.

1.4.4. *Creativity scale (8 items)*

Creativity is using your imagination and intuition, being playful and 'dreaming' new ideas, having hunches, letting answers come to you, rather than just 'racking your brains' or looking things up. It's about going 'off the beaten track' and exploring ideas.

1.4.5. *Curiosity scale (6 items)*

Curiosity is your desire to get beneath the surface, find things out and ask questions, especially 'Why?' If you are a curious learner, you won't simply accept what you are told without wanting to know for yourself whether and why it's true.

1.4.6. *Collaboration scale (3 items*)

Collaboration is how you learn through your relationships with others. It is about knowing who to turn to for advice and how to offer it too. It's about solving problems by talking them through, generating new ideas through listening carefully, making suggestions and responding positively to feedback.

1.4.7. *Belonging scale (3 items)*

Belonging reflects how much you feel you belong as part of a 'learning community' – at work or at home, or in your wider social network. It's about the confidence you gain from knowing there are people you learn well together with and to whom you can turn when you need guidance, support and encouragement.

1.4.8. *Orientation to learning scale (10 items)*

Orientation to Learning is about the degree to which a person is open to new ideas and to challenge and having the 'inner strength' to move towards learning and change, rather than either giving up and withdrawing or 'toughing it out' and getting mad with the world. Becoming more open to learning is like a pathway to all the other Scales of learning power, and just as the other Scales it also help you become more open to learning. This Scale is sometimes referred to simply as 'Openness to Learning'.

1.5. Resilient Agency

The term resilience is much used in various contexts and domains. In the psychological literature resilience refers to those qualities that an individual has that enables them to succeed despite adverse conditions or circumstances (Rutter, 1985; Rutter, 2012; Masten, 2007). In the 2015 revision of the learning power assessment tool, the term was chosen to describe the overall purpose of the whole assessment event, in response to all of the now eight Scales of learning power, which is to empower the individual to understand themselves as a learner and to use that

understanding to explore strategies for change. In the early version, ELLI, the Scale now called orientation to learning was described as Resilience in keeping with commercial applications of learning power (Gornall et al., 2005). However, the data demonstrated that simply persisting in particular behaviours did not necessarily enable one to succeed despite adverse conditions or circumstances. Indeed, in some contexts, it led to more negative outcomes (Deakin Crick and Salway, 2006). Resilience in learning is complex and includes the capacity to persist, but also must include the capacity to explore identity and purpose, to generate questions, utilize one's imagination and develop positive relationships. In the context of developing learning power, Resilient Agency was identified as a descriptor for the purpose of the whole assessment event, which is to stimulate self-leadership and self-directed change strategies which lead towards a more profitable future.

2. METHOD

2.1. Research Model

In this study, it was examined whether the psychometric properties of Crick Learning for Resilient Agency (CLARA) determined by different methods and techniques of Classical Test Theory (CTT) and Item Response Theory (IRT) differed or not in the process of adapting the Inventory to Turkish culture. The study aims to reveal the psychometric properties of the said Inventory as they exist on the basis of two different test theories. In this respect, this study is a descriptive survey research (Büyüköztürk et al., 2014; Erkuş, 2013; Scott & Usher, 2011).

2.2. Study Groups

In scale adaptation studies, due to the limitations in terms of time, money and labor, the sample is chosen from easily accessible and practicable units. For this reason, instead of working with the population and sample, it is preferable to conduct the research with a "study group", which is reached through convenient sampling from individuals similar to the target group. In this study, the target group was determined as undergraduate students, and in line with the purpose of the study, two different study groups were formed from students studying at different departments of Ankara University, and also a group of English lecturers working at Ankara University have participated in the lingual equivalence applications.

2.2.1. *Linguistic equivalence application groups*

Linguistic equivalence applications were carried out on two separate groups that were deemed to be sufficient in both languages. In the first group, there were a total of 31 students from the 2nd and 4th grade students who are continuing their education in Ankara University, Department of English Language and Literature. In the second group there were 35 English lecturers working at Ankara University Turkish and Foreign Language Application and Research Center.

2.2.2. *Validity & reliability studies application group*

It has been taken into consideration that the analyzes to be made in order to determine the psychometric properties of the adapted instrument will be made according to both CTT and IRT. For this reason, taking into account the lower limits of the number of participants suggested by researchers such as Crocker and Algina (1986), Reise and Yu (1990) and De Ayala (2009), which is sufficient for statistical methods to be used and necessary to provide assumptions and to ensure variability, this application was conducted on a group of 1054 students who are continuing their education at 101 different undergraduate programs of Ankara University. 33.11% (n = 349) of the students in this group are male and 66.89% (n = 705) are female. Considering the grade levels, 2.56% (n = 27) of the group was preparatory class, 7.97% (n = 84) were 1st grade, 16.41% (n = 173) 2nd grade, 24.67% (n = 260) 3rd grade, 42.41% (n = 447) 4th grade, 4.74% (n = 50) 5th grade and 1.23% (n = 13) 6th grade students.

2.3. The Adaptation Process of the Crick Learning for Resilient Agency (CLARA)

The following steps have been followed in the process of adapting CLARA, which is planned to be introduced into Turkish psychometry field:

1. Participation in the workshop organized in Bristol / England, in order to receive the necessary training on CLARA's application, scoring and interpretation of the scores.

2. CLARA, was translated from its original language English to Turkish by a group of expert translators who have mastered the language and culture, and then back translated into Turkish by a different group of translators. The back-translations of the Inventory, and the Scale names, and also the items, and the response categories were shared with the developers, and their opinions and approvals were received. The original form, the form translated into Turkish and the back translation form were presented to the evaluation of a group of instructors who know both languages well and who are knowledgeable about measurement and learning. While considering the back-translations, the evaluators were asked to compare the Turkish translation form with the original form, in terms of language and meaning.

3. The necessary corrections were made in line with the suggestions and evaluations of the expert group, and the final version of the Turkish form was presented to the opinion of the Turkish language experts and final checks were carried out.

4. Bilingual group design was used to ensure linguistic equivalence. In this direction, it is necessary to apply the instrument's original and translated forms on a group that is deemed to be sufficient in both languages. For this reason, applications were made in two separate groups in order to test whether linguistic equivalence was achieved. In both groups, the original form and the translation form of the tool were applied every three weeks. After the applications, the relationship between the scores obtained from the original and target language forms of the scale was examined.

In this study, the procedure steps suggested by Hambleton and Swaminathan (1985) for the estimation of psychometric properties of Likert type measuring instruments based on IRT were followed. In the estimation of the psychometric properties of CLARA based on IRT, the inventory was first applied to a group with a high number of participants. It was tested whether the data meet the IRT assumptions; unidimensionality and local independence, and whether the data fit the selected model. Ability levels (θ) and item parameters were estimated with MULTILOG 7.03 program. Also, IBM SPSS 22 and LISREL 8.8 were used for statistical analysis of the data within the scope of the study. Before starting the testing phase of the research questions, it was examined whether the data met the CTT and IRT assumptions required for analysis. Kolmogorov-Smirnov and Shapiro Wilk tests were used together with descriptive statistics, and the histogram graphs in the analysis of whether the data provided the assumption of normality. In testing the assumptions of unidimensionality and local independence, the results of two confirmatory factor analysis were used. In terms of Item Response Theory, data model fit was analyzed using the "-2 lnL" statistic, and also the level of data-model fit was examined by the difference between the observed and expected proportions of responses to the item response categories.

An example of the MULTILOG program output (Belonging Scale) showing the a and b parameters estimated according to the IRT of the CLARA-Tr items used in this study, as well as the model-data fit and marginal reliability coefficient values are given in Appendix 1.

3. RESULT

The results / findings obtained regarding the research questions are given and discussed below respectively.

3.1. Findings Regarding the First Research Question – The Relationship Between the Scores Obtained from the Application of English and Turkish Forms of CLARA

In order to search for an answer to the question "Is there a relationship between the scores obtained from the application of English and Turkish forms of CLARA?" and to test whether linguistic equivalence was achieved between the original and Turkish forms of the Inventory, linguistic equivalence applications were carried out in two separate groups (n1 = 31 and n2 = 35). In both groups, the original and the translation forms of the tool were applied three weeks apart, and the relationship between the scores obtained from these applications was examined with the Pearson Product-Moments Correlation coefficient. The correlation values are presented in Table 1.

Table 1. *Relationship Between Scores Obtained from English and Turkish Forms of CLARA.*

Scales	n_1=31		n_2=35	
(English / Turkish)	*r*	*p*	*r*	*p*
Belonging	0.75	0.000	0.78	0.000
Collaboration	0.72	0.000	0.71	0.000
Creativity	0.76	0.000	0.82	0.000
Curiosity	0.81	0.000	0.87	0.000
Hope & Optimism	0.70	0.000	0.73	0.000
Mindful Agency	0.78	0.000	0.79	0.000
Orientation to Learning	0.71	0.000	0.81	0.000
Sense Making	0.79	0.000	0.80	0.000

When Table 1 is examined, it is determined that there is a positive, high and significant (r = 0.70-0.87, p <0.01) relationship between the scores obtained from the English and Turkish forms of CLARA's both linguistic equivalence applications. Accordingly, it can be accepted that linguistic equivalence is provided between the original and Turkish forms of CLARA (Büyüköztürk et al., 2014). The item examples included in CLARA and CLARA-Tr that emerged as a result of this process are presented in Appendix 2.

3.2. Findings Regarding the Second Research Question – The Structure of CLARA Verified in Turkish Culture

Randomly chosen, with sufficient sample sizes two separate (n_1 = 550 and n_2 = 504) confirmatory factor analyzes were conducted on the data obtained from the validity & reliability studies application to find an answer to the question "Is the original structure of CLARA verified in Turkish culture?" and to determine whether the eight-scale original structure of the Inventory was also confirmed by Turkish undergraduates or not. The analyzes were carried out using LISREL 8.8 program. Covariances were used as the moment matrix, and maximum likelihood (ML) estimation method was used in CFA. Fit indices obtained as a result of the analyzes are given in Table 2.

Table 2. *Confirmatory Factor Analysis of CLARA Turkish Form Fit Indices.*

Fit Indices	CFA 1 (n=550)	CFA 2 (n=504)
	Values	
Chi - Square (X^2)	3956.82	2983.64
Degrees of Freedom (df)	1398	1152
X^2/sd	2.83	2.59
Non-Normed Fit Index (NNFI)	0.95	0.95
Comparative Fit Index (CFI)	0.96	0.95
Root Mean Square Error of Approximation (RMSEA)	0.069	0.066
Root Mean Square Residual (RMSR)	0.016	0.015
Standardized RMR	0.08	0.08

Fit indices of the models obtained from CFA's were examined and Chi-square values (χ^2 = 3956.82, N = 550, *df* = 1398, *p* = 0.00; χ^2 / *df* = 2.83 and χ^2 = 2983.64, N = 504, *df* = 1152, *p* = 0.00; χ^2 / *df* = 2.59) were found to be significant. Fit index values were obtained as RMSEA = .069 and .066, NNFI = .95 and .96, CFI = .95 and .95, RMR = .016 and .015, Standardized RMR = 0.08 and 0.08 respectively. 90% confidence interval of RMSEA are between 0.057-0.071 and 0.054-0.069. According to Jöreskog and Sörbom (1993), Hu & Bentler (1999), Kline (2005), Özdamar (2013), Sümer (2000), Şimşek (2007), Vieira (2011) the values in Table 2 indicate acceptable fit. According to these data, it was decided that the original structure of CLARA was also verified by Turkish undergraduate students, and that data on learning power could be collected from university students in a valid and reliable manner by its application.

3.3. Findings Regarding the Third Research Question – Relationship Between the Values of the Psychometric Properties of the Turkish Form of CLARA

The third research question of the study is "What is the relationship between the values of the psychometric properties of the Turkish form of CLARA, which are estimated based on CTT and IRT?" Findings and comments regarding the sub-questions to be answered within the scope of this question are presented below.

3.3.1. *Research question 3.1. findings – relationship between the item discrimination index values of CLARA-Tr*

"Is there a relationship between the item discrimination index values of CLARA-Tr, which are estimated based on CTT and IRT?" For this question, the relationship between the item discrimination indices of each item estimated according to two theories was tested with the Spearman Rank Difference Correlation Coefficient.

In the estimation of item discrimination index according to CTT, correlation based item analysis technique was used. For this purpose, the relationship between the responses of the participants to the items and their corrected total scores from the scale in which that item is included was calculated with the Pearson Product-Moment correlation coefficient. The corrected total score was calculated by subtracting each participant's relevant item score from his/her raw score obtained from that scale. In IRT, on the other hand, a parameter was estimated for each item according to the Graded Response Model of Samejima (Samejima, 1969) and the relationship between the values obtained according to both theories was examined. The Graded Response Model is an extension of the two-parameter logistic model (2PL). This model is appropriate when the responses of an individual to an item can be classified into more than two ordered categories, such as to represent different levels of agreement or frequency to a certain statement. In Table 3, the discrimination indices of the items in the Inventory, which is estimated based on CTT and IRT, are presented.

Table 3. *Discrimination Index Values of CLARA-Tr Items Estimated According to CTT and IRT.*

Scale	Item No	CTT (Item Score-Corrected Total Score Correlation)	IRT (a parameter)
Belonging	7	0.970	4.12
	17	0.972	4.46
	45	0.964	2.29
Collaboration	6	0.886	2.39
	35	0.930	0.98
	48	0.885	0.92
Creativity	9	0.982	1.80
	29	0.982	1.22
	1	0.983	1.86
	41	0.985	1.12
	16	0.984	0.94
	31	0.986	2.01
	39	0.982	1.58
	11	0.990	1.09
Curiosity	2	0.972	1.10
	47	0.978	1.34
	33	0.977	3.04
	22	0.978	2.90
	5	0.988	1.13
	38	0.986	1.23
Hope & Optimism	13	0.962	3.22
	24	0.967	1.91
	49	0.971	6.00
Mindful Agency	3	0.986	1.09
	10	0.992	1.41
	15	0.990	1.57
	23	0.989	1.32
	26	0.986	1.26
	34	0.985	1.34
	36	0.990	1.69
	43	0.986	1.36
	46	0.993	2.07
Orientation to Learning	14	0.979	1.03
	18	0.981	1.64
	20	0.985	1.98
	21	0.989	1.60
	25	0.986	1.46
	28	0.983	1.98
	30	0.989	1.15
	32	0.975	1.58
	37	0.992	1.61
	42	0.983	1.87
Sense Making	4	0.979	1.11
	8	0.963	1.13
	12	0.948	1.82
	19	0.910	1.61
	27	0.947	1.10
	40	0.973	1.32
	44	0.933	1.61

In Table 4, descriptive statistics of the discrimination index values of the items of the Inventory, which are estimated based on the CTT and IRT, are presented.

Table 4. *Descriptive Statistics of the Discrimination Index Values of CLARA-Tr Items Estimated According to CTT and IRT.*

Descriptive Statistics				
	CTT	Stand. Error	IRT	Stand. Error
Minimum	0.890		0.92	
Maximum	0.990		6.00	
$\bar{X}$	0.973	0.0035	1.78	0.139
Median	0.982		1.58	
SD	0.249		0.97	
Kurtosis	0.668		7.68	0.67
Skewness	0.340		2.55	0.34
Range	0.100		5.08	
Number of Items (k)	49		49	
Number of Students	1054		1054	

When Table 3 and 4 are examined together, it is seen that the discrimination indices of the items of the eight scales that make up the Inventory vary between 0.885 (Collaboration Scale, item 48) and 0.993 (Mindful Agency Scale, item 46) and the median is 0.982. In the analysis of correlation-based item discrimination, it is concluded that as the values approach 1.00, the item measures the feature/trait that is measured with the whole scale to which it belongs, and it can better discriminate the individuals who have this feature/trait and those who do not. Based on this, it was observed that all 49 items in 8 Scales of the Inventory, which was adapted to Turkish culture, had a high level of discrimination.

It is seen that the values of a parameter estimated according to the IRT vary between 0.92 (Collaboration Scale, item 25) and 6.00 (Hope and Optimism Scale, item 49) and the median is 1.58. In the IRT, it is accepted that the items with a discriminative power of 1.00 and above are sufficiently discriminating (Hambleton & Swaminathan, 1985). This can be interpreted as that 49 items of the inventory can discriminate the individuals who have the desired feature/trait to be measured with the scales they belong to, and those who do not.

Despite the good discrimination index values obtained according to the two test theories, only the items of the "Belonging" and "Mindful Agency" scales discrimination values determined according to CTT and IRT showed significant relationship when examined with the Spearman Rank Correlation Coefficient ($p < 0.05$), no significant relationship was found for the items of the other six scales. According to this result, it can be interpreted that the item discrimination indices of "Belonging" and "Mindful Agency" scales estimated according to the two theories are similar to each other and these values are comparable.

3.3.2. *Research question 3.2. findings – relationship between the levels of the features/traits measured by CLARA-Tr*

Another sub-question to be answered within the scope of the third research question of the study is "Is there a relationship between the levels of the features/traits measured by CLARA-Tr items (arithmetic mean and b parameter) estimated according to the CTT and IRT?" For this question, the relationship between the levels of the features/traits measured by each CLARA-Tr item based on two theories, was tested with Spearman Rank Differences Correlation.

According to the CTT, the levels of the features/traits measured by the items were calculated by the arithmetic mean of the responses given to the relevant item by the students in the study group. According to the IRT, the levels of the features/traits measured by each item were determined by taking the arithmetic mean of the b parameter values estimated according to

Samejima's Graded Response Model (Samejima, 1996). The values of the levels of the features/traits measured by the items based on both theories are given in Table 5.

Table 5. *Levels of the features/traits measured by CLARA-Tr Items Estimated According to CTT and IRT.*

Scale	Item No	CTT	IRT					
		Art. Mean	b_1	b_2	b_3	b_4	b_5	b_{AM}
Belonging	7	3.12	-1.38	-0.68	-0.14	0.18	0.72	-0.26
	17	4.43	-1.35	-0.76	-0.16	0.22	0.70	-0.27
	45	4.87	-1.51	-0.49	0.21	0.66	1.27	0.03
Collaboration	48	4.05	-2.94	-1.31	-0.04	0.95	2.29	-0.21
	6	4.05	-2.82	-1.96	-1.30	-0.80	0.00	-1.38
	35	4.31	-3.61	-2.16	-1.06	-0.26	0.80	-1.26
Creativity	9	4.63	-2.43	-1.26	-0.45	0.21	0.99	-0.59
	29	4.70	-2.61	-1.00	0.10	0.90	1.95	-0.13
	1	3.55	-5.60	-3.17	-1.19	-0.24	1.23	-1.79
	41	4.03	-4.36	-2.56	-1.11	-0.17	1.02	-1.44
	16	4.25	-4.66	-2.84	-1.48	-0.43	0.97	-1.69
	31	5.03	-2.74	-1.54	-0.68	0.02	0.85	-0.82
	39	3.38	-2.34	-0.92	-0.12	0.53	1.48	-0.27
	11	5.00	-6.28	-3.82	-2.29	-1.25	0.06	-2.72
Curiosity	2	3.23	-2.52	-0.66	0.59	1.40	2.68	0.30
	47	4.10	-3.77	-2.25	-1.23	-0.48	0.44	-1.46
	33	4.45	-2.01	-1.03	-0.28	0.21	0.83	-0.46
	22	4.17	-1.89	-1.04	-0.39	0.16	0.95	-0.44
	5	5.16	-5.25	-4.18	-2.26	-1.25	-0.01	-2.59
	38	5.01	-4.93	-3.47	-1.98	-0.92	0.26	-2.21
Hope & Optimism	49	5.64	-2.13	-1.16	-0.46	0.08	0.79	-0.58
	13	4.56	-2.02	-1.20	-0.47	0.13	0.93	-0.53
	24	5.16	-3.51	-2.51	-1.64	-0.95	0.00	-1.72
Mindful Agency	3	3.86	-3.60	-2.03	-0.81	0.25	1.87	-0.86
	15	4.71	-3.25	-2.44	-1.29	-0.38	0.82	-1.31
	43	4.24	-2.36	-0.73	0.39	1.15	2.26	0.14
	36	3.29	-3.76	-2.07	-1.01	-0.19	0.89	-1.23
	46	4.23	-3.46	-2.51	-1.48	-0.64	0.39	-1.54
	23	4.92	-4.15	-2.65	-1.67	-0.72	0.35	-1.77
	34	3.70	-2.62	-1.46	-0.40	0.48	1.70	-0.46
	26	3.86	-3.46	-2.13	-1.07	-0.25	0.90	-1.20
	10	4.99	-4.64	-3.49	-1.89	-0.90	0.34	-2.12
Orientation to Learning	20	4.56	-1.59	-0.37	0.47	1.08	1.80	0.28
	30	3.14	-0.63	1.22	2.13	2.81	3.82	1.87
	25	2.08	-8.20	-5.87	-3.16	-1.04	1.60	-3.33
	28	4.46	-1.40	-0.34	0.48	1.00	1.80	0.31
	14	5.41	-0.72	0.58	1.33	1.85	2.64	1.14
	42	3.89	-2.49	-1.36	-0.46	0.05	0.74	-0.70
	21	2.58	-1.16	0.18	1.15	1.89	2.70	0.95
	18	4.04	-5.08	-2.61	-0.81	0.50	2.31	-1.14
	32	3.73	-3.58	-1.70	-0.27	0.77	2.62	-0.43
	37	5.20	-8.78	-6.12	-4.18	-2.25	-0.06	-4.28
Sense Making	19	4.05	-3.32	-2.15	-1.23	-0.32	0.89	-1.23
	40	3.61	-8.51	-4.09	-1.58	0.60	3.74	-1.97
	4	2.51	-5.36	-4.54	-3.22	-1.96	-0.51	-3.12
	27	4.51	-5.07	-4.41	-3.83	-2.81	-1.23	-3.47
	8	4.46	-4.53	-2.44	-0.96	0.14	1.65	-1.23
	12	5.12	-3.79	-2.92	-1.75	-0.78	0.35	-1.78
	44	4.09	-3.14	-1.66	-0.52	0.31	1.25	-0.75

When Table 5 is examined, it is seen that the levels of the features/traits measured by the items according to the CTT vary between 2.51 (Sense Making Scale, item 4) and 5.64 (Hope and Optimism Scale, item 49) and the median is 4.24. It is seen that the vast majority of the items (38 items) have a negative skewness value and when all items are considered, the average skewness value is -4.32. When all these findings are evaluated together, it has been determined that both the items generally measure the feature/trait to be measured with the scales they belong to at a high level and all the items have a relatively high approval rate. In other words, it can be said that participating students have chosen the high-level end of the response categories.

According to the IRT, one less number of b parameters were estimated from the number of response categories of the items. Since the inventory has a six-point Likert response format, the number of b parameters estimated was five (b1 - b5). The b1 parameter estimated for an item is the ability (θ) level, which corresponds to the preference of the other five answer categories of the item to the first answer category, in other words, the choice of the second, third, fourth, fifth and sixth answer categories with a probability of 0.50. The b2 parameter is the ability (θ) level, which corresponds preferring the third, fourth, fifth and sixth answer categories with a probability of 0.50 instead of the first and second answer categories. The b3 parameter is the ability (θ) level, which corresponds to choosing the fourth, fifth and sixth answer categories with a probability of 0.50 instead of the first, second and third answer categories. With a similar logic, the b4 and b5 parameters also express the ability (θ) level, which corresponds to the preference of the relevant answer category and subsequent answer categories/category with a probability of 0.50 instead of the previous answer categories. When the item boundary parameter, that is, the b parameter values, are examined, it is seen that they mostly have negative values. Based on this, it can be said that the answers are mostly supported by the low level of the measured feature/trait (θ<0) (Uyar et al., 2013).

In this context, when the levels of the feature/trait measured by the items according to IRT is examined, the arithmetic mean values of five b parameters estimated for each items vary between -3.47 (Sense Making Scale, item 27) and 0.30 (Curiosity Scale, item 2), and the median is -1.140. According to the IRT, the low levels of the features/traits measured by the items are an indication that the higher level response categories are selected, the higher levels of the features/traits measured by the items are also the indicators that the lower level response categories are selected. The average of the arithmetic means of the levels of the feature/trait measured by the items estimated within the scope of the study is -1.056. Usually the b parameter can take a value between ± 3, with probability 0.50 representing the required θ level of feature/trait for the approval of the item. A negative b value can be interpreted as the items are better at distinguishing those with a low level of the trait of interest from those with a moderate level (Flannery et al., 1995).

When the frequency distribution of the responses to the items is examined, it is seen that although the students prefer each of the answer options at varying rates, they generally choose the high-level response categories. For example, the distribution of the answers according to the response categories for the 46th item in the Mindful Agency Scale, of which the item score average is 4.23 according to the CTT is; 1 = 6 (0.60%), 2 = 24 (2.30%), 3 = 100 (9.50%), 4 = 187 (17.70%), 5 = 334 (31.70%), 6 = 403 (38.20%). A similar trend to this item was observed in the rest of the items.

When Table 5 is examined, another point that stands out is that some b1 and b2 parameters are less than -3. It was stated by Embretson and Reise (2000) that this may be due to the low number of respondents who preferred the first response categories of these items or the fact that the item could not accurately measure the desired feature/trait. Accordingly, when the distribution of

response categories is examined, it is seen that the students who prefer the first categories are much less than the other categories.

While introducing the scales of the CLARA Inventory, it was emphasized that the low or high score obtained from the "Learning Orientation Scale" reflects a rigid persistence in the sense of not deviating from what he/she knows at one end; and reflects a dependent fragility, a feeling of being vulnerable in the slightest challenging situation at the other. For this reason, while the highest and lowest values of the levels of the feature/trait measured by the items according to both theories were reported, the values of the "Learning Orientation Scale" were ignored in order not to be misleading.

In Table 6, descriptive statistics of the levels of the features/trait measured by CLARA-TR items, estimated based on the CTT and IRT, are presented.

Table 6. *Descriptive Statistics of the Levels of the Features/Traits Measured by CLARA-Tr Items Estimated According to CTT and IRT.*

Descriptive Statistics				
	CTT	Stand. Error	IRT	Stand. Error
Minimum	2.08		-4.28	
Maximum	5.64		1.87	
$\bar{X}$	4.21	0.11	-1.06	0.17
Median	4.24		-1.14	
SD	0.77		1.20	
Kurtosis	0.338	0.67	0.65	0.67
Skewness	-0.619	0.34	-0.30	0.34
Range	3.56		6.15	
Number of Items (k)	49		49	
Number of Students	1054		1054	

The correlation between the level of the features/traits measured by the items determined according to CTT and IRT was calculated with the Spearman Rank-Differences Correlation Coefficient and it was determined that there was a negative and highly significant relationship between these two values (r = -0.830, p <0.05). If individuals prefer higher response categories while answering the items, the item score average, i.e. the value of the level of the features/traits measured by the items, increases according to the CTT. According to the IRT on the other hand, the boundary location parameter value, which is accepted as the level of the feature/trait measured by the items, decreases. The boundary location parameter is the required feature/trait level for responders to react above the limit of a response category with a probability of 0.50 (Ostini & Nering, 2006), and when individuals prefer higher response categories, the boundary location parameter, or b parameter, takes lower values. According to this result, it can be interpreted that the feature/trait levels of the items determined according to the CTT and IRT are similar to each other, and this result is consistent with the previous study results in which polytomous items statistics based on two test theories are compared (Karakılıç, 2009; Koch, 1983; Nartgün, 2002; Uysal, 2015).

3.3.3. *Research question 3.3. findings & comments – relationship between the study group's learning power levels estimated by CLARA-Tr*

Another sub-question to be answered within the scope of the third research question is "Is there a relationship between the study group's learning power levels estimated by CLARA-Tr based on CTT and IRT? For this sub-question, the relationship between the scores obtained by the students from eight Scales, which together make up the learning power profile, based on CTT and IRT was examined with the Pearson Product-Moments Correlation.

In this context, the relationship between the eight Scales that constitute the CLARA-Tr Inventory, the levels of features/traits measured according to CTT and IRT was examined. Based on the CTT, the raw scores of the students from each scale were transformed into a 100-point system with a simple formulation. In doing so, firstly, the arithmetic mean of the answers given by the students to the items in each scale was taken. Then, the base score that could be obtained from an item was subtracted from this average, and finally, this score was divided by five and multiplied by 100. According to the IRT, for each scale the trait levels of the students measured with that scale were estimated according to the Graded Response Model. Students' estimated scale scores belonging to eight scales according to CTT and IRT are given in Appendix 3. Descriptive statistics of these scores are presented in Table 7.

Table 7. *Descriptive Statistics of the Students' Scores Estimated According to CTT and IRT.*

Descriptive Statistics					
Scale		CTT	Stand. Er-ror	IRT	Stand. Er-ror
Belonging	Minimum	0.00		-1.547	
	Maximum	100.00		1.547	
	$\bar{X}$	54.22	0.91	0.134	0.025
	Median	53.33		0.072	
	SD	29.43		0.819	
	Kurtosis	-1.03	0.151	-0.542	0.151
	Skewness	-0.53	0.075	0.003	0.075
	Range	100.00		3.094	
	Number of Items (k)	3		3	
	Number of Students	1054		1054	
Collaboration	Minimum	0,00		-1.547	
	Maximum	100		1.547	
	$\bar{X}$	67.12	0.64	0.461	0.175
	Median	66.67		0.444	
	SD	20.67		0.569	
	Kurtosis	0.17	0.151	0.178	0.151
	Skewness	-0,55	0.075	-0.263	0.075
	Range	100		3.094	
	Number of Items (k)	3		3	
	Number of Students	1054		1054	
Creativity	Minimum	0.00		-2.436	
	Maximum	100		2.436	
	$\bar{X}$	65.48	0.53	0.597	0.021
	Median	65.00		0.524	
	SD	17.11		0.691	
	Kurtosis	-0.38	0.151	0.153	0.151
	Skewness	-0.17	0.075	0.307	0.075
	Range	100		4.872	
	Number of Items (k)	8		8	
	Number of Students	1054		1054	

Table 7. *Continues*

Curiosity	Minimum	0.00		-2.175	
	Maximum	100		2.175	
	$\bar{X}$	67.34	0.58	0.623	0.022
	Median	70.00		0.607	
	SD	18.75		0.702	
	Kurtosis	-0.45	0.151	-0.277	0.151
	Skewness	-0.36	0.075	0.035	0.075
	Range	100		4.350	
	Number of Items (k)	6		6	
	Number of Students	1054		1054	
Hope & Optimism	Minimum	0.00		-1.547	
	Maximum	100		1.547	
	$\bar{X}$	70.49	0.70	0.553	0.020
	Median	73.33		0.548	
	SD	22.80		0.645	
	Kurtosis	-0.26	0.151	-0.261	0.151
	Skewness	-0.59	0.075	-0.197	0.075
	Range	100		3.094	
	Number of Items (k)	3		3	
	Number of Students	1054		1054	
Mindful Agency	Minimum	0.00		-2.542	
	Maximum	100.00		2.542	
	$\bar{X}$	68.11	0.50	0.709	0.021
	Median	68.89		0.675	
	SD	16.16		0.682	
	Kurtosis	0.00	0.151	0.417	0.151
	Skewness	-0.37	0.075	0.204	0.075
	Range	100.00		5.084	
	Number of Items (k)	9		9	
	Number of Students	1054		1054	
Orientation to Learning	Minimum	0,00		-2.637	
	Maximum	100		2.289	
	$\bar{X}$	49.10	0.46	0.033	0.20
	Median	48.89		0.000	
	SD	14.91		0.644	
	Kurtosis	0.01	0.151	0.823	0.151
	Skewness	0.16	0.075	0.092	0.075
	Range	100		4.926	
	Number of Items (k)	10		10	
	Number of Students	1054		1054	
Sense Making	Minimum	0,00		-2.315	
	Maximum	100		2.315	
	$\bar{X}$	74.12	0.39	0.897	0.017
	Median	74.29		0.846	
	SD	12.59		0.546	
	Kurtosis	0.91	0.151	0.932	0.151
	Skewness	0.45	0.075	0.150	0.075
	Range	100		4.630	
	Number of Items (k)	7		7	
	Number of Students	1054		1054	

When the values in Table 7 and Appendix 3 were examined together, it was determined that there is a similarity between the students' levels of the features/traits participating in the study, which were estimated based on both theories for all scales. It was observed that a student who got a lower score from a scale according to CTT had a similarly low score estimated according to IRT.

Within the scope of the third research question of the study, the relationship between the scores obtained by the students from eight scales estimated according to CTT and IRT was examined with Pearson Product-Moments Correlation Coefficient. Correlation values are presented in Table 8.

Table 8. *Relationship Between Students CLARA-Tr Scores Estimated from CTT and IRT.*

Scales (CTT / IRT)	n=1054	
	r	*p*
Belonging	0.992	0.000
Collaboration	0.991	0.000
Creativity	0.983	0.000
Curiosity	0.986	0.000
Hope & Optimism	0.987	0.000
Mindful Agency	0.979	0.000
Orientation to Learning	0.975	0.000
Sense Making	0.973	0.000

When Table 8 is examined, it is seen that there is a positive, high and significant (r = 0.973-0.992, p <.01) relationship between the scores of the students obtained from CLARA-Tr's eight scales estimated based on CTT and IRT. Based on these correlation coefficients, it can be inferred that the scores estimated according to both theories are similar and comparable.

3.3.4. *Research question 3.4. findings – reliability of clara-tr*

The last answer will be sought within the scope of the third research question is "How is the reliability of CLARA-Tr according to the Classical Test Theory and Item Response Theory?" For this sub-question; the reliability of the instrument was determined by calculating the Cronbach alpha internal consistency coefficient value according to the CTT and the marginal reliability coefficient according to the IRT.

Each reliability levels of the eight scales in the Turkish form of CLARA were examined both according to CTT and IRT. Calculated Cronbach alpha internal consistency coefficients and marginal reliability coefficients are summarized in Table 9.

Table 9. *Cronbach Alpha Internal Consistency & Marginal Reliability Coefficients of CLARA-Tr.*

Scale	Number of Items (k)	Cronbach Alpha	McDonald's omega	Marginal Reliability
Belonging	3	0.871	0.871	0.874
Hope & Optimism	3	0.833	0.869	0.873
Mindful Agency	9	0.812	0.813	0.842
Creativity	8	0.790	0.795	0.814
Curiosity	6	0.785	0.806	0.850
Sense Making	7	0.754	0.759	0.756
Orientation to Learning	10	0.742	0.741	0.834
Collaboration	3	0.730	0.734	0.721

When Table 9 was examined, it was seen that the Cronbach alpha reliability coefficient for all scales varied between 0.871 and 0.730, and McDonald's omega coefficients varied between 0.871 and 0.734. For scales, reliability coefficient values above 0.70 are accepted as high reliability levels (Nunnally, 1978; Özdamar, 2013). According to these values, it can be said

that all scales are consistent within themselves and have a high level of reliability according to CTT. When the marginal reliability coefficients estimated according to IRT are examined, it is seen that these values change between 0.874 and 0.721. The marginal reliability coefficient is defined as the arithmetic mean of the reliability coefficients estimated separately for the different levels of the measured psychological feature/trait (Thissen, 1991; Flannery et al., 1995). In this respect, the marginal reliability coefficient is accepted as a reliability coefficient calculated for the whole of a measurement tool. The high value of this coefficient is an indication that the results obtained from the measurement tool used are reliable. It is seen that the reliability coefficient values estimated according to both theories presented in Table 9 are quite high and similar to each other. These values can be interpreted as all eight scales of CLARA-Tr can make reliable measurements.

4. DISCUSSION and CONCLUSION

In this study, it was aimed to adapt Crick Learning for Resilient Agency (CLARA) Inventory to Turkish culture, also to analyze and compare the psychometric properties of the Inventory in the adaptation process according to the methods and techniques of both Classical Test Theory (CTT) and Item Response Theory (IRT). The results achieved are listed below in items.

1. It was determined that there is a positive, high and significant relationship between the scale scores obtained from the English and Turkish forms of CLARA's language equivalence applications. Based on this finding, it was accepted that linguistic equivalence was provided between the original form of CLARA and its Turkish form.

2. Two separate confirmatory factor analyses were conducted on the data obtained from the pilot application to determine whether the eight-scale original structure of the Inventory was also verified by Turkish university students. It was decided that the original factor structure of the Inventory was also verified in Turkish undergraduate students, and that data on learning power could be collected from university students in a valid and reliable manner with the Inventory.

3. Before starting the analysis to determine the psychometric properties of CLARA-Tr according to CTT and IRT, it was tested whether the data obtained as a result of the pilot application provided the assumptions of both theories. As a result of meeting the assumptions, analyses were carried out regarding the research questions.

4. The item discrimination index (item corrected total score correlation and a parameter) of the items of CLARA-Tr was estimated according to the Classical Test Theory and Item Response Theory, and a decent level discrimination index values were obtained according to both theories.

5. The levels of the features/traits (arithmetic mean and b parameter) measured by the items that constitute the Inventory were determined according to both test theories. It has been determined that the items generally measure the features/traits to be measured with the scales they belong to at a high level and all items have a relatively high approval rate, in other words, the participants have responded to high-level categories. The relationship between the levels of the features/traits measured by the items determined according to CTT and IRT was examined and a highly significant negative relationship was found. According to this result, it has been interpreted that the levels of the features/traits measured by the items determined according to CTT and IRT are similar to each other.

6. The relationship between the estimated scores, based on both test theories, of the undergraduates' obtained from eight scales, which together constitute the learning power profile, was examined. In this context, a high level of relationship was found between undergraduates' learning power levels predicted according to both theories, and from this point of view, it was concluded that the scores obtained from the two theories were similar.

7. The reliability of the scales that constitute the CLARA-Tr has been examined by calculating the Cronbach alpha internal consistency coefficient according to the CTT, and by the marginal reliability coefficient according to the IRT. It has been observed that the reliability coefficient values estimated according to both theories are quite high and similar to each other. As a result of these values, it was concluded that all eight scales of CLARA-Tr make reliable measurements.

8. It can be said that with CLARA-Tr, obtained by adapting CLARA, a valid and reliable tool has been provided to the Turkish literature to be used in the future studies.

The recommendations made as a result of the adaptation process and the comparisons made according to different test theories in this process are presented below in items.

1. Within the scope of this study, it is revealed that CLARA-Tr, whose psychometric properties were examined by adapting into Turkish culture, make valid and reliable measurements according to both test theories; and in the light of this result, it can be said that researchers who aim to reveal the undergraduate students' learning power profiles will be able to use the Inventory.

2. Whether the item parameters of CLARA-Tr show invariance between different samples according to both test theories can be discussed in a separate study.

3. In order to compare with the results of this study, the Inventory can be applied to samples of different sizes and different characteristics.

4. Within the scope of this study, it was determined that the values of the psychometric properties of the Inventory estimated according to both theories were similar. In this context, studies can be carried out on the basis of both theories as currently applied in scale development studies. On the other hand, it is recommended that researchers who want to reach more explanatory information at the item and test level should especially prefer the IRT. The fact that IRT gives different error estimates at different levels of the psychological feature/trait to be measured, and that items which give information with higher precision can be selected, will enable researchers to develop scales suitable for their purposes.

The CLARA learning power profile tool was designed to enable an individual learner to develop their capacity for self-leadership in learning which is a crucial 21st Century life competence (Sala et al., 2020). It was, at the same time, a deliberate attempt on the part of researchers to challenge the dominant ‘performativity’ discourse in educational assessment (Broadfoot, 1998). The accuracy, reliability and validity of the measurement model as reported here provides the foundation for this personal, social and political development, supported most effectively through coaching relationships. Since the first learning power model was developed in 2002 there has been significant user led demand for the tool which has been and practiced extensively in education, community and corporate contexts around the world, for example (Crick and Bentley, 2020).

However, it also brings with it the inherent challenge of forging pathways to impact for research outputs, moving beyond academia into digital learning analytics and also into practice led improvement in different contexts. Such pathways to impact require new business models which can integrate the differing requirements, funding mechanisms and lifecycles of research, policy, practice and commercial enterprise. The digital capability for the assessment tool, built on a data architecture which has a ‘single view of the learner’, uses one data point to provide rapid feedback to the individual, the team and the organisation as well as raw data for ongoing research. This is beyond the traditional capacity of a single research or educational institution and requires ethical quality assurance derived from a not for profit entity, funding for user services as well as digital entrepreneurship in a world which tends towards an individualist and reductionist ideology and practice. Twenty-one years of experience have led to the current

business model – which also provided the basis for this research study. The next steps are to take CLARA-Tr and explore whether and how it can add value in practice, through Work Integrated Learning Design (WILD) in Türkiye.

Declaration of Conflicting Interests and Ethics

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

Authorship Contribution Statement

This study has been generated from the doctoral dissertation named "Adaptation of CLARA Inventory to Turkish Culture and Investigation of the Factors Affecting Learning Power with Hierarchical Linear Models" which was submitted by **Hasan Fehmi OZDEMIR** under the supervision of **Omer KUTLU** at Ankara University, Institute of Educational Sciences. The adapted Inventory was originally developed by **Ruth CRICK** and revised by **Shaofu HUANG** and **Ruth CRICK**. They have also contributed to the translation, online data gathering and writing the original draft processes.

Orcid

Hasan Fehmi Ozdemir https://orcid.org/0000-0002-0705-8032
Omer Kutlu https://orcid.org/0000-0003-4364-5629
Shaofu Huang https://orcid.org/0000-0002-9887-2434
Ruth Crick https://orcid.org/0000-0003-1082-9560

REFERENCES

Arthur, James & Crick, Ruth & Samuel, Elspeth & Wilson, Kenneth & Mcgettrick, Bart. (2006). *Character Education: The Formation of Virtues and Dispositions in 16-19 Year Olds with particular reference to the religious and spiritual*. West Conshohocken, PA: John Templeton Foundation. Retrieved from: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.473.7540&rep=rep1&type=pdf

Broadfoot, P. (1998). Records of Achievement and the Learning Society: a tale of two discourses. *Assessment in Education, 5*(3), 447-477. https://doi.org/10.1080/0969595980050307

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö.E., Karadeniz, Ş., & Demirel, F. (2014). *Bilimsel Araştırma Yöntemleri* [*Scientific Research Methods*]. Pegem Akademi.

Carr, M., & Claxton, G. (2002). Tracking the development of learning dispositions. *Assessment in Education*, 9 (1), pp. 9-37. https://doi.org/10.1080/09695940220119148

Claxton, G. (1999). *Wise Up: The Challenge of Lifelong Learning*. Bloomsbury.

Cohen, R.J., & Swerdlik, M.E. (2010). *Psychological testing and assessment*. McGrawHill Companies.

Crick, R., & Bentley, J. (2020). Becoming a resilient organisation: integrating people and practice in infrastructure services. *International Journal of Sustainable Engineering, 13*(6), 423-440. https://doi.org/10.1080/19397038.2020.1750738

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt, New York.

Cronbach, L.J. (1960). *Essentials of psychological testing*. Harper Collins Publishers.

Deakin Crick, R. (2005). *Learning Power: Dynamic Assessment for Learning*. The Leader: Secondary Heads Association.

Deakin Crick, R. (2006). *Learning power in practice: A guide for teachers*. Paul Chapman.

Deakin Crick, R. (2007). Learning how to learn: the dynamic assessment of learning power. *Curriculum Journal, 18*(2), 135-153. https://doi.org/10.1080/09585170701445947

Deakin Crick, R. (2009a). Assessment in Schools - Dispositions. In: McGaw, B., Peterson, P., & Baker, E. (eds.) *The International Encyclopedia of Education* (3rd Edition). Elsevier.

Deakin Crick, R. (2009b). Inquiry-based learning: reconciling the personal with the public in a democratic and archaeological pedagogy. *Curriculum Journal*, 20(1), 73-92. https://doi.org/10.1080/09585170902764021

Deakin Crick, R., Broadfoot, P., & Claxton, G. (2004). Developing an effective lifelong learning inventory: The ELLI Project. *Assessment in Education*, *11*(3), 248-272. https://doi.org/10.1080/0969594042000304582

Deakin Crick, R., & McCombs, B. (2006). The Assessment of Learner Centered Principles: An English Case Study. *Educational Research and Evaluation*, *12*(5), 423-444. https://doi.org/10.1080/13803610600697021

Deakin Crick, R., & Salway, A. (2006). *Locked Up Learning: learning power and young offenders*. ViTaL Partnerships Ltd.

Deakin Crick, R., & Wilson, K. (2005). Being a Learner: A Virtue for the 21st Century. *British Journal of Educational Studies*, *53*(5), 359-374. https://doi.org/10.1111/j.1467-8527.2005.00300.x

Deakin Crick, R., & Yu, G. (2008). Assessing learning dispositions: is the effective lifelong learning inventory valid and reliable as a measurement tool? *Educational Research*, *50*(4), 387-402. https://doi.org/10.1080/00131880802499886

Deakin Crick, R., Haigney, D., Huang, S., Coburn, T., & Goldspink, C. (2013). Learning power in the workplace: the effective lifelong learning inventory and its reliability and validity and implications for learning and development. *The International Journal of Human Resource Management*, *24*(11), 2255-2272. https://doi.org/10.1080/09585192.2012.725075

Deakin Crick, R., Huang, S., Shafi, A.A., & Goldspink, C. (2015). Developing resilient agency in learning: The internal structure of learning power. *British Journal of Educational Studies*, *63*(2), 121-160. https://doi.org/10.1080/00071005.2015.1006574

De Ayala, R.J. (2009). The theory and practice of item response theory. U.S.A.: Guilford Press.

Diker Coşkun, Y., & Demirel, M. (2012). Üniversite Öğrencilerinin Yaşam Boyu Öğrenme Eğilimleri [Lifelong Learning Dispositions of University Students]. *Hacettepe University Journal of Education*, 42, 108-120.

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

Erkuş, A. (2013). *Davranış Bilimleri İçin Bilimsel Araştırma Süreci* [*Scientific Research Process for Behavioral Sciences*]. Seçkin Yayıncılık.

Flannery, W.P., Reise, S.P., & Widaman, F.K. (1995). An item response theory analysis of the general and academic scales of self-description questionnaire-II. *Journal of Research in Personality*, *29*(1), 168-188.

Gornall, S., Chambers, M.R., & Claxton, G. (2005). Building Learning Power in Action. TLO Ltd. Retrieved from: https://www.buildinglearningpower.com/wp-content/uploads/2014/11/BLPIA_A4_v1l-sample.pdf

Hambleton, R., & Swaminathan, R. (1985). *Fundamentals of Item Response Theory*. Sage Publishers.

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

Hambleton, R.K. & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology*, *1*(1), 1-12.

Hambleton, R.K., Merenda, P.F., & Spielberger, C.D. (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Lawrence Erlbaum.

Harlen, W., & Deakin Crick, R. (2003a). *A systematic review of the impact of summative assessment and testing on pupils' motivation for learning*. Retrieved from: https://dspace.stir.ac.uk/bitstream/1893/19607/1/SysRevImpSummativeAssessment2002.pdf

Harlen, W., & Deakin Crick, R. (2003b). Testing and Motivation for Learning. *Assessment in Education, 10*(2), 169-207. https://doi.org/10.1080/0969594032000121270

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

International Test Commission, ITC. (2018). Guidelines for translating and adapting tests (Second edition). *International Journal of Testing, 18*(2), 101-134.

Jöreskog, K.G., & Sörbom, D. (1993). *Lisrel 8: structural equation modeling with the simplis command language*. Erlbaum Associates Publishers.

Karakılıç, M. (2009). *Beden eğitimi dersi için hazırlanan tutum ölçeğinin psikometrik kuramlar açısından incelenmesi* [*Examination of the attitude scale prepared for the physical education lesson in terms of psychometric theories*] [Unpublished Doctoral Thesis]. Ankara University.

Kelecioğlu, H. (2001). Örtük Özellikler Teorisindeki b ve a Parametreleri İle Klasik Test Teorisindeki p ve r İstatistikleri Arasındaki İlişki [The relationship between b ve a parameters in latent trait theory and p and r statistiscs in classical test theory]. *Hacettepe University Journal of Education, 20*(1), 104-110.

Kline, R.B. (2005). *Principles and practice of structural equation modeling*. The Guilford Press.

Koch, W.R. (1983). Likert scaling using the graded response latent trait model. *Applied Psychological Measurement, 7*(1), 15-32. https://doi.org/10.1177/014662168300700104

Masten, A. (2007). Resilience in developing systems: progress and promise as the fourth wave rises. *Developmental Psychopathology, 19*(3), 921-930. https://doi.org/10.1017/S0954579407000442

MoNE (2014). Türkiye hayat boyu öğrenme strateji belgesi ve eylem planı 2014-2018 [Türkiye Lifelong Learning Strategy Document and Action Plan 2014-2018]. Retrieved from: http://www.resmigazete.gov.tr/eskiler/2014/07/20140716-8-1.pdf

MoNE (2018). *2023 Eğitim Vizyon Belgesi* [*2023 Education Vision Document*]. Retrieved from: http://2023vizyonu.meb.gov.tr/

Nartgün, Z. (2002). *Aynı Tutumu Ölçmeye Yönelik Likert Tipi Ölçek ile Metrik Ölçeğin Madde ve Ölçek Özelliklerinin Klasik Test Kuramı ve Örtük Özellikler Kuramına Göre İncelenmesi* [*The investigation of item and scale properties of likert type scale and metric scale measuring the same attitude according to classisical test theory and item response theory*] [Unpublished PhD Thesis]. Hacettepe University.

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

Ostini, R., & Nering, M.L. (Eds.). (2006). *Polytomous item response theory models*. Sage Publication.

Özdamar, K. (2013). *Paket Programlar ile İstatistiksel Veri Analizi* [*Statistical Data Analysis with Package Programs*] (9. Edition). Nisan Kitabevi.

Özdemir, H.F., Toraman, Ç., & Kutlu, Ö. (2019). The use of polychoric and Pearson correlation matrices in the determination of construct validity of Likert type scales. *Turkish Journal of Education, 8*(1), 180-195. https://doi.org/10.19128/turje.519235

Reise, S.P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*(2), 133-144. https://doi.org/10.1111/j.1745-3984.1990.tb00738.x

Rutter, M. (1985). Resilience in the face of adversity. Protective factors and resistance to psychiatric disorder. *British Journal of Psychiatry*, *147*(6), 598-611. https://doi.org/10.1192/bjp.147.6.598

Rutter, M. (2012). Resilience as a Dynamic Concept. *Development and Psychopathology*, 24 335-344. https://doi.org/10.1017/S0954579412000028

Sala, A., Punie, Y., Garkov, V., & Cabrera Giraldez, M. (2020). LifeComp: The European Framework for Personal, Social and Learning to Learn Key Competence. Luxembourg: Publications Office of the European Union. https://data.europa.eu/doi/10.2760/302967

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(4), 5-17. https://www.psychometricsociety.org/sites/main/files/file-attachments/mn17.pdf

Samejima, F. (1996). *Polychotomous responses and the test score*. Tennessee: The University of Tennessee.

Savaşır, I. (1994). Ölçek Uyarlamasındaki Sorunlar ve Bazı Çözüm Yolları [Problems in Scale Adaptation and Some Solutions]. *Turkish Journal of Psychology*, *9*(33), 27-32. https://www.psikolog.org.tr/tr/yayinlar/dergiler/1031828/tpd13004433199400000m000295.pdf

Scott, D., & Usher, R. (2011). *Researching education: Data, methods and theory in educational enquiry*. Continuum.

Sümer, N. (2000). Yapısal Eşitlik Modelleri: Temel Kavramlar ve Örnek Uygulamalar [Structural Equation Models: Basic Concepts and Applications]. *Turkish Psychological Articles*, *3*(6), 49-73. https://www.psikolog.org.tr/tr/yayinlar/dergiler/1031828/tpy1301996120000000m000225.pdf

Şimşek, Ö.F. (2007). *Yapısal Eşitlik Modellemesine Giriş: Temel İlkeler ve LISREL Uygulamaları* [*Introduction to Structural Equation Modeling: Fundamentals and Applications of LISREL*]. Ekinoks Yayınları.

The Assessment Reform Group. (2010, December). Retrieved from https://www.nuffieldfoundation.org/project/the-assessment-reform-group

Thissen, D. (1991). *Multilog user's guide*. Scientific Software

Tunca, N., Alkın Şahin, S., & Aydın, Ö. (2015). Öğretmen adaylarının yaşam boyu öğrenme eğilimleri. *Mersin Üniversitesi Eğitim Fakültesi Dergisi*, *11*(2), 432-446. https://doi.org/10.17860/efd.92694

Uyar, Ş., Öztürk Gübeş, N., & Kelecioğlu H. (2013). PISA 2009 Tutum Anketi Madde Puanlarının Aşamalı Tepki Modeli ile İncelenmesi [Examinig the Pisa 2009 Attitude Questionnaire Item Scores with Using Graded Response Model]. *Journal of Research in Education and Teaching*, *2*(4), 125-134. http://www.jret.org/FileUpload/ks281142/File/15.uyar.pdf

Uysal, M. (2015). *Araştırma özyeterlik ölçeğinin psikometrik özelliklerinin klasik test kuramı ve madde tepki kuramına göre incelenmesi* [*An investigation of psychometric properties of research self-efficacy scale according to Classical Test Theory and Item Response Theory*] [Unpublished Master Thesis]. Gazi University.

Vieira, A.L. (2011). *Preparation of the analysis. Interactive LISREL in practice* (First Edition). Springer.

APPENDIX

Appendix 1. An example of the MULTILOG program output (Belonging Scale)

MULTILOG--FOR MULTIPLE CATEGORICAL ITEM RESPONSE DATA--VERSION 7.0.3

MULTILOG for Windows 7.00.2327.2

Created on: 19 September 2018, 12:55:24

```
>PROBLEM RANDOM,

      INDIVIDUAL,

      DATA = 'C:\Users\kullanici\Desktop\ait\ait.DAT',

      NITEMS = 3,

      NGROUPS = 1,

      NEXAMINEES = 1054,

      NCHARS = 4;


 DATA FILE NAME IS
      C:\USERS\KULLANICI\DESKTOP\AIT\AIT.DAT


TYPE OF INPUT:
  INDIVIDUAL RESPONSE VECTORS


>TEST ALL,

    GRADED,

    NC = (6(0)3);

NUMBER OF CODES  6
```

123456
VECTOR OF CATEGORIES FOR CODE=1

111
VECTOR OF CATEGORIES FOR CODE=2

222
VECTOR OF CATEGORIES FOR CODE=3

333
VECTOR OF CATEGORIES FOR CODE=4

444
VECTOR OF CATEGORIES FOR CODE=5

555
VECTOR OF CATEGORIES FOR CODE=6

666
(4A1,T5,3A1)

MULTILOG--FOR MULTIPLE CATEGORICAL ITEM RESPONSE DATA--VERSION 7.0.3
MULTILOG for Windows 7.00.2327.2
Created on: 19 September 2018, 12:55:24

DATA PARAMETERS:
NUMBER OF LINES IN THE DATA FILE: 1054
NUMBER OF CATEGORICAL-RESPONSE ITEMS: 3
NUMBER OF CONTINUOUS-RESPONSE ITEMS, AND/OR GROUPS: 1
TOTAL NUMBER OF "ITEMS" (INCLUDING GROUPS): 4
NUMBER OF CHARACTERS IN ID FIELDS: 4
MAXIMUM NUMBER OF RESPONSE-CODES FOR ANY ITEM: 6
THE MISSING VALUE CODE FOR CONTINUOUS DATA: 9.0000
THE DATA WILL BE STORED IN MEMORY

ESTIMATION PARAMETERS:

THE ITEMS WILL BE CALIBRATED--
BY MARGINAL MAXIMUM LIKELIHOOD ESTIMATION
MAXIMUM NUMBER OF EM CYCLES PERMITTED: 25
NUMBER OF PARAMETER-SEGMENTS USED IS: 3
NUMBER OF FREE PARAMETERS IS: 18
MAXIMUM NUMBER OF M-STEP ITERATIONS IS 4 TIMES
THE NUMBER OF PARAMETERS IN THE SEGMENT
THE M-STEP CONVERGENCE CRITERION IS: 0.000100
THE EM-CYCLE CONVERGENCE CRITERION IS: 0.001000
THE RK CONTROL PARAMETER (FOR THE M-STEPS) IS: 0.9000
THE RM CONTROL PARAMETER (FOR THE M-STEPS) IS: 1.0000
THE MAXIMUM ACCELERATION PERMITTED IS: 0.0000
THETA-GROUP LOCATIONS WILL REMAIN UNCHANGED

QUADRATURE POINTS FOR MML,
AT THETA:
-4.500
-4.000
-3.500
-3.000
-2.500
-2.000
-1.500
-1.000
-0.500
0.000
0.500
1.000
1.500
2.000
2.500
3.000
3.500
4.000
4.500

MULTILOG for Windows 7.00.2327.2

READING DATA...

KEY-

CODE CATEGORY

1 111
2 222
3 333
4 444
5 555
6 666

FORMAT FOR DATA-

(4A1,T5,3A1)

FIRST OBSERVATION AS READ-

ID 0001
ITEMS 555
NORML 0.000

FINISHED CYCLE 25
MAXIMUM INTERCYCLE PARAMETER CHANGE= 0.00344 P(7)

ITEM SUMMARY

MULTILOG for Windows 7.00.2327.2

ITEM 1: 6 GRADED CATEGORIES
  P(#) ESTIMATE (S.E.)
A 1 4.12 (0.17)
B(1) 2 -1.38 (0.06)

B(2) 3 -0.68 (0.03)
B(3) 4 -0.14 (0.03)
B(4) 5 0.18 (0.03)
B(5) 6 0.72 (0.04)

@THETA: INFORMATION: (Theta values increase in steps of 0.2)
-3.0 - -1.6 0.021 0.048 0.109 0.245 0.538 1.131 2.165 3.499
-1.4 - 0.0 4.406 4.409 4.279 4.602 4.825 4.837 5.043 5.202
 0.2 - 1.6 5.090 4.850 4.754 4.311 3.126 1.819 0.918 0.430
 1.8 - 3.0 0.194 0.086 0.038 0.017 0.007 0.003 0.001

OBSERVED AND EXPECTED COUNTS/PROPORTIONS IN
CATEGORY(K): 1 2 3 4 5 6
OBS. FREQ. 104 171 201 126 192 260
OBS. PROP. 0.0987 0.1622 0.1907 0.1195 0.1822 0.2467
EXP. PROP. 0.1031 0.1627 0.1825 0.1176 0.1799 0.2541

ITEM 2: 6 GRADED CATEGORIES
P(#) ESTIMATE (S.E.)
A 7 4.46 (0.20)
B(1) 8 -1.35 (0.05)
B(2) 9 -0.76 (0.04)
B(3) 10 -0.16 (0.03)
B(4) 11 0.22 (0.03)
B(5) 12 0.70 (0.04)
@THETA: INFORMATION: (Theta values increase in steps of 0.2)
-3.0 - -1.6 0.013 0.031 0.074 0.180 0.428 0.981 2.076 3.724
-1.4 - 0.0 5.105 5.336 5.279 5.540 5.383 5.300 5.731 5.967
 0.2 - 1.6 5.944 5.739 5.606 4.927 3.317 1.758 0.810 0.349
 1.8 - 3.0 0.146 0.060 0.025 0.010 0.004 0.002 0.001

OBSERVED AND EXPECTED COUNTS/PROPORTIONS IN
CATEGORY(K): 1 2 3 4 5 6
OBS. FREQ. 106 141 218 152 174 263
OBS. PROP. 0.1006 0.1338 0.2068 0.1442 0.1651 0.2495
EXP. PROP. 0.1053 0.1353 0.1996 0.1404 0.1622 0.2571

ITEM 3: 6 GRADED CATEGORIES
P(#) ESTIMATE (S.E.)
A 13 2.29 (0.11)
B(1) 14 -1.51 (0.08)
B(2) 15 -0.49 (0.05)
B(3) 16 0.21 (0.05)
B(4) 17 0.66 (0.05)
B(5) 18 1.27 (0.08)

@THETA: INFORMATION: (Theta values increase in steps of 0.2)
-3.0 - -1.6 0.162 0.246 0.369 0.535 0.746 0.982 1.204 1.366
-1.4 - 0.0 1.444 1.461 1.467 1.495 1.540 1.579 1.605 1.626
0.2 - 1.6 1.647 1.660 1.659 1.641 1.602 1.524 1.386 1.184
1.8 - 3.0 0.943 0.704 0.500 0.342 0.227 0.149 0.096

OBSERVED AND EXPECTED COUNTS/PROPORTIONS IN
CATEGORY(K): 1 2 3 4 5 6
OBS. FREQ. 120 252 243 142 140 157
OBS. PROP. 0.1139 0.2391 0.2306 0.1347 0.1328 0.1490
EXP. PROP. 0.1158 0.2314 0.2198 0.1344 0.1422 0.1564

ITEM 4: GRP1, N[MU: 0.00 SIGMA: 1.00]
P(#);(S.E.): 20; (0.00) 21; (0.00)

@THETA: INFORMATION: (Theta values increase in steps of 0.2)
-3.0 - -1.6 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000
-1.4 - 0.0 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000
0.2 - 1.6 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000
1.8 - 3.0 1.000 1.000 1.000 1.000 1.000 1.000 1.000

TOTAL TEST INFORMATION

@THETA: INFORMATION:
-3.0 - -1.6 1.195 1.325 1.552 1.960 2.712 4.094 6.446 9.589
-1.4 - 0.0 11.956 12.206 12.025 12.637 12.749 12.717 13.379 13.795

0.2 - 1.6 13.681 13.249 13.019 11.879 9.045 6.101 4.114 2.962
1.8 - 3.0 2.283 1.851 1.563 1.369 1.239 1.154 1.098

@THETA: POSTERIOR STANDARD DEVIATION:
-3.0 - -1.6 0.915 0.869 0.803 0.714 0.607 0.494 0.394 0.323
-1.4 - 0.0 0.289 0.286 0.288 0.281 0.280 0.280 0.273 0.269
0.2 - 1.6 0.270 0.275 0.277 0.290 0.333 0.405 0.493 0.581
1.8 - 3.0 0.662 0.735 0.800 0.855 0.898 0.931 0.954
MARGINAL RELIABILITY: 0.8741

NEGATIVE TWICE THE LOGLIKELIHOOD= -5344.0
(CHI-SQUARE FOR SEVERAL TIMES MORE EXAMINEES THAN CELLS)

NORMAL PROGRAM TERMINATION

START DATE: 09-19-2018
START TIME: 12:58:28
END TIME: 12:58:29

Appendix 2. Sample scale items of CLARA and CLARA-Tr

Mindful Agency

I know I can find a way of solving a problem if I have enough time to think.
(Düşünmek için yeterli zamanım olursa, karşılaştığım sorunu çözmenin bir yolunu bulabilirim.)
I think about everything that I will need before I begin a task.
(Bir işe girişmeden önce ihtiyaç duyacağım her şey hakkında düşünürüm.)

Hope and Optimism

I know I am changing and growing over time.
(Zamanla değiştiğimi ve geliştiğimi biliyorum.)
I am getting better at learning all the time.
(Öğrenme işinde sürekli daha iyiye gidiyorum.)

Sense Making

I make connections between what I am learning and what I have learned before.
(Yeni öğrendiğim şeylerle, önceden öğrendiklerim arasında bağlantı kurarım.)
I often look back and think about what I have learned.
(Öğrenmiş olduğum şeyler hakkında sıkça geçmişi hatırlar ve düşünürüm.)

Creativity

Sometimes good ideas just come into my head.

(Bazen, güzel fikirler ansızın aklıma geliverir.)

I tend to use my imagination to help me learn.

(Öğrenmeme yardımcı olması için hayal gücümü kullanma eğilimindeyimdir.)

Curiosity

I prefer learning something when I have to try really hard to understand it.

(Gerçekten çok çaba harcayarak anlayabileceğim şeyleri öğrenmeyi tercih ederim.)

I am more stimulated by interesting questions than easy answers.

(İlginç sorular, kolay cevaplara göre beni daha çok teşvik eder.)

Collaboration

I enjoy solving problems together with other people.

(Sorunları diğer insanlarla birlikte çözmekten hoşlanırım.)

I find it helps me to learn if I can talk about it with colleagues.

(Arkadaşlarımla, zorlayıcı sorunlar hakkında ayrıntılı bir şekilde tartışmayı severim.)

Belonging

There is at least one person close to me who has helped me to learn.

(Ben öğrenirken yardım etmiş olan bana yakın en az bir kişi var.)

I have at least one person close to me who I can turn to for guidance in my learning.

(Öğrenirken beni yönlendirmesi için başvurabileceğim, bana yakın en az bir kişi var.)

Orientation to Learning

I find it difficult to know what to do when I get stuck.

(Bir konuya takılıp kaldığımda ne yapacağımı bilmekte zorlanırım.)

Because I dislike feelings of confusion and uncertainty I generally steer clear of learning something new.

(Kafa karışıklığı ve belirsizlik duygularını sevmediğimden, genellikle yeni bir şey öğrenmekten kaçınırım.)

Appendix 3. Students' estimated scale scores belonging to eight scales according to CTT and IRT

Appendix 3. has been given as a separate document due to the number of pages.