e-ISSN: 2148-7456

a peer-reviewed
online journal

hosted by DergiPark

International Journal of Assessment Tools in Education

Volume: 11 Issue: 2 June 2024

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

IJATE

e-ISSN: 2148-7456

Volume 11 Issue 2 June 2024

Editor Dr. Omer KUTLU

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

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

Publisher Dr. Izzet KARA

Address Pamukkale University, Education Faculty, Kinikli Campus, 20070 Denizli, Türkiye

Phone +90 258 296 1036

Fax +90 258 296 1200

E-mail ikara@pau.edu.tr
ijate.editor@gmail.com

Journal Contact Dr. Eren Can AYBEK

Address Department of Educational Sciences, Pamukkale University, Faculty of Education, Kinikli Yerleskesi, Denizli, 20070, Türkiye

Phone +90 258 296 31050

Fax +90 258 296 1200

E-mail erencanaybek@gmail.com

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

Frequency 4 issues per year (March, June, September, December)

Online ISSN 2148-7456

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

Cover Design IJATE

International Journal of Assessment Tools in Education (IJATE) is a peer-reviewed and academic online journal. The scientific and legal responsibility for manuscripts published in our journal belongs to the authors(s).

International Journal of Assessment Tools in Education

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

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

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

Indexes and Platforms:

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

CONTENTS

Research Articles

International Journal of Assessment Tools in Education

2024, Vol. 11, No. 2, 213–248

https://doi.org/10.21449/ijate.1290831

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

Comparison of item response theory ability and item parameters according to classical and Bayesian estimation methods

Eray Selçuk [1*], **Ergül Demir** [2]

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

ARTICLE HISTORY

Received: May 01, 2023

Accepted: Feb. 13, 2024

Keywords:

IRT parameter estimation,

Maximum likelihood estimation,

Bayesian estimation method,

MCMC,

RMSE.

Abstract: This research aims to compare the ability and item parameter estimations of Item Response Theory according to Maximum likelihood and Bayesian approaches in different Monte Carlo simulation conditions. For this purpose, depending on the changes in the priori distribution type, sample size, test length, and logistics model, the ability and item parameters estimated according to the maximum likelihood and Bayesian method and the differences in the RMSE of these parameters were examined. The priori distribution (normal, left-skewed, right-skewed, leptokurtic, and platykurtic), test length (10, 20, 40), sample size (100, 500, 1000), logistics model (2PL, 3PL). The simulation conditions were performed with 100 replications. Mixed model ANOVA was performed to determine RMSE differentiations. The prior distribution type, test length, and estimation method in the differentiation of ability parameter and RMSE were estimated in 2PL models; the priori distribution type and test length were significant in the differences in the ability parameter and RMSE estimated in the 3PL model. While prior distribution type, sample size, and estimation method created a significant difference in the RMSE of the item discrimination parameter estimated in the 2PL model, none of the conditions created a significant difference in the RMSE of the item difficulty parameter. The priori distribution type, sample size, and estimation method in the item discrimination RMSE were estimated in the 3PL model; the a priori distribution and estimation method created significant differentiation in the RMSE of the lower asymptote parameter. However, none of the conditions significantly changed the RMSE of item difficulty parameters.

1. INTRODUCTION

Test development consists of sequential activities (Thorndike, 1982). Test development processes are carried out within the framework of various theories aimed at minimizing error. In this context, test theories use various methods and models to ensure the reliability and validity of the measurement process. Test theories are an overview that connects observed variables to latent variables. The general purpose of test theories is to estimate the true score. While making this estimation, it is also to determine how much the measurement scores of the defined construct are affected by measurement errors and to find methods to minimize these

*CONTACT: Eray Selçuk ✉ erayselcuk84@gmail.com Republic of Türkiye Ministry of National Education, Ankara, Türkiye

e-ISSN: 2148-7456

errors. Another purpose of test theory is to help experts become aware of the logical and mathematical models underlying standard practices in test use and construction (Crocker & Algina, 1986).

Two common measurement theories are used in the historical process of the science of psychometrics, which deals with the test development processes and the problems related to their psychometric properties. These are the Classical Test Theory (CTT), which was first developed, and the Item Response Theory (IRT), also called the Latent Trait Theory (LTT), which is increasingly used.

According to IRT, ability or latent trait is performance on test items. IRT is defined as a model that shows the procedure to be followed to establish the consistency between the latent variables and the findings obtained from these variables. IRT should not be seen as a hypothetical theory because this theory does not explain why a person gives an answer to an item or how he/she decides to answer an item. IRT is more of a model based on statistical estimations. IRT uses latent traits of individuals and items to estimate observed responses (Hambleton et al., 1991). In other words, IRT is a statistical theory about how the item under investigation and test performance relates to the abilities measured by the items in the test (Hambleton & Jones, 1993).

The advantages of IRT models can be achieved only when the fit between the model and test data is satisfactory (De Mars, 2010). The most important conditions for ensuring this harmony are appropriate sample size, adequate test length, and a normal priori distribution type. These conditions significantly affect the amount of error, especially in parameter estimation. In addition, although the number of standard error rates of parameter estimations depends on sample size and test length, estimation methods also affect this amount of standard error. In addition, there are some assumptions that estimation methods can work effectively. In terms of data, if these assumptions are ignored and neglected, the error rates in the estimations increase (Hambleton & Swaminathan, 1985).

There are different methods for estimating item and ability (person) parameters within the framework of IRT. Most of these methods are based on calculating the maximum likelihood (ML) function. The ML function is calculated by estimating the probabilities of the values, maximizing the item and ability parameters over the observed data. These estimation methods perform a solution with an iterative process. The most critical limitation of ML functions, in general, is that it is not possible to estimate the ability parameters of individuals with a full or zero score on a test or to estimate the parameters of the items that are correctly or incorrectly made by everyone (Lord, 1983; Samejima, 1993a, 1993b). In addition, the priori distribution type (normal, skewed left, skewed right, leptokurtic, and platykurtic) effectively estimates item and ability parameters and determines the standard errors of these estimations. In cases where the distribution becomes skewed or when the aforementioned general problems of the ML methods are encountered, “Bayesian Estimation Methods" are recommended to make estimations meticulously (with a lower standard error rate) (Bock & Mislevy, 1982; De Ayala, 2009; Hambleton & Swaminathan, 1985; Hambleton et al., 1991).

ML methods cannot accurately estimate item and ability parameters in generally small samples, short tests, and especially in skewed data. Likewise, the increase in the number of parameters in the IRT model (as in the 2 PL and 3 PL models) also increases the error in these estimations. The literature recommends parameter estimation with the Bayesian approach for such problems.

Most likelihood methods used in IRT are based on the frequency approach. However, the frequency approach has shortcomings because it depends on a fixed value and does not provide distribution information. The Bayesian approach allows estimations by including a priori distribution information. In the Bayesian approach, the variance of the prior distribution represents the uncertainties of the parameter estimates. If the variance of the prior distribution is low, the error rates of the parameter estimates will be lower (van de Schoot & Depaoli, 2014).

Using a Bayesian approach will solve some of the difficulties encountered with the ML approach. Bayesian estimates for the level of ability (ϴ) can be obtained for zero correct response item patterns, fully correct response item patterns, and anomaly response patterns (Hambleton et al., 1991).

Bayesian IRT estimation methods can provide advantages over ML IRT estimation methods (Bock & Mislevy, 1982; De Ayala, 2009; Hambleton & Swaminathan, 1985; Hambleton et al., 1991). The essence of the Bayesian approach is to know the individual's point in the distribution in terms of a trait before obtaining any data. This distribution is called a priori distribution. Therefore, restricting parameter estimations to specific ranges using a priori distribution is essential for Bayesian estimations of IRT (Gao & Chen, 2005).

Gao and Chen (2005) conducted a large-scale simulation study on 3 PL models. In their study, authors used uniform distribution data sets with test lengths of 10, 20, and 60 items and sample sizes of 100, 200, and 500. The authors compared the marginal maximum likelihood (MML) estimation method and Bayesian estimation methods on these data. As a result of the research, the authors concluded that the marginal maximum likelihood method tends to estimate out of the true item parameter values in small samples. Moreover, the authors stated that Bayesian estimation yielded more accurate estimates than marginal maximum likelihood estimation when the sample size was as low as 100. The authors emphasized that the results of Bayesian estimation are more satisfactory regarding the root mean standard error of the estimates (RMSE). However, the error amounts of the marginal maximum likelihood estimation methods also tend to decrease when the test length and sample size increase.

Sass et al. (2005) compared the estimation errors of the latent trait distribution under normal and non-normal distributions. The authors simulatively generated data for 1000 samples, 30 items, and 2 PL models. They used maximum likelihood (ML), Bayesian MAP, and EAP as parameter estimation methods. They also examined true and estimated item parameters to distinguish item parameter estimation from latent trait estimation errors. They stated that non-normal latent trait distributions produce higher estimation errors than normal distributions.

Accordingly, while estimating the parameters based on IRT, the data are the problem of this research is whether there will be a difference between the RMSE of the estimations when the priori distribution type is manipulated in terms of sample size, test length, and logistics model compared to ML and Bayesian IRT. For this purpose, answers to the following research problems were sought through the data generating according to simulation conditions:

1. Is there a significant difference between the RMSE of the ability parameters ($ϴ_{RMSE}$) estimated by ML and Bayesian methods in the generated datasets in 2 PL models according to simulation conditions?
2. Is there a significant difference between the RMSE of the ability parameters ($ϴ_{RMSE}$) estimated by ML and Bayesian methods in the generated datasets in 3 PL models according to simulation conditions?
3. Is there a significant difference between the RMSE of item discrimination (a_{RMSE}), RMSE of item difficulty (b_{RMSE}) and RMSE of lower asymptote (c_{RMSE}) estimated by ML and Bayesian methods in the generated datasets in 2 PL models according to simulation conditions?
4. Is there a significant difference between the RMSE of item discrimination (a_{RMSE}), RMSE of item difficulty (b_{RMSE}) and RMSE of lower asymptote (c_{RMSE}) estimated by ML and Bayesian methods in the generated datasets in 3 PL models according to simulation conditions?

Estimation methods are affected by the distributional types of persons' abilities and item parameters. It is also assumed that most traits (ϴ) are normally distributed in the universe. This assumption reveals the strengths of IRT and affects the estimation of parameters. Therefore, skewed distributions cause some issues in parameter estimation. This is because the accurate estimation of parameters depends on the variance not being sufficiently large at some levels of ϴ. If such distributional assumptions are not satisfied, the accuracy of parameter estimation based on maximum likelihood methods of IRT is questionable. In conclusion, this research is essential in the sense that it acknowledges that parameters estimated with different a priori ability distributions other than the normal distribution (left and right skewed, leptokurtic and platykurtic) have high RMSE and proposes an alternative estimation method to reduce this error and Bayesian approach provides advantages in parameter estimation compared to the ML approach.

1.1. Significance of the Research

The studies by Swaminathan and Gifford (1986), Harwell and Janosky (1991), Gao and Chen (2005), Sass et al. (2005), Finch and Edwards (2015), Çelikten and Çakan (2019) and Kıbrıslıoğlu Uysal (2020) compared different estimation methods on IRT parameter estimation under different conditions. It is seen that most comparisons were made under the conditions of sample size and test length, and the most used estimation methods were likelihood (ML), MAP, and EAP. Studies also investigate the effect of latent trait or item parameter distributions. These studies were generally conducted on simulative data.

This research aims to compare different sample sizes, test lengths, latent trait distributions, and parameter estimation methods with the effect of manipulating conditions as in the previous studies. The research is similar to other studies in this respect. However, the distinguishing feature of this research is that five different types of a priori ability distributions were generated; accordingly, the total test scores also had this distribution type. However, there are some studies in which the latent distribution is skewed. This study analyzed skewness as bidirectional (left-skewed and right-skewed), and leptokurtic and platykurtic distributions were also examined. In addition, in some previous studies, Bayesian estimation has usually been analyzed in Rasch or 2 PL models. This research also examined the results of Bayesian MCMC parameter estimation in the 3 PL model.

As a result of the research, it is foreseen that using Bayesian estimation methods in situations where sample size and test length are not enough for a priori distributions of ability in different patterns will lead to low RMSE in parameter (ability and item) estimations. From this point of view, this research is thought to provide a different viewpoint on the parameter estimation methods used in IRT and contribute to the literature.

2. METHOD

2.1. Research Design

This research created data sets with different the priori distribution types following the simulation conditions. Estimations of ability and item parameters were made using ML and Bayesian (MCMC) methods on these data sets. Simulation studies can use data generated in simulative conditions to investigate certain variables. The simulation approach creates an artificial condition where relevant information and data can be generated. This enables us to observe the dynamic behavior of a system (or sub-system) under controlled conditions (Fraenkel & Wallen, 2009; Kothari, 2004). The literature argues that simulation studies are empirical experiments (Morris et al., 2017) and should be considered statistical sampling, depending on the research design and data analysis principles determined (Hoaglin & Andrews, 1975). Accordingly, this research uses a statistically experimental method to compare estimation methods by manipulating various conditions through simulatively generated data. In this respect, this research is a simulation-based experimental study.

2.2. Generating Data

Monte Carlo (MC) simulation generates the data within the scope of this study following the conditions manipulated in different ways according to the prior distribution types, sample size, test length, logistics model and parameter estimation method specified in the research problem.

Monte Carlo (MC) simulation is used in many applications, such as evaluating new methods in IRT parameter estimation, performance comparison of different item analysis programs, and parameter estimation in multidimensional data. Accordingly, IRT applications using the MC simulation technique should include at least one of the following (Harwell et al., 1996):

1. Evaluation of parameter recovery or parameter estimation methods,
2. Evaluation of the properties of IRT-based statistics,
3. Methodological comparison by combining different IRT applications.

The R programming language generated the data depending on the simulation conditions. In R, mirt (Chalmers, 2012), e1071 (Meyer, 2022), psych (Revelle, 2022) and lattice (Sarkar, 2022) packages were run. The simdata function in the mirt package generated binary (1-0) score matrices with the "Önsel (Prior)" script block written by the researchers, according to the simulation conditions. The "Önsel (Prior)" script block is given in Appendix. While generating the binary score matrices, the priori ability scores produced by the distribution types were placed in the latent distribution argument within the simdata function.

In generating the data in the "Önsel (Prior)" script block, previous research in the literature was referred to for the initial item parameters. Accordingly, log-normal distribution [a~lnN(0.3, 0.2)] was used to generate the item discrimination parameter, standard normal distribution [b~N(0, 1)] was used to generate the item difficulty parameter, and uniform distribution [c~U(0.01, 0.25)] was used to generate the item chance parameter (lower asymptote) (Baker, 2001; Feinberg & Rubright, 2016; Bulut & Sünbül, 2017; Soysal, 2017; Pekmezci, 2018). In generating the a priori ability parameter, more than one and different (normal and uniform) distribution types were combined. In the generation of skewed, leptokurtic, and platykurtic distributions other than the normal distribution, outliers were generated at Z scores above ±4. Accordingly, Θ~N(0, 1) if the distribution is normal; Θ~N(2, 1), Θ~U(-5.0, -4.0) and Θ~U(-4.0, -3.0) if the distribution is left skewed; Θ~N(-2, 1), Θ~U(3.0, 4.0) and Θ~U(4. 0, 5.0); Θ~N(-1, 100), Θ~N(1, 100) and Θ~N(0, 0.00001) if leptokurtic; Θ~N(0, 1), Θ~U(-3.0, -1.0) and Θ~U(1.0, 3.0) if platykurtic.

Considering skewed distributions with normal distribution assumptions leads to incorrect results (Kolen, 1985). Deviations from the normal distribution cause various problems when estimating parameters with ML estimation methods (Hambleton & Swaminathan, 1985). For this reason, the problem of this research is how different a priori ability distribution types will affect parameter estimation methods.

2.3. Simulation Conditions

In the simulation model created to solve the problems in this research, some conditions were fixed while others were manipulated. According to the literature, the selection of each condition in the research was determined by examining similar previous studies. The conditions that were fixed and manipulated are given in Table 1.

Table 1. *Conditions of simulation.*

Conditions of Simulation							
Fixed conditions		Manipulated conditions					
Model Parameters		Parameter estimation methods (x2)		Sample size (x3)	Test length (x3)	Logistics model (x2)	Prior distribution type (x5)
Initial of ability parameters (Θ_i)	Initial of item parameters (a_i, b_i, c_i)	Maximum likelihood (ML)	Bayesian (MCMC)	100 500 1000	10 20 40	2 PL 3 PL	Normal Left-skewed Right-skewed Leptokurtic Platykurtic

Table 1 shows that the research conditions consist of fixed and manipulated conditions. Fixed conditions, initial of model parameters, and manipulated conditions were determined as estimation method, sample size, test length, logistics model, and priori distribution type. Accordingly, parameter estimation methods (ML x Bayesian), sample size (100 x 500 x 1000), test length (10 x 20 x 40), logistics model (2 PL x 3 PL), and priori distribution type (normal x left-skewed x right-skewed x leptokurtic x platykurtic) 180 simulation conditions were carried out with 100 replications. Accordingly, 18000 data sets were used in the research process.

Determining the simulation conditions is essential in reviewing previous research in the literature and determining which factors should be selected to contribute to the literature. In the simulation model developed to solve the research problems in this study, some conditions were kept fixed while others were manipulated.

2.3.1. *Fixed conditions*

Model parameters (ability and item parameters): The initial parameters used to generate ability and item parameters are given in the data generation section.

2.3.2. *Manipulated conditions*

Parameter estimation method: Maximum likelihood (ML) and Bayesian MCMC methods were used to estimate the ability and item parameters. These estimation methods were used for each simulation condition and replications separately. Moreover, this condition is one of the most critical problems the research aims to address.

Sample size: For each simulation condition, three different sample sizes of 100, 500, and 1000 participants were selected. Sample size is considered an essential variable for IRT estimation (Hambleton, 1989; Orlando, 2004). The strengths of IRT depend on the sample size, and it is suggested that it should be applied in large samples (DeMars, 2010). Linacre (1994) stated that small samples are needed when the number of parameters in the model is less, while more complicated models need larger samples. In the literature, there are some studies indicating that sample sizes of 200 (Wright & Stone, 1979) or 500 (Hulin et al., 1982) for 1 PL model, 1000 (Ree & Jensen, 1980) for 2 PL model, and 1000 (Lord, 1968) or 10000 or more (Thissen & Wainer, 1983) for 3 PL model are adequate. In addition, De Ayala (2009) stated that sample sizes of 250 or 500 are adequate for parameter estimation, whereas Hulin et al. (1982) concluded that a sample size of more than 2000 is unnecessary for parameter estimation using ML methods in general. Mislevy (1986) used a sample of 1000 in his study on parameter estimation using Bayesian approach. In this study, we want to utilize the advantages of Bayesian approach by using different sample sizes. Therefore, data sets of 100 for a small sample size, 500 for a medium sample size, and 1000 for a large sample size were used.

Test Length: Three different test lengths were selected for each simulation condition: 10, 20, and 40 items. Using different test lengths leads to a variation in the item response patterns. This variation is especially crucial for the accuracy of item parameter estimates (Hulin et al., 1982). As the test length increases, the accuracy of ϴ estimations increases. Accordingly, increasing the sample size and test length will increase the accuracy of the estimation item parameters (ai, bi, and ci) and thus increase the accuracy of the ability parameter (ϴ) estimates (Reise & Yu, 1990). DeMars (2010) stated that for 2 PL and 3 PL models, the test length should be 20 when using a sample of 500, 40 items when using a sample of 1000, and 50 to 80 items when using a sample of 2000-3000. Hulin et al. (1982) suggest that using a 30-item test in a sample of 500 in 2 PL models and a 60-item test in a sample of 1000 in 3 PL models would be adequate in terms of the accuracy of parameter estimations. Hambleton and Cook (1983) stated that a 20-item test in a sample of 500 in the 3 PL model is adequate for parameter estimation. However, Hambleton and Cook (1983) stated that the estimation error was negatively affected when the test length increased to 40. Akour and Al-Omari (2013) stated that a test length of 15 items in a sample of 200 is sufficient for parameter estimation in the 3 PL model. Mislevy (1986) used 20 and 40 items as test lengths in his study on parameter estimation with the Bayesian approach.

This study generated data sets of 10 items for short tests, 20 for medium length tests, and 40 for longer tests. Although short tests are mostly teacher-made tests in classroom assessments, these tests are now also used in secondary education entrance examinations in Turkey. In these examinations, the number of items in the Turkish History of Turkish Revolution and Kemalism subtests, Religious Culture and Ethics, and Foreign Language, is 10 (MoNE LGS Guide, 2022). For this purpose, 10 items were selected as test length, one of the simulation conditions.

Priori Distribution of Ability (Theta): Each simulation condition used five different types of distributions, keeping the standard deviation values fixed. The simulation conditions were selected as normal and non-normal (left-skewed, right-skewed, leptokurtic, and platykurtic) distribution types. The skewness coefficient's absolute value means that the samples' distribution types are highly skewed when greater than 1.00, moderately skewed between 0.50 and 1.00, and approximately symmetric when less than 0.50. For kurtosis, it is stated that the distribution is normal if the coefficient is 3, leptokurtic if it is greater than 3, and platykurtic if it is less than 3 (Bulmer, 1979). However, with the addition of -3 to the formula, this value becomes 0. This means that a kurtosis coefficient of 0 indicates that the distribution is normal, a coefficient greater than 0 indicates that the distribution is leptokurtic, and a coefficient less than 0 indicates that the distribution is platykurtic. Tabachnick and Fidell (2014) stated that when the skewness and kurtosis values are between -1.50 and +1.50, the distribution is assumed to be normal. Evaluating skewed distributions with normal distribution assumptions causes incorrect conclusions (Kolen, 1985). It is known that deviations from the normal distribution cause various problems when estimating parameters with maximum likelihood estimation methods (Hambleton & Swaminathan, 1985). For this reason, the issue of this study is how different a priori ability distribution types will affect parameter estimation methods.

IRT Model: This research selected 2 PL and 3 PL models for parameter estimations. According to Hulin et al. (1982), these logistic models are robust and the most widely used models.

Accordingly, two different references were considered when setting the simulation conditions. The first one is to benefit from similar studies in the literature while setting each condition, and the second one is to consider the advantages of the Bayesian estimation method depending on the purpose of the research. In the first reference, the previous research related to the literature is discussed in detail under the topic of each condition. In the second reference, these conditions were selected by considering the problems of ML estimation and the advantages of Bayesian estimation. Since this study aims to determine how the ML and Bayesian estimation results will change, especially in cases where the sample becomes smaller, the number of items decreases. The prior ability distribution becomes skewed; this is another significant reason for choosing the simulation conditions in this way.

Harwell et al. (1996) suggested that at least 25 replications should be used in studies where the IRT parameters are manipulated. However, Seong (1990) used 5 replications, Stone (1992) used 100 replications, Kirisci et al. (2001) used 10 replications, Sass et al. (2008) used 100 replications, Finch and Edwards (2015) used 1000 replications, Bulut and Sünbül (2017) used 100 replications, Karadavut (2019) used 25 replications, and Kıbrıslıoğlu Uysal (2020) used 100 replications in various simulation studies given in related studies.

A literature review shows that similar simulation studies use different numbers of replications when generating data. There are two factors affecting this issue. The first is that the degree of accuracy of the data generated because of a low number of replications is insufficient, and the second is that the simulation program is inadequate and time costly because of many replications (Bulut & Sünbül, 2017). Moreover, Feinberg and Rubright (2016) proposed a formulation for the number of replications in IRT simulations based on the standard deviation of the estimated parameters. This equation is given below:

$$\sigma_M = \frac{\hat{\sigma}}{\sqrt{R-1}} \qquad \text{(Equation 1)}$$

where $\hat{\sigma}$ is the standard deviation of the estimated parameter across replications, R is the number of replications, and σM is the standard error of the mean. Accordingly, researchers determine an initial number of replications, and after computing the standard deviation of the data, they set a new number of replications. If the standard deviation is larger than expected Feinberg and Rubright (2016) recommend increasing the number of replications. However, there is no acceptable value for the estimated standard deviation value. Therefore, Barış-Pekmezci and Şengül-Avşar (2021) state that it is not practical to use this equation. Therefore, considering the research previously cited in the literature, it was decided to use 100 replications in this study to produce accurate results and not to increase the simulation time.

2.4. Analysis of Data

First, basic assumptions were checked to determine the fit of the generated datasets for IRT parameter estimation. These assumptions are unidimensionality, local independence, and model-data fit (Baker, 2001; Baker & Kim, 2004; De Ayala, 2009; Hambleton & Swaminathan, 1985; Hambleton et al., 1991; Hambleton & Jones, 1993). The psych (Revelle, 2022), sirt (Robitzsch, 2022), and mirt (Chalmers, 2012) packages were used to test the basic assumptions.

Second, the R programming language was used in the analysis of the data as well as in the generating of the data. The R software version used is R Studio, Version: 2022.12.0+353. Researchers generally use statistics such as correlation, covariance, bias, absolute bias, standard error of estimate (SE), mean square error (MSE) and root mean square error (RMSE). The statistics to be used and how to interpret them depend on the problems of the research. A review of the literature shows that bias, standard error (SE) and root mean square error of the mean square error (RMSE) are the most used ones (Feinberg & Rubright, 2016). RMSE was used in this research.

Root means square error (RMSE) between the ability and item parameters and the initial parameters estimated on the data generated according to the simulation conditions were calculated. This is because biased values can take both positive and negative values. This situation affects the mean of bias. In addition, there is a relationship between RMSE and bias. This relationship is given in the equation below (Atar, 2007; Bilir, 2009; Feinberg & Rubright, 2016):

$$RMSE^2 = Bias^2 + SE^2 \qquad \text{(Equation 2)}$$

In this equation, the sum of the bias's square and the standard error's square equals the square of the RMSE. Accordingly, the negative and positive biases created by the bias have disappeared. While analyzing the data, the ML estimation was first performed using the irtplay package (Lim & Wells, 2020) compared to ML approaches, followed by standard Bayesian estimations using Monte Carlo Markov Chain (MCMC) methods using the bairt (Martinez, 2017) and sirt (Robitzsch, 2022) packages for Bayesian approaches. In Bayesian estimations, the burning was defined as 1000, and the iteration was defined as 3000. The number of burn-in and iterations are set at these values due to the procedures performed in the algorithm of the method. Because in the MCMC method, the first chain generated up to the burn-in value is subtracted from the whole chain generated later. Thus, parameter estimation is performed from the sample generated by the number of iterations (Martin & Quinn, 2006; SAS Institute, 2020). These values are determined according to the conditions of the simulation to provide unbiased results at the expected level.

Third, the significance of the differences between the RMSE values of the parameters was tested by mixed model ANOVA according to sample size, test length, logistic model, a priori distribution type, and estimation method. Assumptions were checked before analyzing the mixed model ANOVA. Afex (Singmann, 2022) and emmeans (Lenth, 2022) packages were used for this analysis. For the mixed model ANOVA, the main effects (between) variables were the simulation conditions that were manipulated (sample size, test length, a priori ability distribution types, parameter estimation methods) and fixed (initial values of ability and item parameters), and the number of simulation replications was assigned as the interaction (within) variable. According to the analysis results, the significant conditions' effect sizes (generalized eta-square coefficient) were computed and assessed according to Cohen's (1988) proposal. Accordingly, the size of the effect size was interpreted as weak if it was less than 0.0099, moderate if it was 0.0588, and strong if it was greater than 0.1379. At the same time, since the generalized eta-square coefficient takes a value between 0 and 1 when this value is multiplied by 100, it shows how much of the variance of the dependent variable is explained by the independent variables (Lakens, 2013). Statistically significant conditions were compared using the Bonferroni post hoc comparison method, included by default in the emmans package. According to the analysis results, ggplot2 (Wickham, 2016) and ggbeeswarm (Clarke, 2022) packages were used to visualize significant conditions.

3. RESULTS

Analysis was conducted to determine whether the datasets meet the assumptions of the IRT. Accordingly, for the unidimensionality assumption, the ratio of the explained variance, the averages of the first eigenvalues and the ratio of the first eigenvalue to the second eigenvalue were calculated according to the explanatory factor analysis results. It was accepted that this assumption was fulfilled if a dominant factor was found (Lord, 1980). Accordingly, it is seen that the data fulfills the unidimensionality assumption in all conditions.

The Q3 statistic of Yen (1984) is used to test the local independence assumption. Accordingly, it is determined that the local independence assumption is mostly fulfilled for the data in all conditions.

M2 values were examined to test the assumption of model-data fit. As a fit criterion, the M2 statistic is expected to be non-significant (Maydeu-Olivares & Joe, 2006). Accordingly, it is seen that model-data fit is fulfilled in all the data.

Normality and homogeneity of variances test results of the data were analyzed. In big samples, it is more practical to use descriptive statistics and graphical analysis to check the normality assumption. In big samples, normality tests with hypothesis tests risk increasing the probability of Type I error (Demir, 2019). Accordingly, it is seen that the skewness and kurtosis coefficients and histogram graphs of the data fulfill the normality assumption. Examining the hypothesis of homogeneity of variances test results shows that this assumption is fulfilled ($F_{(2PL.\theta.RMSE)} = 0.13$; p>.05, $F_{(2PL.a.RMSE)} = 0.51$; p>.05, $F_{(2PL.b.RMSE)} = 0.78$; p>.05, $F_{(3PL.\theta.RMSE)} = 0.06$; p>.05, $F_{(3PL.a.RMSE)} = 0.21$; p> .05, $F_{(3PL.b.RMSE)} = 0.99$; p>.05, $F_{(3PL.c.RMSE)} = 0.59$; p>.05). Then, the findings related to the research problems are presented under headings.

3.1. Investigation of θ_{RMSE} Estimated by ML and Bayesian Methods in 2 PL Model

In the first problem of the study, the RMSE changes of ability parameters according to sample size, test length, and estimation method were analyzed with mixed model ANOVA in the data in the 2 PL model with normal and non-normal priori distribution (left-skewed, right-skewed, leptokurtic and platykurtic). Accordingly, the results of the mixed model ANOVA performed for the ability parameters according to the sample size, test length, and estimation method in the data in the 2 PL model with normal and non-normal priori distribution (left-skewed, right-skewed, leptokurtic, and platykurtic) are given in Table 2.

Table 2. *Mixed model ANOVA results for ability parameters RMSE in data in 2 PL models with normal and non-normal priori distribution.*

Independent variables	Mean squares of error	Degrees of freedom	F	*p*	Generalized η^2
Estimation method (K)	72.05	1	7.83	0.006**	0.078
Sample size (S)	79.36	2	0.00	0.997	0.001
Test length (M)	65.24	2	9.42	0.001**	0.171
Prior distribution type (D)	42.74	4	1.88	0.001**	0.456
K*S	75.46	2	0.01	0.994	0.001
K*M	58.50	2	1.69	0.191	0.037
K*D	31.90	4	4.05	0.005**	0.155
Error	0.30	198			
Total	425.55				

$^{*}p< .05$, $^{**}p< .01$

Table 2 shows that the main effects of the estimation method ($F_{(1, 88)} = 7.83$; $p<.01$, $\eta^2 = .078$), test length ($F_{(2, 84)} = 9.42$; $p<.01$, $\eta^2 = .171$) and priori distribution type ($F_{(4, 80)} = 1.88$; $p<.01$, $\eta^2 = .456$) seem to have a significant effect. However, the sample size ($F_{(2, 87)} = 0.00$; $p>.05$, $\eta^2 = .001$) did not have a significant effect. Significantly, the estimation method has a medium effect size, the test length is high, and the priori distribution type has a high effect size. When the interactions were examined, the interaction between the estimation method and the priori distribution type was significant ($F_{(4, 80)} = 4.05$; $p<.01$, $\eta^2 = .155$). The effect size of the interaction is high. Pairwise comparisons of the ability parameter estimation method in 2 PL models are given in Table 3.

Table 3. *Ability parameter estimation method pair comparisons in 2 PL models.*

Estimation method	Difference	Standard error	*t*	*p*
Bayes-ML	-0.501	0.179	-2.799	0.001**

$^{*}p< .05$, $^{**}p< .01$

Table 3 shows that Bayesian estimation, the ability parameter estimation method in the 2 PL model, produced lower and more significant RMSE than the ML (t=-2.799; p<.01). The RMSE changes of the ability parameter estimation methods in the 2 PL model are given in Figure 1.

Figure 1. *The change of ability parameter RMSE in 2 PL models by estimation methods.*

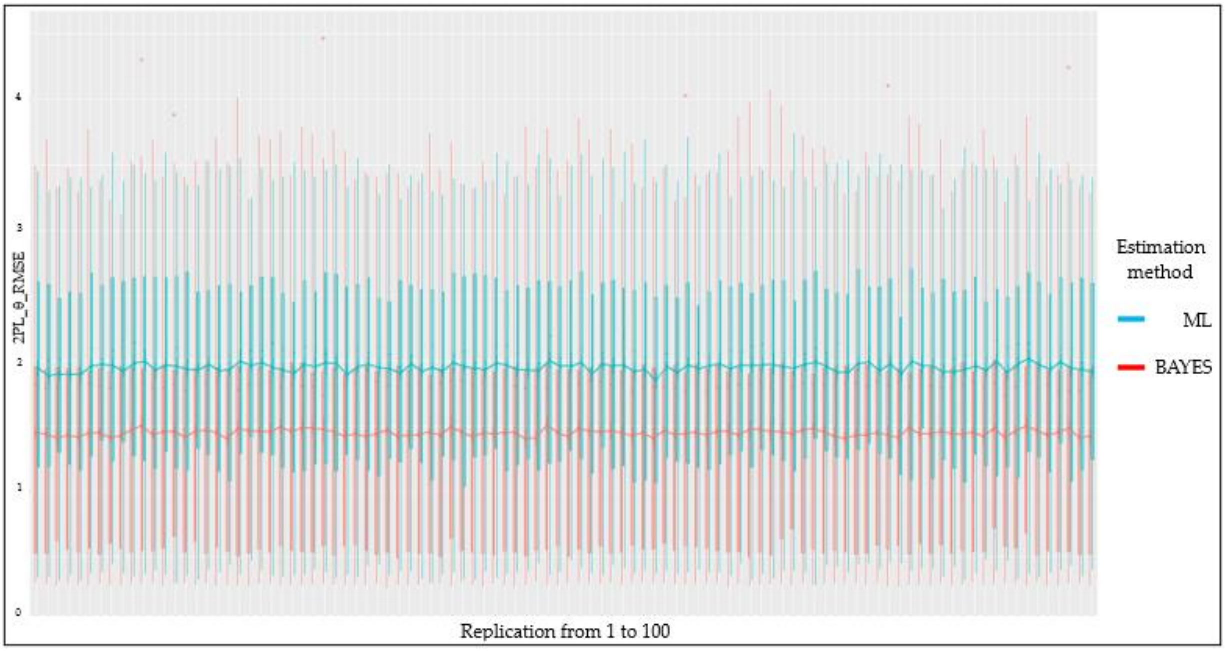


Figure 1 shows that the RMSE of the ability parameters obtained from all data sets in the 2 PL model, regardless of the research conditions, change. Accordingly, while the ability parameter was estimated in the 2 PL model, the Bayesian method produced lower and more significant RMSE than the ML method. Pairwise comparisons according to the number of items on the ability parameter in the 2 PL model are given in Table 4.

Table 4. *Pairwise comparisons of ability parameter RMSE in 2 PL models by test length.*

Test Length	Difference	Standard error	*t*	*p*
10 – 20	0.272	0.209	1.304	0.397
10 – 40	0.884		4.236	0.001**
20 – 40	0.612		2.933	0.012*

*p< .05, **p< .01

Table 4 shows that there are significant differences between test lengths 10 and 40 (t=4.236; p<.01) and 20 and 40 (t=2.933; p<.05) on ability parameter RMSE in the 2 PL model. The RMSE change according to test length on the ability parameter in the 2 PL model is given in Figure 2.

Figure 2. *The change of ability parameter RMSE in 2 PL models by the test length.*

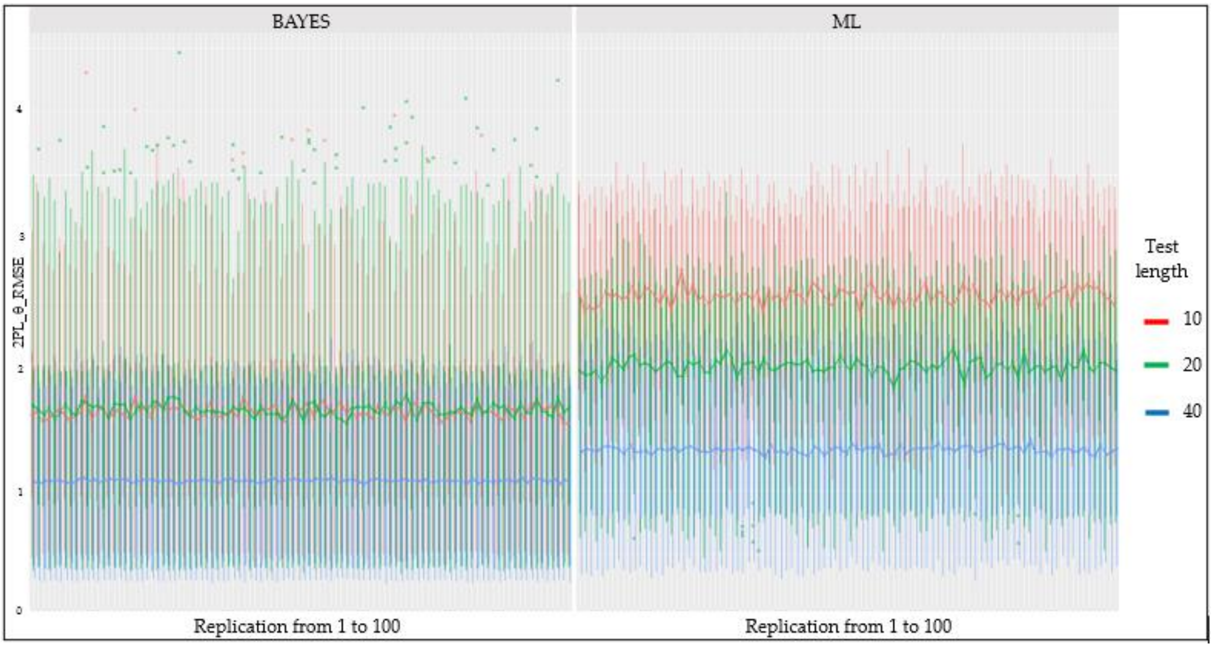


Figure 2 shows that RMSE decreases as the test length increases on the ability parameter estimations in the 2 PL model. As a result of the estimation made with the ML, the RMSE of the ability parameters decreases as the test length increases. The same situation is seen in the Bayesian estimation method. In the Bayesian estimation method, there is no difference in the test length between 10 and 20, but a lower RMSE is obtained in case the test length is 40. However, the RMSE of ability parameters obtained according to test length in Bayesian estimation was lower than in ML estimation. Pairwise comparisons according to priori distribution on the ability parameter in the 2 PL model are given in Table 5.

Table 5 shows that the priori distribution type on the ability parameter RMSE in the 2 PL model is normal to left skewed (t=-7.292; p<.01), normal to right skewed (t=-7.321; p<.01), normal to leptokurtic (t=-5.434; p<.01), normal to platykurtic (t=-3.267; p<.05), left skewed to platykurtic (t=4.026; p<.01), right skewed to platykurtic (t=4.054; p<.01) significant differences were found. These differences are in favor of the Bayesian estimation method. In the 2 PL model, Bayesian estimation produces lower RMSE as the priori distribution type differs from the normal. The RMSE change according to the priori distribution type on the ability parameter in the 2 PL model is given in Figure 3.

Table 5. *Pairwise comparisons of ability parameter RMSE in 2 PL models by prior distribution.*

Prior talent distribution type	Difference	Standard error	*t*	*p*
Normal – Left skewed	-1.589	0.218	-7.292	0.000**
Normal – Right skewed	-1.595		-7.321	0.000**
Normal – Leptokurtic	-1.184		-5.434	0.000**
Normal – Platykurtic	-0.712		-3.267	0.013*
Left skewed – Right skewed	-0.006		-0.029	0.999
Left skewed – Leptokurtic	0.405		1.858	0.347
Left skewed – Platykurtic	0.877		4.026	0.001**
Right skewed – Leptokurtic	0.411		1.887	0.332
Right skewed – Platykurtic	0.884		4.054	0.001**
Leptokurtic – Platykurtic	0.472		2.267	0.202

*$p < .05$, **$p < .01$

Figure 3. *The change of ability parameter RMSE in 2 PL models by prior distribution type.*

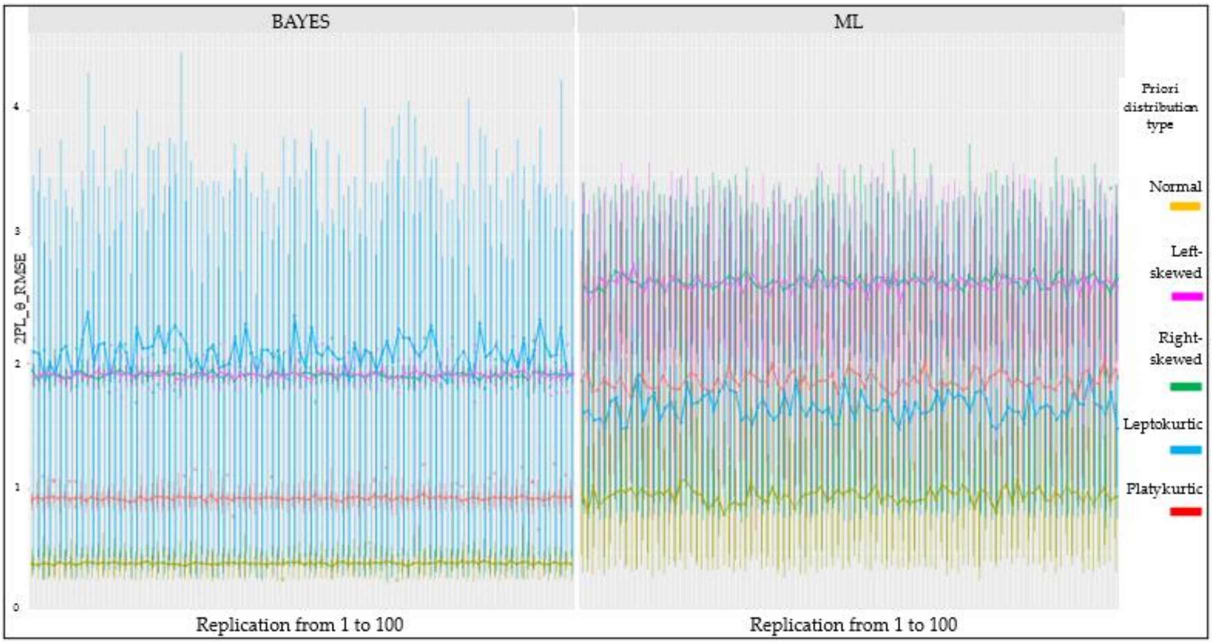


Figure 3 shows that the priori distribution in the 2 PL model becomes skewed from normal (left skewed, right skewed, leptokurtic, and platykurtic), and the RMSE of the ability parameters increases in the ML estimation. The lowest RMSE on the ability parameters was obtained in ML estimation when the prior distribution was normal. As the distribution becomes skewed, the error values increase. The RMSE is highest when the distribution is left skewed and right skewed and lower when it is leptokurtic and platykurtic. As the distribution normalizes, these values show a further decrease. In the 2 PL model, when the Bayesian method performs the ability parameters estimation, RMSE is lower than the ML estimation.

Similarly, the lowest RMSE is in the normal, platykurtic, left and right skewed distribution and the leptokurtic distribution, respectively. In all the priori distribution types, except for the leptokurtic distribution, the RMSE decreases in Bayesian estimation. In contrast, in the leptokurtic distribution, they have higher values than the ML estimation. When the prior distribution is produced, since the leptokurtic distribution has a lower standard deviation than the normal distribution and remains relatively between -1 and +1 as a distribution range, it takes shape in a broader range as a posterior distribution compared to the prior distribution. Therefore, the RMSE differences between the initial and estimated ability parameters increase.

Accordingly, while estimating the ability parameters in the 2 PL model, using the Bayesian estimation method in other distribution types provides lower RMSE, except when the priori distribution is leptokurtic.

3.2. Investigation of θ_{RMSE} Estimated by ML and Bayesian Methods in 3 PL Model

In the second problem of the study, the RMSE changes of ability parameters according to sample size, test length, and estimation method were analyzed with mixed model ANOVA in the data in the 3 PL model with normal and non-normal priori distribution (left-skewed, right-skewed, leptokurtic and platykurtic). Accordingly, the mixed model ANOVA results were performed for the ability parameters according to the sample size, test length, and estimation method in the data in the 3 PL model with normal and non-normal priori distribution (left-skewed, right-skewed, leptokurtic, and platykurtic) are given in Table 6.

Table 6. *Mixed model ANOVA results for ability parameters RMSE in the data in the 3 PL model with normal and non-normal priori distribution.*

Independent variables	Mean squares of error	Degrees of freedom	F	p	Generalized η^2
Estimation method (K)	2769.62	1	0.27	0.607	0.003
Sample size (S)	2747.44	2	0.99	0.376	0.022
Test length (M)	2488.40	2	5.62	0.005**	0.111
Priori distribution type (D)	2315.05	4	5.15	0.001**	0.189
K*S	2836.73	2	0.00	0.999	0.001
K*M	2568.23	2	0.00	0.996	0.001
K*D	2441.93	4	0.07	0.991	0.003
Error	0.96	198			
Total	18168.36				

$^{*}p< .05$, $^{**}p< .01$

Table 6 shows that the test length is the main effect of the independent variables ($F_{(2, 84)} = 5.62$; $p<.01$, $\eta^2 = .111$) according to the mixed model ANOVA results for the ability parameters RMSE in the data in the 3 PL model with normal and non-normal priori distribution and priori distribution type ($F_{(4, 80)} = 5.15$; $p<.01$, $\eta^2 = .189$) were found to be significant. The estimation method ($F_{(1, 88)} = 0.27$; $p>.05$, $\eta^2 = .003$) and sample size ($F_{(2, 87)} = 0.99$; $p>.05$, $\eta^2 = .022$) do not have a significant difference. Significantly, the test length is medium, and the priori distribution type has a high effect size. When the interactions were examined, no condition was found to be significant. Pairwise comparisons according to the test length on the ability parameter in the 3 PL model are given in Table 7.

Table 7. *Pairwise comparisons of ability parameter RMSE in 3 PL models by test length.*

Test length	Difference	Standard error	t	p
10 – 20	3.429	1.288	2.663	0.025*
10 – 40	3.988		3.096	0.007**
20 – 40	0.558		0.434	0.902

$^{*}p< .05$, $^{**}p< .01$

Table 7 shows that there are significant differences between test lengths 10 and 20 (t=2.663; $p<.05$) and 10 and 40 (t=3.096; $p<.01$) on ability parameter RMSE in the 3 PL model. The RMSE change according to test length on the ability parameter in the 3 PL model is given in Figure 4.

Figure 4. *The change of ability parameter RMSE in 3 PL model by estimation methods*.

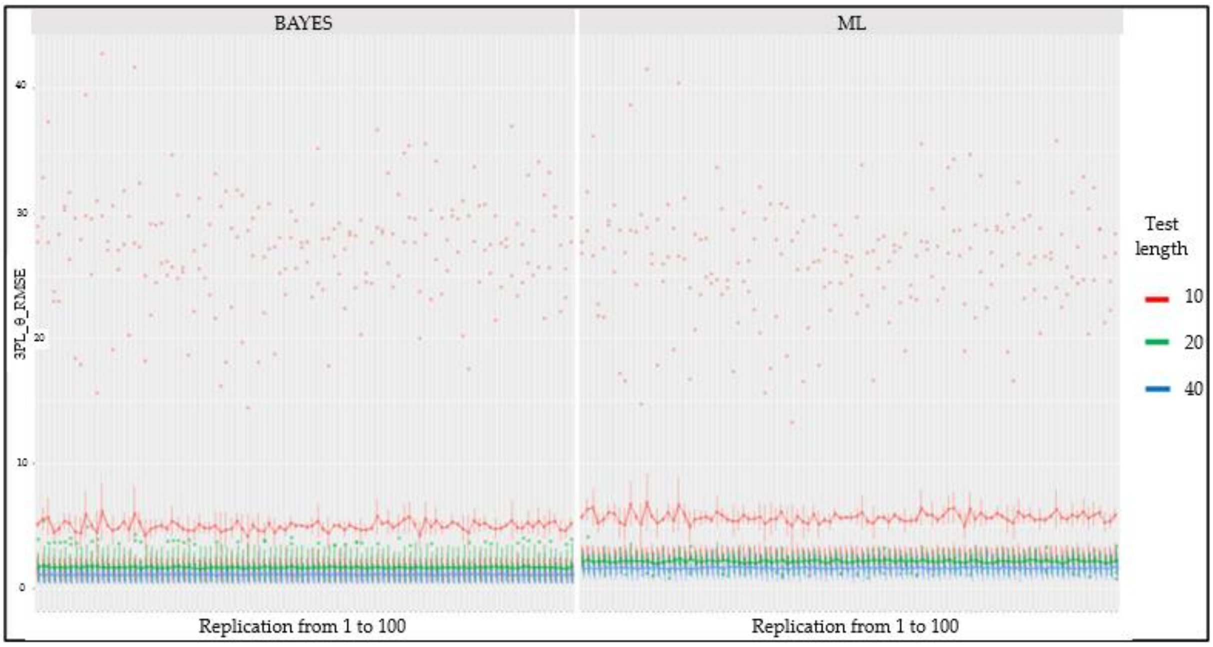


Figure 4 shows that the RMSE decreases as the test length increases on the ability parameter estimations in the 3 PL model. At the same time, the Bayesian estimation method took lower values than ML estimation in cases where test length decreased. However, this situation was not found to be significant. Therefore, using ML or Bayesian methods does not make a difference when estimating ability parameters in the 3 PL model. However, regardless of the estimation method used, the increase in test length causes a decrease in the RMSE of the ability parameters. For example, when the test length decreased to 10, RMSE in the ability parameters increased significantly. Accordingly, lower RMSE for ability parameters in the 3 PL model was observed when the test length was 20 and 40. Pairwise comparisons according to priori distribution type on the ability parameter in the 3 PL model are given in Table 8.

Table 8. *Pairwise comparison of ability parameter RMSE in 3 PL models by prior distribution type*.

Prior distribution type	Difference	Standard error	*t*	*p*
Normal – Left skewed	-1.376	1.604	-0.858	0.911
Normal – Right skewed	-1.300		-0.811	0.926
Normal – Leptokurtic	-6.463		-4.030	0.001**
Normal – Platykurtic	-0.697		-0.434	0.992
Left skewed – Right skewed	0.076		0.047	0.999
Left skewed – Leptokurtic	-5.088		-3.172	0.017*
Left skewed – Platykurtic	0.679		0.424	0.993
Right skewed – Leptokurtic	-5.163		-3.219	0.015*
Right skewed – Platykurtic	0.604		-0.376	0.996
Leptokurtic – Platykurtic	5.767		3.595	0.004**

$^{*}p< .05$, $^{**}p< .01$

Table 8 shows that the priori distribution type on the ability parameter RMSE in the 3 PL model is normal to leptokurtic (*t*=-4.030; *p*<.01), left skewed to leptokurtic (*t*=-3.172; *p*<.05), right skewed to leptokurtic (*t*=-3.219; *p*<.05), significant differences were found between leptokurtic and platykurtic (*t*=3.595; *p*<.01). In the 3 PL model, RMSE increase as the priori distribution becomes leptokurtic on the ability parameters. The RMSE change according to the priori distribution type on the ability parameter in the 3 PL model is given in Figure 5.

Figure 5. *The change of ability parameter RMSE in the 3 PL model by prior distribution type.*

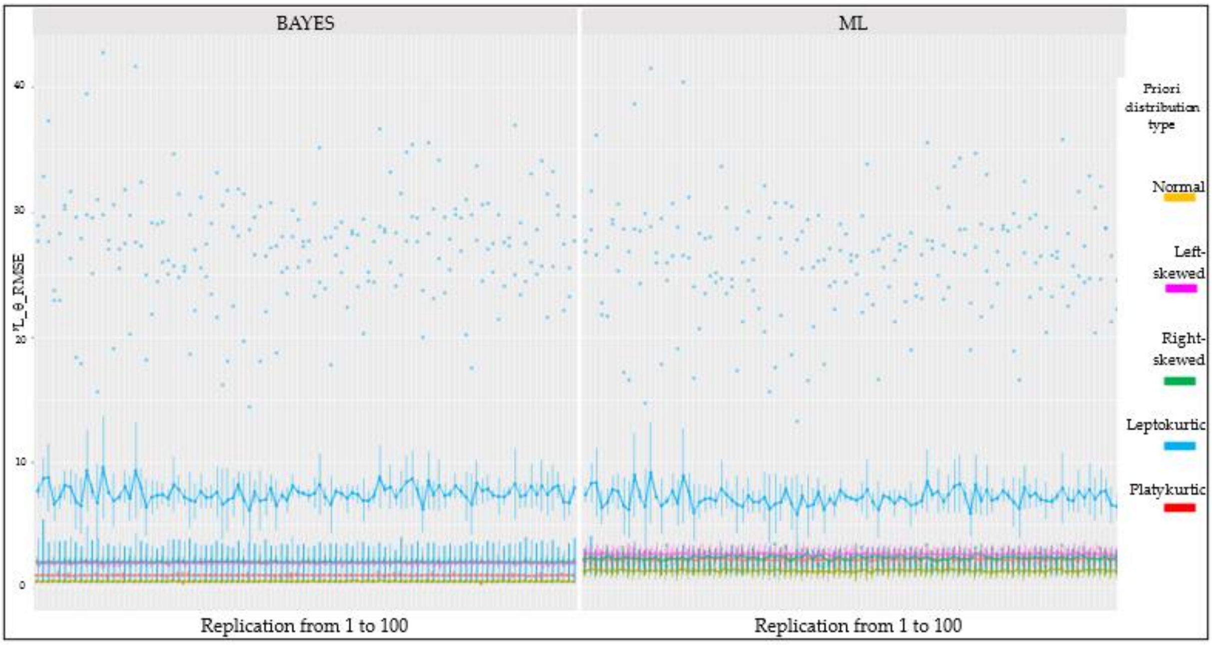


Figure 5 shows that the priori distribution type becomes leptokurtic in the 3 PL model, and the RMSE of ability parameters takes higher values. However, according to the ability parameters estimation method, other distribution types did not differentiate on the priori distribution type, except for the leptokurtic distribution. Therefore, as in the 2 PL model, the leptokurtic priori distribution on the estimations of the ability parameters significantly affects the RMSE. This is seen in both ML and Bayesian estimation methods. Accordingly, the leptokurtic of the priori distribution harms the RMSE of the ability parameters, regardless of the model (2 PL or 3 PL). This situation is likely caused by the leptokurtic distribution (lower standard deviation and narrow ranges) and the data structure generated while performing the simulation. For this reason, cases where priori is leptokurtic should be examined in more detail within the framework of IRT parameter estimations.

3.3. Investigation of a_{RMSE}, b_{RMSE}, c_{RMSE} Estimated by ML and Bayesian Methods in 2 PL Model

RMSE changes of item parameters according to sample size, test length, and estimation method in 2 PL models with normal and non-normal (left skewed, right skewed, leptokurtic, and platykurtic) priori distribution stated in the third problem of the study were analyzed by mixed model ANOVA. Accordingly, the mixed model ANOVA results were performed for the item parameters according to sample size, test length, and estimation method in the data in 2 PL models with normal and non-normal priori distribution (left-skewed, right-skewed, leptokurtic, and platykurtic) are given in Table 9.

Table 9 shows that according to the mixed model ANOVA results for the item discrimination parameter RMSE in the data in the 2 PL models with normal and non-normal priori distribution, the main effects of independent variables as estimation method ($F_{(1, 88)} = 8.17$; $p<.01$, $\eta^2 = .045$), sample size ($F_{(2, 87)} = 8.97$; $p<.01$, $\eta^2 = .090$) and priori distribution type ($F_{(4, 85)} = 3.93$; $p<.01$, $\eta^2 = .083$) have significant effects. Test length ($F_{(2, 87)} = 0.10$; $p>.05$, $\eta^2 = .001$) did not have a significant effect. Among the independent variables found to be statistically significant, the estimation method has a small effect size, the sample size has a medium effect size, and the priori ability distribution has a medium effect size.

Table 9. *Mixed model ANOVA results for item parameters RMSE in data in 2 PL models with normal and non-normal priori distribution.*

Independent variables	Mean squares of error	Degrees of freedom	F	*p*	Generalized η^2
Item discrimination (a_{RMSE})					
Estimation method (K)	5385.60	1	8.17	0.005**	0.045
Sample size (S)	4935.31	2	8.97	0.001**	0.090
Test length (M)	5939.41	2	0.10	0.905	0.001
Prior distribution type (D)	5141.71	4	3.93	0.006**	0.083
K*S	3891.39	2	7.52	0.001**	0.070
K*M	5621.35	2	0.05	0.952	0.001
K*D	4210.31	4	3.34	0.014*	0.069
Error	53.68	198			
Total	35178				
Item difficulty (b_{RMSE})					
Estimation method (K)	5827.21	1	1.26	0.264	0.002
Sample size (S)	5706.21	2	2.08	0.131	0.007
Test length (M)	5900.27	2	0.58	0.562	0.002
Prior distribution type (D)	5606.80	4	1.94	0.111	0.014
K*S	5597.01	2	1.69	0.191	0.006
K*M	5931.07	2	0.65	0.523	0.003
K*D	5244.57	4	2.37	0.060	0.017
Error	311.54	198			
Total	40124.68				

*p< .05, **p< .01

According to the mixed model ANOVA results, none of the independent variables created a significant difference for the item difficulty parameter RMSE values in the data in the 2 PL model with and without normal a priori ability distribution. Therefore, only significant conditions on the item discrimination parameter RMSE were given in the third research problem.

In the 2 PL model, sample size with estimation method ($F_{(2, 84)} = 7.52$; $p<.01$, $\eta^2 = .070$) and priori distribution type with estimation method ($F_{(4, 80)} = 3.34$; $p<.05$, $\eta^2 = .069$) were significant differences on item discrimination parameter RMSE. However, these pairwise interactions had moderate effect sizes. Therefore, for the data in the 2 PL model, the pairwise comparisons of the estimation method having a significant effect on the item discrimination parameter are given in Table 10.

Table 10. *Pairwise comparisons of item discrimination parameter RMSE in 2 PL model by method of estimation.*

Estimation method	Difference	Standard error	*t*	*p*
ML-Bayes	4.421	1.547	2.858	0.005**

*p< .05, **p< .01

Table 10 shows that the estimation method on the item discrimination parameter RMSE in the 2 PL model data with normal and non-normal priori distribution type is in favor of the Bayesian estimation method and significant (t=2.858; p<.01). RMSE changes of the item discrimination parameter estimation methods in the 2 PL model are given in Figure 6.

Figure 6. *The change of item discrimination parameter RMSE in 2 PL model by method of estimation*

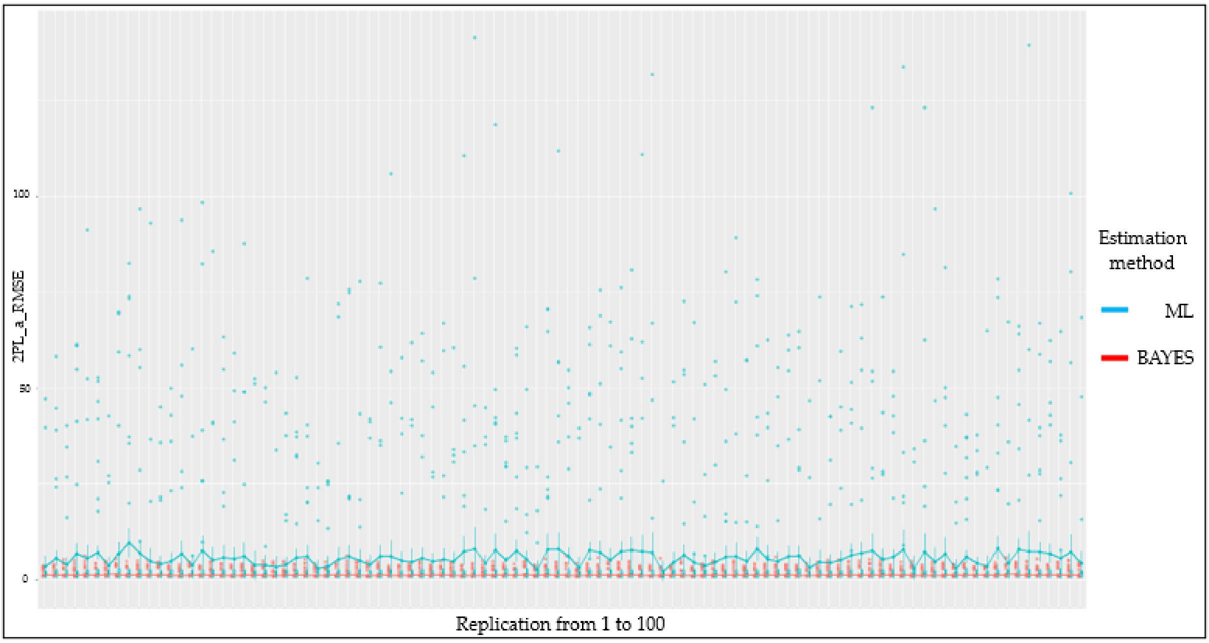


Figure 6 shows that the item discrimination parameter RMSE in the 2 PL model, independent of all research conditions, takes lower Bayesian estimation values than ML estimation. Furthermore, while the item discrimination parameter RMSE (a_{RMSE}) shows a scattering according to the estimation results of the ML method, these values are more linear and stable in Bayesian estimation. Accordingly, the Bayesian approach provides advantages over the ML procedure in estimating item discrimination parameters. Pairwise comparisons of the sample size significantly affected the item discrimination parameters for the data in the 2 PL models, which are given in Table 11.

Table 11. *Pairwise comparisons of item discrimination parameter RMSE in 2 PL model by sample size and estimation method.*

Estimation method	Sample size		Difference	Standard error	*t*	*p*
ML	100	500	11.958	2.278	5.250	0.000**
		1000	12.164		5.340	0.000**
	500	1000	0.207		0.091	0.999
Bayes	100	500	1.198		0.526	0.995
		1000	1.290		0.566	0.993
	500	1000	0.092		0.040	0.999
ML*Bayes	100	100	-11.633		-5.107	0.000**
	500	500	-0.873		-0.383	0.999
	1000	1000	-0.758		-0.333	0.999

$^{*}p< .05$, $^{**}p< .01$

Table 11 shows a significant difference between the RMSE of the item discrimination parameter estimated by the ML method in the 2 PL model between sample sizes of 100 and 500 (t=5.250; p<.01) and between 100 and 1000 (t=5.340; p<.01). However, there was no difference between sample sizes in Bayesian estimation. Accordingly, the significant RMSE in small samples in ML estimation decreased in the Bayesian method. Nevertheless, the RMSE of the item discrimination parameter estimated by different methods at the same sample sizes showed a significant difference at a sample size of 100 (t=-5.107; p<.01). This difference was eliminated as the sample size increased. Accordingly, using the Bayesian estimation method to obtain item discrimination parameters with low RMSE in small samples is more suitable. RMSE change according to sample size on item discrimination parameter in the 2 PL model is given in Figure 7.

Figure 7. *The change of item discrimination parameter RMSE in 2 PL model by sample size.*

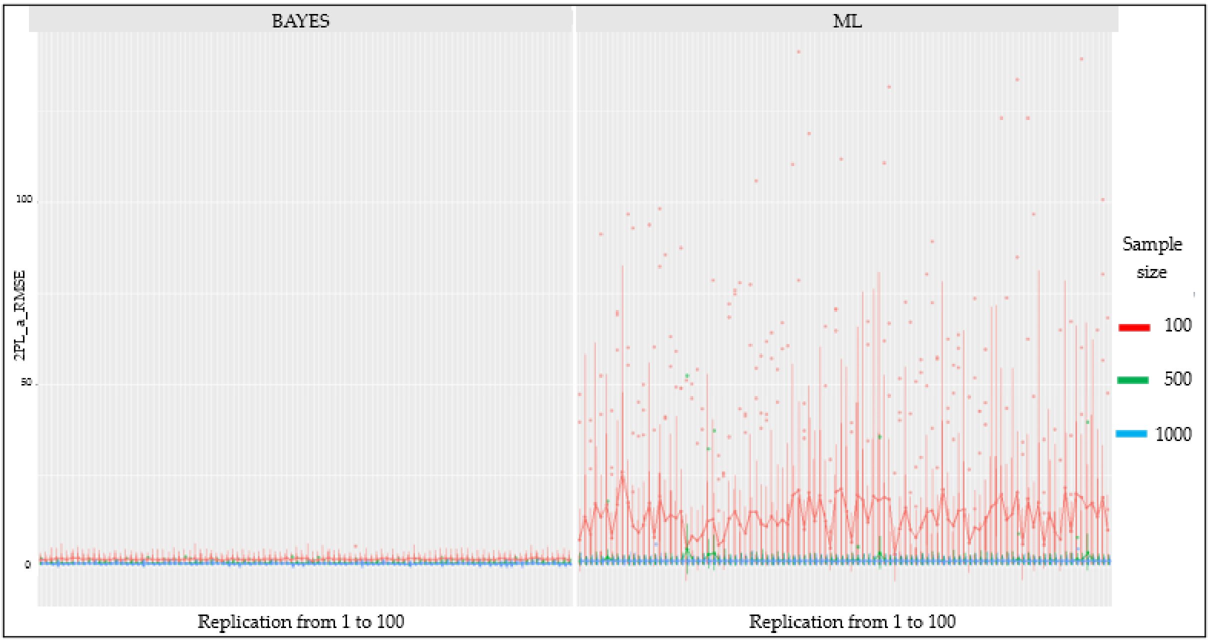


Figure 7 shows that the Bayesian estimation method produced lower values on item discrimination parameter RMSE (a_{RMSE}) when the sample size decreased compared to ML estimation. When the sample size decreased to 100 in the ML estimation, the item discrimination parameter RMSE increased excessively and created scattering. In this case, when the Bayesian estimation method was used, RMSE tended to decrease and showed a linear distribution. When the sample size was 500 or 1000, RMSE did not show a significant difference according to the estimation method. As can be understood from this, when the ML estimation method is used in the 2 PL model, a sample of at least 500 sample size should be used to reduce the item discrimination parameter RMSE. When the sample size drops to 100, the Bayesian estimation method should be used. Pairwise comparisons on the item discrimination parameter in the 2 PL model according to the priori distribution form are given in Table 12.

Table 12 shows that significant differences were found between the item discrimination parameter RMSE estimated by ML method in the 2 PL model between normal and left skewed (t=-4.031; p<.01), normal and right skewed (t=-3.754; p<.05), left skewed and leptokurtic (t=3.815; p<.01), left skewed and platykurtic (t=3.513; p<.05), right skewed and leptokurtic (t=3.538; p<.05) and right skewed and platykurtic (t=3.236; p<.05) according to the distribution types. These differences were eliminated in Bayesian estimation. The RMSE of the item discrimination parameter estimated by Bayesian method in the 2 PL model were not significantly affected by the type of prior distribution. In the same type of a priori distributions, item discrimination parameter RMSE estimated by ML and Bayesian methods differed significantly when the distribution was left skewed (t=3.569; p<.05) or right skewed (t=3.300; p<.05). However, according to the estimation methods, no difference was found for the other distribution types. In the 2 PL model, RMSE on the item discrimination parameter according to the priori ability distribution types are given in Figure 8.

Table 12. *Pairwise comparisons of item discrimination parameter RMSE in 2 PL model by priori distribution type and estimation method.*

Estimation method	Prior distribution type		Difference	Standard error	*t*	*p*
ML	Normal	Left skewed	-12.330	3.059	-4.031	0.004^{**}
		Right skewed	-11.482		-3.754	0.011^{*}
		Leptokurtic	-0.659		-0.216	0.999
		Platykurtic	-1.584		-0.518	0.999
	Left skewed	Right skewed	0.847		0.277	0.999
		Leptokurtic	11.670		3.815	0.009^{**}
		Platykurtic	10.746		3.513	0.024^{*}
	Right skewed	Leptokurtic	10.823		3.538	0.022^{*}
		Platykurtic	9.898		3.236	0.050^{*}
	Leptokurtic	Platykurtic	-0.925		-0.302	0.999
Bayes	Normal	Left skewed	-1.162		-0.380	0.999
		Right skewed	-1.137		-0.372	0.999
		Leptokurtic	-0.329		-0.108	0.999
		Platykurtic	-0.061		0.020	0.999
	Left skewed	Right skewed	0.025		0.008	0.999
		Leptokurtic	0.833		0.272	0.999
		Platykurtic	1.101		0.360	0.999
	Right skewed	Leptokurtic	0.808		0.264	0.999
		Platykurtic	1.076		0.352	0.999
	Leptokurtic	Platykurtic	0.268		0.088	0.999
ML*Bayes	Normal	Normal	-0.252		-0.082	0.999
	Left skewed	Left skewed	10.916		3.569	0.020^{*}
	Right skewed	Right skewed	10.093		3.300	0.044^{*}
	Leptokurtic	Leptokurtic	0.078		0.026	0.999
	Platykurtic	Platykurtic	1.271		0.416	0.999

$^{*}p< .05$, $^{**}p< .01$

Figure 8. *The change of item discrimination parameter RMSE in 2 PL model by priori distribution type.*

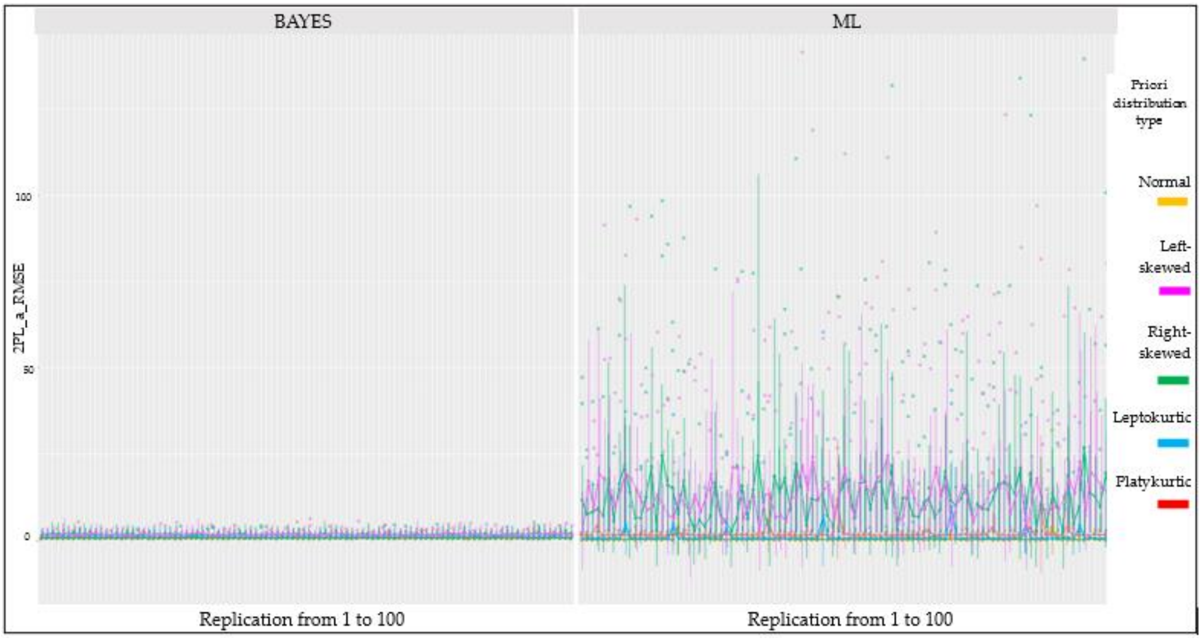


Figure 8 shows that the item discrimination parameter RMSE (a_{RMSE}) is higher as the priori distribution becomes skewed in the 2 PL model. In the ML estimation method, the item discrimination parameter RMSE (a_{RMSE}) increases as the priori distribution becomes skewed to the left or right. The leptokurtic or platykurtic of the prior distribution does not have an increasing effect on the item discrimination parameter RMSE. However, the Bayesian estimation method reduced the high RMSE of the item discrimination parameter if the priori ability distribution was skewed to the left or right. Accordingly, in the 2 PL model, the item discrimination parameter RMSE (a_{RMSE}) is affected by the differentiation of the priori distribution type. As a result, it shows low values when using the Bayesian estimation method.

3.4. Investigation of a_{RMSE}, b_{RMSE}, c_{RMSE} Estimated by ML and Bayesian Methods in 3 PL Model

RMSE changes of item parameters according to sample size, test length, and estimation method in 3 PL models with normal and non-normal (left skewed, right skewed, leptokurtic, and platykurtic) priori distribution stated in the fourth problem of the research were analyzed by mixed model ANOVA. Accordingly, the mixed model ANOVA results were performed for the item parameters according to sample size, test length, and estimation method in the data in 3 PL models with normal and non-normal priori distribution (left skewed, right skewed, leptokurtic, and platykurtic) are given in Table 13.

Table 13 shows that according to the mixed model ANOVA results for the item discrimination parameter RMSE in the data in the 3 PL models with normal and non-normal priori distribution, the main effects of independent variables as estimation method ($F_{(1, 88)} = 28.61$; $p<.01$, $\eta^2 = .203$), sample size ($F_{(2, 87)} = 4.55$; $p<.05$, $\eta^2 = .078$) and priori distribution type ($F_{(4, 85)} = 6.40$; $p<.01$, $\eta^2 = .192$) had significant effects. Test length ($F_{(2, 87)} = 0.53$; $p>.05$, $\eta^2 = .010$) did not show a significant difference. Among the independent variables found to be statistically significant, the estimation method is high, the sample size is medium, and the priori ability distribution type has a high effect size. In the 3 PL model, sample size ($F_{(2, 84)} = 5.22$; $p<.01$, $\eta^2 = .085$) has a significant and moderate effect size and priori distribution type ($F_{(4, 80)} = 13.46$; $p<.01$, $\eta^2 = .295$) has a significant and high effect size on item discrimination parameter RMSE. According to the mixed model ANOVA results for the item difficulty parameter RMSE in the 3 PL models with normal and non-normal priori distribution, none of the independent variables created a significant difference.

According to the mixed model ANOVA results for lower asymptote parameter RMSE in the data in 3 PL models with normal and non-normal priori distribution, estimation method ($F_{(1, 88)} = 9.10$; $p<.01$, $\eta^2 = .074$) and priori distribution type ($F_{(4, 80)} = 13.00$; $p<.01$, $\eta^2 = .306$) as the main effects of independent variables created significant differences. Sample size ($F_{(2, 87)} = 2.49$; $p>.05$, $\eta^2 = .043$) and test length ($F_{(2, 87)} = 0.50$; $p>.05$, $\eta^2 = .009$) were not significantly different. The estimation method that created a significant difference had a medium effect size, and the priori distribution type had a high effect size. In the 3 PL model, the estimation method from interactions and priori distribution type ($F_{(4, 80)} = 4.11$; $p<.01$, $\eta^2 = .117$) had a significant and medium effect size on lower asymptote parameter RMSE. Pairwise comparisons of the estimation method's significant difference in the item discrimination parameter for the data in the 3 PL model are given in Table 14.

Table 13. *Mixed model ANOVA results for item parameters RMSE in 3 PL models with normal and non-normal priori distribution.*

Independent variables	Mean squares of error	Degrees of freedom	F	*p*	Generalized η ^{2}
Item discrimination (a_{RMSE})					
Estimation method (K)	98140.76	1	28.61	0.001^{**}	0.203
Sample size (S)	119088.76	2	4.55	0.013^{*}	0.078
Test length (M)	129940.17	2	0.53	0.588	0.010
Prior distribution type (D)	103482.30	4	6.40	0.001^{**}	0.192
K*S	79972.48	2	5.22	0.007^{**}	0.085
K*M	99643.32	2	0.64	0.530	0.012
K*D	44745.66	4	13.46	0.001^{**}	0.295
Error	273.87	198			
Total	675287.32				
Item difficulty (b_{RMSE})					
Estimation method (K)	1149170.54	1	1.82	0.180	0.001
Sample size (S)	1132417.18	2	2.08	0.132	0.001
Test length (M)	117180079	2	0.54	0.582	0.001
Prior distribution type (D)	1106887.16	4	2.06	0.093	0.001
K*S	1095077.60	2	2.03	0.138	0.001
K*M	1173689.34	2	0.54	0.586	0.001
K*D	1037641.31	4	2.16	0.081	0.001
Error	797927.50	198			
Total	8664611.42				
Lower asymptote (c_{RMSE})					
Estimation method (K)	0.06	1	9.10	0.003^{**}	0.074
Sample size (S)	0.06	2	2.49	0.089	0.043
Test length (M)	0.07	2	0.50	0.606	0.009
Prior distribution type (D)	0.04	4	13.00	0.001^{**}	0.306
K*S	0.06	2	2.11	0.127	0.037
K*M	0.06	2	0.32	0.727	0.006
K*D	0.03	4	4.11	0.004^{**}	0.117
Error	0.00	198			
Total	0.38				

$^{*}p< .05$, $^{**}p< .01$

Table 14. *Pairwise comparisons of item discrimination parameter RMSE in 3 PL model by method of estimation.*

Estimation method	Difference	Standard error	*t*	*p*
Bayes-ML	-35.323	6.604	-5.348	0.001^{**}

$^{*}p< .05$, $^{**}p< .01$

Table 14 shows that the item discrimination parameter RMSE of the data in the 3 PL models with normal and non-normal priori distribution were significant in favor of the Bayesian estimation method (t=-5.348; p<.01). Bayesian estimation method produced lower RMSE than the ML estimation method. RMSE changes of the item discrimination parameter (aRMSE) estimation methods in the 2 PL model are given in Figure 9.

Figure 9. *The change of item discrimination parameter RMSE values in 3 PL model by estimation methods.*

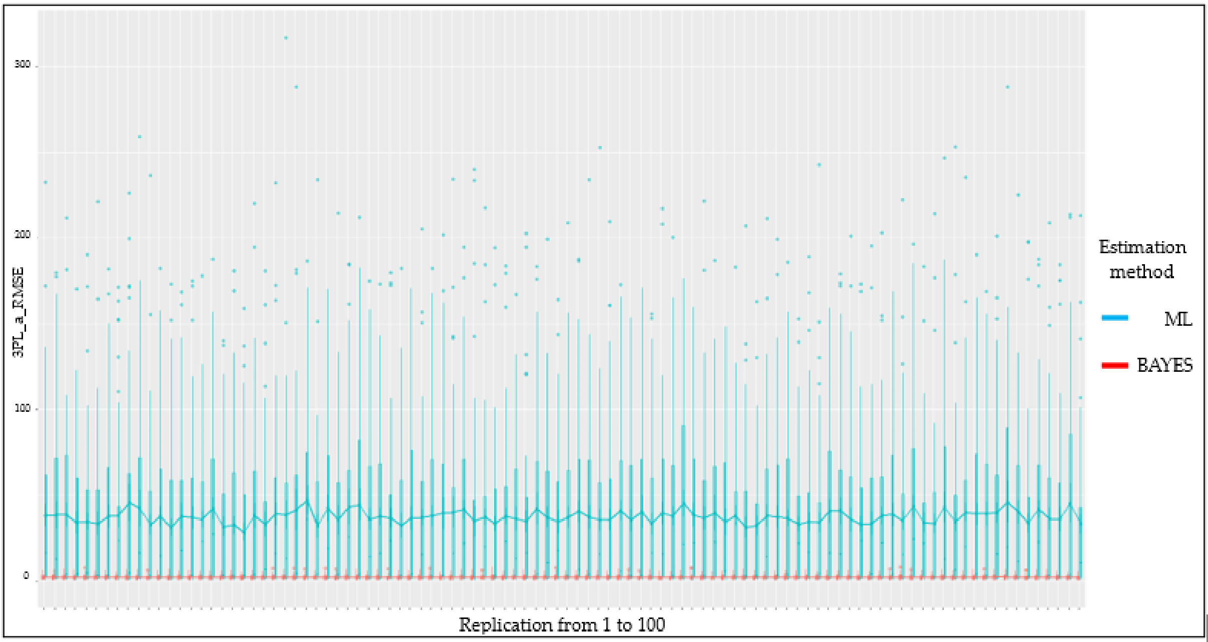


Figure 9 shows that the item discrimination parameter RMSE (a_{RMSE}) in the 3 PL model, independent of all simulation conditions, takes lower Bayesian estimation values than ML estimation. For the data in the 3 PL model, the pairwise comparisons of the sample size having a significant effect on the item discrimination parameter RMSE are given in Table 15.

Table 15. *Pairwise comparisons of item discrimination parameter RMSE values by sample size and estimation method in 3 PL model.*

Estimation method	Sample size		Difference	Standard error	*t*	*p*
ML	100	500	40.622	10.326	3.934	0.002**
		1000	46.255		4.479	0.001**
	500	1000	5.633		0.546	0.994
Bayes	100	500	2.711		0.263	0.999
		1000	2.941		0.285	0.999
	500	1000	0.231		0.022	0.999
ML*Bayes	100	100	62.398		6.043	0.001**
	500	500	24.487		2.371	0.178
	1000	1000	19.085		1.848	0.441

$^{*}p< .05$, $^{**}p< .01$

Table 15 shows that there is a significant difference between the item discrimination parameter RMSE estimated by ML method in the 3 PL model between sample sizes 100 and 500 (*t*=3.934; *p*<.01) and 100 and 1000 (*t*=4.479; *p*<.01), but no significant difference between 500 and 1000 (*t*=0.546; *p*>.05). However, using Bayes as the estimation method eliminated the significant differences between the sample sizes. Accordingly, using the Bayesian estimation method to estimate the item discrimination parameter more accurately in 3 PL models and small samples is more appropriate. Supporting this, when the sample size was 100 (*t*=6.043; *p*<.01), a significant difference was found between the RMSE of the item discrimination parameter according to the ML and Bayesian estimation method analyses. However, this significant difference is not observed as the sample size increases. RMSE change according to sample size on item discrimination parameter in the 3 PL model is given in Figure 10.

Figure 10. *The change of item discrimination parameter RMSE by sample size in the 3 PL model.*

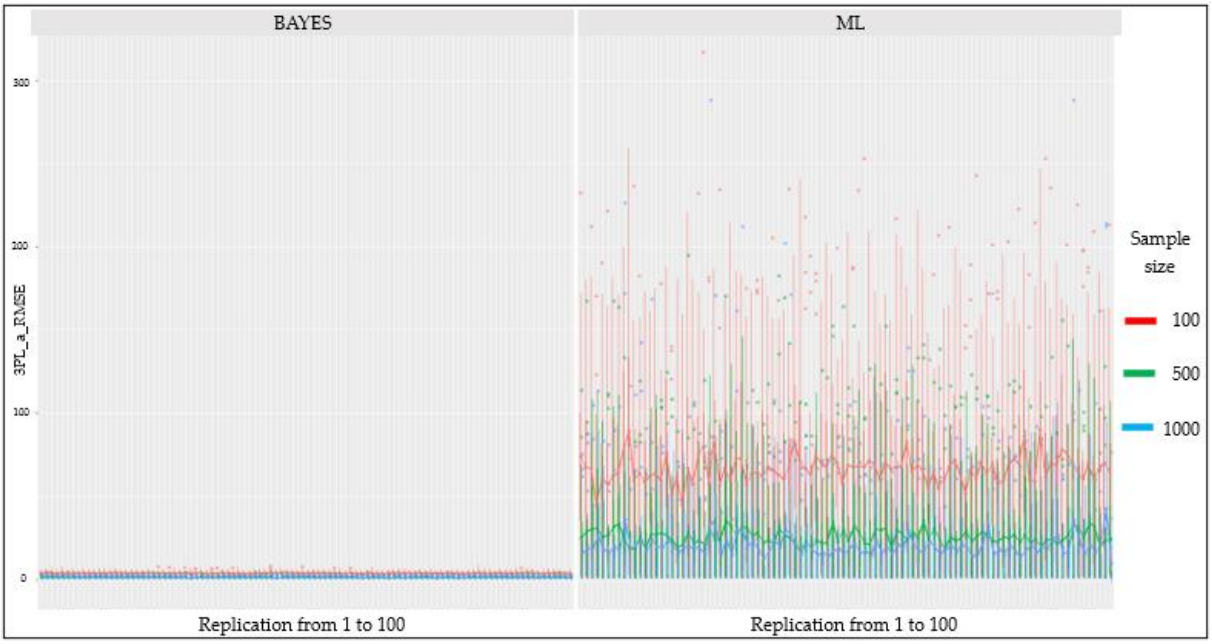


Figure 10 shows that when the sample size decreased, the Bayesian estimation method produced lower RMSE on item discrimination parameters than ML estimation. In ML estimation, item discrimination RMSE increases as the sample size decreases. In addition, these values show scattering. This situation is similar to the results obtained in the 2 PL model. These values decrease as the sample size increases. However, the Bayesian method tends to reduce the item discrimination parameter RMSE compared to the ML method. In Bayesian estimation, the increase in sample size did not make a difference in the item discrimination parameter RMSE (a_{RMSE}). RMSE obtained according to sample size is linear. In other words, the Bayesian method reduced and stabilized the item discrimination parameter RMSE (a_{RMSE}) compared to the ML estimation. Pairwise comparisons on the item discrimination parameter in the 3 PL model according to the priori distribution type are given in Table 16.

Table 16 shows that significant differences were found between the RMSE of the item discrimination parameter estimated by the ML method in the 3 PL model between normal and right-skewed (*t*=-8.852; *p*<.01), normal and leptokurtic (*t*=-4.516; *p*<.01), left-skewed and right-skewed (*t*=-7.960; *p*<.01), left-skewed and leptokurtic (*t*=-3.624; *p*<.05), right-skewed and leptokurtic (*t*=4.337; *p*<.01), right-skewed and platykurtic (*t*=8.400; *p*<.01), leptokurtic and platykurtic (*t*=4.063; *p*<.01) according to the distribution types. However, no significant difference was found between the a priori distribution types when the same parameter was estimated with the Bayesian method. Bayesian estimation method eliminated the significant difference depending on the a priori distribution type. Confirming this, the item discrimination parameter RMSE estimated by ML and Bayesian methods in the same a priori distribution types show a significant difference when the distribution is right skewed (*t*=9.274; *p*<.01) or leptokurtic (*t*=5.162; *p*<.01). Here, unlike in the 2 PL model, a distribution of a priori leptokurtic in the 3 PL model was found to cause differentiation. No differentiation was observed for the other distribution types. RMSE on item discrimination parameters in the 3 PL model according to priori distribution type is given in Figure 11.

Table 16. *Pairwise comparisons of item discrimination parameter RMSE in 3 PL model by priori distribution type and estimation method.*

Estimation method	Prior distribution type		Difference	Standard error	*t*	*p*
ML	Normal	Left skewed	-8.894	9.972	-0.892	0.996
		Right skewed	-88.270		-8.852	0.001^{**}
		Leptokurtic	-45.027		-4.516	0.001^{**}
		Platykurtic	-4.508		0.452	0.999
	Left skewed	Right skewed	-79.377		-7.960	0.001^{**}
		Leptokurtic	-36.134		-3.624	0.016^{*}
		Platykurtic	4.386		0.440	0.999
	Right skewed	Leptokurtic	43.243		4.337	0.001^{**}
		Platykurtic	83.763		8.400	0.001^{**}
	Leptokurtic	Platykurtic	40.520		4.063	0.004^{**}
Bayes	Normal	Left skewed	-0.536		-0.054	0.999
		Right skewed	-2.614		-0.262	0.999
		Leptokurtic	-0.374		-0.038	0.999
		Platykurtic	-0.653		-0.065	0.999
	Left skewed	Right skewed	-2.077		-0.208	0.999
		Leptokurtic	0.162		0.016	0.999
		Platykurtic	-0.117		-0.012	0.999
	Right skewed	Leptokurtic	2.239		0.225	0.999
		Platykurtic	1.961		0.197	0.999
	Leptokurtic	Platykurtic	-0.279		-0.028	0.999
ML*Bayes	Normal	Normal	6.819		0.684	0.999
	Left skewed	Left skewed	15.176		1.522	0.879
	Right skewed	Right skewed	92.476		9.274	0.001^{**}
	Leptokurtic	Leptokurtic	51.472		5.162	0.001^{**}
	Platykurtic	Platykurtic	10.673		1.070	0.986

$^{*}p< .05$, $^{**}p< .01$

Figure 11. *The change of item discrimination parameter RMSE in the 3 PL model by priori distribution types.*

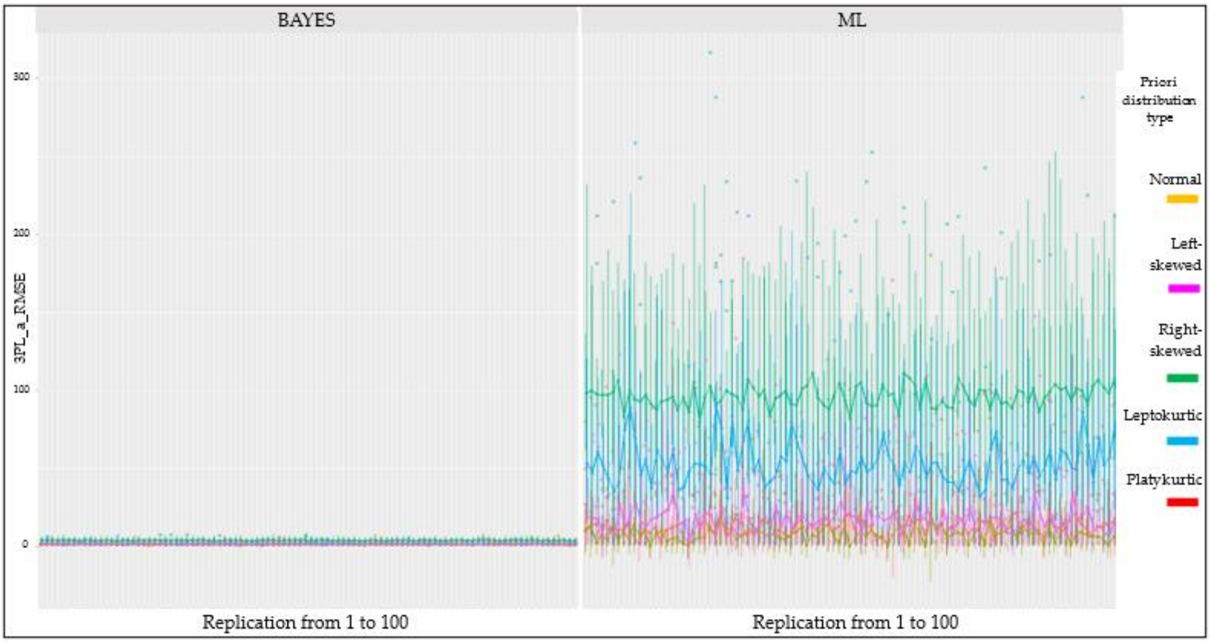


Figure 11 shows that the priori distribution becomes skewed in the 3 PL model, and the item discrimination parameter RMSE takes higher values. These high RMSE were reduced by the Bayesian estimation method. In the ML estimation in the 3 PL model, the item discrimination parameter RMSE gave the highest results when the priori distribution was skewed to the right. This was followed by leptokurtic, left skewed, platykurtic, and normal distributions. The fact that the model is 3 PL is an essential factor for the item discrimination parameter RMSE (a_{RMSE}) to be the highest when the priori distribution is skewed to the right. Unlike the 2 PL model, by adding a third parameter, the lower asymptote parameter (c_i) in this model changes the starting point of the priori distributions. Therefore, the most affected by this situation are the right-skewed priori parameters. When Bayesian estimation was used, the item discrimination parameter RMSE (a_{RMSE}) produced lower RMSE in all a priori distributions compared to ML estimation, which was stably distributed. Pairwise comparisons of the estimation method's significant effect on the lower asymptote parameter for the data in the 3 PL model are given in Table 17.

Table 17. *Pairwise comparisons of lower asymptote parameter RMSE in 3 PL model by estimation method.*

Estimation method	Difference	Standard error	*t*	*p*
Bayes-ML	0.016	0.005	3.016	0.003**

*p< .05, **p< .01

Table 17 shows that the lower asymptote parameter is significant and in favor of the Bayesian estimation method on RMSE in 3 PL models with normal and non-normal priori distribution (t=3.016; p<.01). Bayesian estimation method produced lower RMSE than the ML estimation method. RMSE changes of the lower asymptote parameter estimation methods in the 3 PL model are given in Figure 12.

Figure 12. *The change of the lower asymptote parameter RMSE in the 3 PL model by estimation methods.*

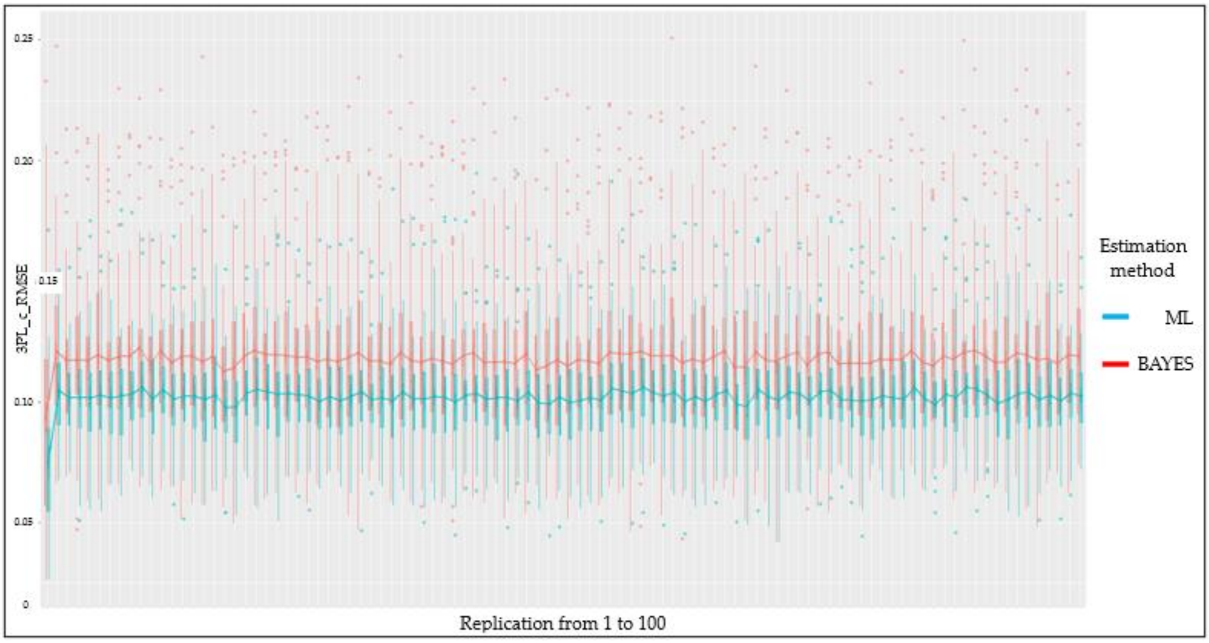


Figure 12 shows that the RMSE of the lower asymptote parameter in the 3 PL model takes higher values in Bayesian estimation regardless of the research conditions. Unlike other parameters, ML estimation was more effective than Bayesian estimation in decreasing the RMSE of the lower asymptote parameters. There are few studies on the lower asymptote parameter in the literature. This result is likely due to the distribution type defined for the lower asymptote parameter while creating the function for the priori distribution. The data in the 3 PL model's pairwise comparisons of the priori distribution type that significantly affect the lower asymptote parameter is given in Table 18.

Table 18. *Pairwise comparisons of the lower asymptote parameter RMSE in the 3 PL model by priori distribution type and estimation method.*

Estimation method	Prior distribution type		Difference	Standard error	*t*	*p*
ML	Normal	Left skewed	0.017	0.009	2.012	0.593
		Right skewed	-0.007		-0.765	0.998
		Leptokurtic	-0.015		-1.742	0.768
		Platykurtic	-0.003		-0.333	0.999
	Left skewed	Right skewed	-0.024		-2.777	0.163
		Leptokurtic	-0.032		-3.754	0.011^{*}
		Platykurtic	-0.020		-2.346	0.372
	Right skewed	Leptokurtic	-0.008		-0.977	0.993
		Platykurtic	0.004		0.431	0.999
	Leptokurtic	Platykurtic	0.012		1.409	0.921
Bayes	Normal	Left skewed	0.043		5.102	0.001^{**}
		Right skewed	0.024		2.816	0.149
		Leptokurtic	-0.023		-2.693	0.194
		Platykurtic	0.020		2.402	0.339
	Left skewed	Right skewed	-0.019		-2.286	0.509
		Leptokurtic	-0.066		-7.795	0.001^{**}
		Platykurtic	0.043		5.094	0.001^{**}
	Right skewed	Leptokurtic	-0.047		-5.509	0.001^{**}
		Platykurtic	-0.004		-0.414	0.999
	Leptokurtic	Platykurtic	0.043		5.094	0.001^{**}
ML*Bayes	Normal	Normal	-0.030		-3.544	0.022^{*}
	Left skewed	Left skewed	-0.004		-0.454	0.999
	Right skewed	Right skewed	0.000		0.037	0.999
	Leptokurtic	Leptokurtic	-0.038		-4.494	0.001^{**}
	Platykurtic	Platykurtic	-0.007		-0.809	0.998

$^{*}p< .05$, $^{**}p< .01$

Table 18 shows a significant difference between the lower asymptote parameter RMSE values estimated by ML method in 3 PL models between left skewed and leptokurtic (*t*=-3.754; *p*<.05) according to distribution types. In Bayesian estimation, there is a significant difference between normal and left skewed (*t*=5.102; *p*<.01), left skewed and leptokurtic (*t*=-7.795; *p*<.01), left skewed and platykurtic (*t*=5.094; *p*<.01), right skewed and leptokurtic (*t*=-5.509; *p*<.01), leptokurtic and platykurtic (*t*=5.094; *p*<.01) according to distribution types. As with the other parameters, no significance is expected for this parameter. However, the advantages of Bayesian estimation over ML estimation were not observed at lower asymptote parameters. The lower asymptote parameter RMSE estimated by ML and Bayesian methods in the same priori distribution types showed a significant difference in the normal (*t*=-3.544; *p*<.05) and leptokurtic (*t*=-4.494; *p*<.01) distributions. No difference was observed in other distribution types. Lower asymptote parameter RMSE according to the priori ability distribution type in the 3 PL model are given in Figure 13.

Figure 13. *The change of the lower asymptote parameter RMSE in the 3 PL model by priori distribution types.*

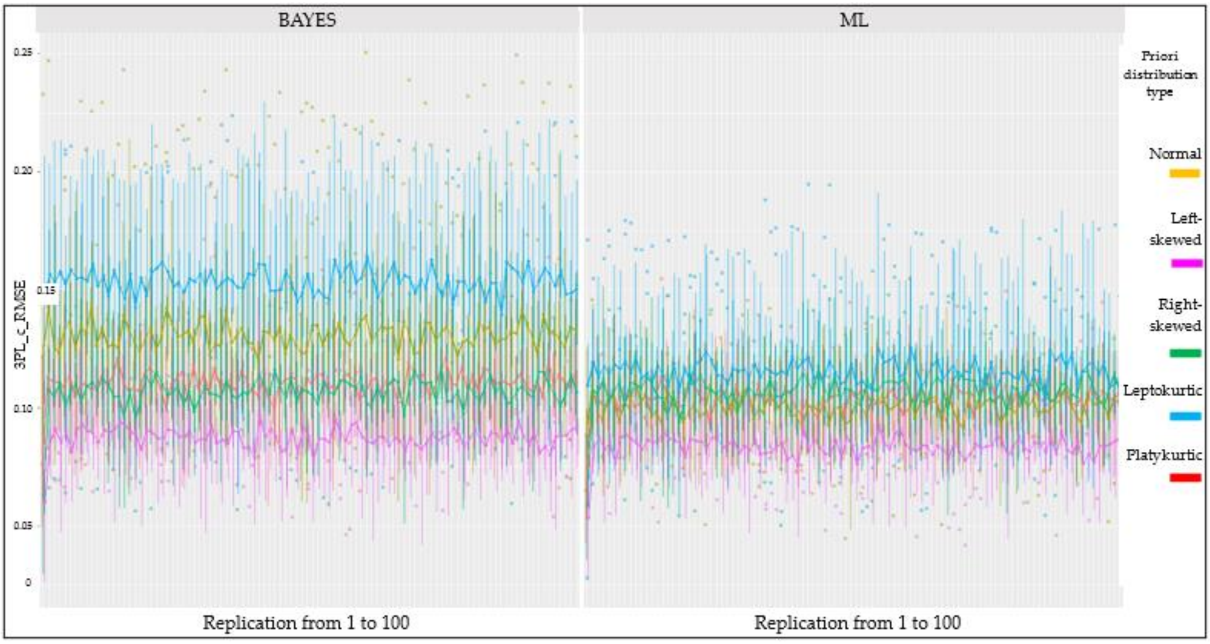


Figure 13 shows that the priori distribution becomes skewed in the 3 PL model; the lower asymptote parameter RMSE takes higher values. However, as the type of priori distribution becomes leptokurtic, the lower asymptote parameter RMSE increases in Bayesian estimation, unlike the other item parameters. Accordingly, the ML estimation method produced lower RMSE as the priori distribution became leptokurtic in the 3 PL model. In addition, the RMSE obtained in the ML estimation for all priori distribution types was distributed in a narrower area than Bayesian estimation. The lower asymptote parameter RMSE (c_{RMSE}) obtained from Bayesian estimation is spread over a wider area because the initial parameter values are generated with a distribution other than the normal distribution. Standard Bayesian estimations tend to normalize the posterior distribution because the priori distribution is normal.

4. DISCUSSION and CONCLUSION

Considering the conditions in all the problems of the research, in the first research problem in which the RMSE of the ability parameters were examined, In the data in the 2 PL model, the estimation method on the RMSE of the ability parameters, test length, the type of priori distribution, and the interaction between estimation method and the priori distribution type created significant differentiation. These results are like the results of Finch and Edwards (2015) when examined in general terms. Likewise, Bayesian estimations give more accurate results in cases where the latent feature is non-normally distributed in the 2 PL model. A similar situation in terms of test length is also seen in Köse (2010)'s study. The change in test length affects the estimation results in ability parameters. An increase in test length decreases the RMSE of ability parameters.

In the second research problem, test length and priori distribution type created significant differences in the RMSE of the ability parameters in the data in the 3 PL model. The general results for this problem are like the results of Swaminathan and Gifford (1986). They suggested that their study use Bayesian estimation instead of ML for the 3 PL model. In addition, Karadavut (2019) stated in her research that when estimating the ability parameter in the 3 PL model, not knowing the priori distribution type would lead to erroneous estimations. A similar situation can be seen in this study's differentiation of the priori distribution type.

In estimating ability parameters and RMSE, the estimation method made a significant difference only in 2 PL models. This significance is in favor of the Bayesian estimation method.

Because Bayesian estimation reduced the high error values obtained in ML to lower values. Similar results were obtained in studies in the literature (Swaminathan & Gifford, 1986; Harwell & Janosky, 1991; Gao & Chen, 2005; Finch & Edwards, 2015).

In the third research problem, in which item parameters RMSE were examined, estimation method, sample size, priori distribution type, the interaction of estimation method and sample size, and interaction of estimation method and priori distribution type on item discrimination parameter RMSE in the data in the 2 PL model created significant differences. In the 2 PL model data, no condition caused a significant difference in the RMSE of the item difficulty parameter. These results are like Harwell and Janosky's (1991) results. Accordingly, Bayesian estimation is considered sufficient for small samples and short tests in the 2 PL model. It is stated in Stone's (1992) study that as the priori distribution for the item discrimination parameter becomes skewed, the bias in the ML estimation increases. In this study, the RMSE for the item discrimination parameter is also affected by the skewness of the prior distribution type. In this respect, these two studies showed similar results. It is also seen in the study of Sass et al. (2008) that item parameters are affected by priori distribution and produce high error values.

In the fourth research problem, the estimation method, sample size, priori distribution type, estimation method and sample size interaction, and estimation method and priori distribution type interaction on the item discrimination parameter RMSE in the data in the 3 PL model created significant differences. In the 3 PL model data, no conditions were significant on the item difficulty parameter RMSE. However, in the 3 PL model data, the estimation method on the RMSE of the lower asymptote parameter, the priori distribution type, and the interaction of the estimation method and the priori distribution type created significant differences. When these results are examined, it is seen that the suggestion of Swaminathan and Gifford (1986) is correct. Accordingly, this related research proposes the Bayesian method for parameter estimation for the 3 PL model. In this study, using the Bayesian estimation method in estimating item parameters in the 3 PL model, especially in the item discrimination parameter, provides an advantage. Likewise, as in the study of Gao and Chen (2005), Bayesian estimation gave more precise results in estimating item parameters when the sample size decreased to 100.

In estimating item parameters and RMSE, the estimation method generally showed a significant differentiation. This differentiation is significant for item discrimination RMSE (a_{RMSE}) and lower asymptote RMSE (c_{RMSE}) parameters regardless of the model. Bayesian estimation method for this significant differentiation item discrimination parameter; for the lower asymptote parameter, the ML estimation method is in favor. However, according to the estimation method for the item difficulty RMSE (b_{RMSE}) parameter, there is no differentiation between 2 PL and 3 PL models. This situation in the item difficulty parameter yielded similar results to the study of Kıbrıslıoğlu Uysal (2020).

While the sample size did not make a significant difference in estimating the ability parameter, the test length, the priori distribution type, and the estimation method (only in the 2 PL model) created significant differences in the RMSE. The sample size does not affect the ability of parameter estimation and error values because the number of estimated parameters is only one. This is similar to the research of Goldman and Raju (1986) and Harwell and Janosky (1991). The study of Goldman and Raju (1986) stated that the sample size of 250 would be sufficient when the estimated parameters were reduced to 1. Harwell and Janosky (1991) concluded that samples of 15 items and 250 people were sufficient. A similar situation is seen in the study of Şahin and Anıl (2017). Şahin and Anıl (2017) concluded that a sample of 150 people would be sufficient to make parameter estimation in 1 PL model.

The sample size was only effective in the RMSE estimations of the item discrimination parameter. This applies when both the 2 PL and 3 PL models are used. The increase in sample size positively affected the item discrimination parameter RMSE (a_{RMSE}), and these values decreased. However, as the sample size decreased, especially the RMSE of the item discrimination parameter showed excessive swelling. The swelling in the RMSE of the item discrimination parameter (a_{RMSE}) due to estimation with the ML method was also seen in the studies of Chuah et al. (2006). However, the Bayesian estimation method played an important role in reducing this swelling. A similar situation is seen in the study of Gao and Chen (2005). In this study, it has been stated that Bayesian estimations give more accurate results than marginal maximum likelihood estimations when the sample size drops to 100.

Increasing the test length only decreased the ability parameter RMSE (Θ_{RMSE}). Moreover, in some cases where the test length is 40, the results of the ML and Bayesian methods for estimating ability have taken values close to each other. Similarly, Gao and Chen (2005) emphasized in their study that increasing test length and sample size tends to reduce the standard errors of estimations. However, when the test length decreased to 10, it caused swelling in the RMSE of the ability parameters in the ML estimation. However, this situation was reduced by the Bayesian estimation method. Item discrimination (a_i), item difficulty (b_i), and lower asymptote (ci) parameters RMSE were not affected in any way by the test length change.

The priori distribution type ability parameters have significant differences in RMSE. According to the logistic model, the priori distribution type did not significantly differ in ability parameters. In both 2 PL and 3 PL models, the priori distribution type, item discrimination (a_i), and lower asymptote (c_i) parameters showed a significant difference in RMSE. In 2 PL and 3 PL models, there was no significant difference in item difficulty parameter RMSE (b_{RMSE}) values according to the priori distribution type. Differentiation of item parameters according to priori distribution type is more significant on the left and right skewed distributions than other distribution types. In a similar study conducted by Doğan (2002), distribution types (skewed or leptokurtic and platykurtic) affected the parameter invariance of the IRT. It was stated that the differentiation was higher in skewed distributions. A similar situation is observed in the studies of Seong (1990), Stone (1992), Kirisci et al. (2001), Sass et al. (2005) and Karadavut (2019).

The logistic model was significant on the RMSE of ability and item parameters. The 3 PL model produced higher prediction RMSE than the 2 PL model. The Bayesian estimation method decreased these values more than the ML.

The parameter estimation method, ability, and item parameters created a significant difference in the RMSE in different conditions that constitute the research's aim. In addition, it was shown that the Bayesian estimation method obtained lower RMSE than the ML estimation method in all simulation conditions. However, the significance of these RMSEs was observed in only some simulation conditions.

RMSE is the total error indicator of parameter estimation's precision and estimation bias (Thissen & Wainer, 1983). When the literature was reviewed, the standard errors of parameter estimation for commonly used models (Rasch, 1 PL, 2 PL, and 3 PL) needed to be comprehensively addressed (Lord, 1980). As stated in the study results, the Bayesian method reduced the RMSE of ability and item parameters to lower levels than the ML method.

Accordingly, the Bayesian estimation method seems advantageous since it produces lower parameter RMSE than the ML estimation method. Moreover, especially when the ML estimation method is used, it is seen that it tends to reduce the excessive increase in parameter RMSE that occurs in small samples and short tests.

Nowadays, it is possible to use IRT to develop classroom achievement tests. However, the first issue is how to do this with small samples and short tests. The Bayesian approach makes this possible and reduces the estimation errors to acceptable levels. In addition, it is only sometimes possible for the distribution under study to be normal. The ML estimation method does not give accurate results in such a case. At this point, the advantages of Bayesian estimation are utilized. The results of this study show that Bayesian estimation can be offered as a solution where ML estimation cannot obtain accurate results.

Acknowledgments

This research article was produced from the doctoral dissertation of first author under the supervision of second author.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Ankara University, 04.11.2019, 13-339.

Contribution of Authors

Eray Selçuk: Investigation, Resources, Visualization, Software, Coding, Analysis, and Writing-original draft. **Ergül Demir**: Investigation, Methodology, Supervision, Critical Review and Validation.

Orcid

Eray Selçuk https://orcid.org/0000-0003-4033-4219
Ergül Demir https://orcid.org/0000-0002-3708-8013

REFERENCES

Akour, M., & Al-Omari, H. (2013). Empirical investigation of the stability of IRT item-parameters estimation. *International Online Journal of Educational Sciences, 5*(2), 291-301.

Atar, B. (2007). *Differential item functioning analyses for mixed response data using IRT likelihood-ratio test, logistic regression, and GLLAMM procedures* [Unpublished doctoral dissertation, Florida State University]. http://purl.flvc.org/fsu/fd/FSU_migr_etd-0248

Baker, F.B. (2001). *The basics of item response theory* (2nd ed.). College Park, (MD): ERIC Clearinghouse on Assessment and Evaluation.

Baker, F.B., & Kim, S. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). Marcel Dekker.

Barış-Pekmezci, F. & Şengül-Avşar, A. (2021). A guide for more accurate and precise estimations in simulative unidimensional IRT models. *International Journal of Assessment Tools in Education*, *8*(2), 423-453. https://doi.org/10.21449/ijate.790289

Bilir, M.K. (2009). *Mixture item response theory-mimic model: Simultaneous estimation of differential item functioning for manifest groups and latent classes* [Unpublished doctoral dissertation, Florida State University]. http://diginole.lib.fsu.edu/islandora/object/fsu:182011/datastream/PDF/view

Bock, R.D., & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431-444. https://doi.org/10.1177/014662168200600405

Bulmer, M.G. (1979). *Principles of statistics*. Dover Publications.

Bulut, O. & Sünbül, Ö. (2017). R programlama dili ile madde tepki kuramında monte carlo simülasyon çalışmaları [Monte carlo simulation studies in item response theory with the R programming language]. *Journal of Measurement and Evaluation in Education and Psychology*, *8*(3), 266-287. https://doi.org/10.21031/epod.305821

Chalmers, R.P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Chuah, S.C., Drasgow F., & Luecht, R. (2006). How big is big enough? Sample size requirements for cast item parameter estimation. *Applied Measurement in Education*, *19*(3), 241-255. https://doi.org/10.1207/s15324818ame1903_5

Clarke, E. (2022, December 22). ggbeeswarm: Categorical scatter (violin point) plots. https://cran.r-project.org/web/packages/ggbeeswarm/index.html

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

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

Çelikten, S. & Çakan, M. (2019). Bayesian ve nonBayesian kestirim yöntemlerine dayalı olarak sınıflama indekslerinin TIMSS-2015 matematik testi üzerinde incelenmesi [Investigation of classification indices on Timss-2015 mathematic-subtest through bayesian and nonbayesian estimation methods]. *Necatibey Faculty of Education Electronic Journal of Science and Mathematics Education, 13*(1), 105-124. https://doi.org/10.17522/balikesirnef.566446

De Ayala, R.J. (2009). *The theory and practice of item response theory.* The Guilford Press.

DeMars, C. (2010). *Item response theory: understanding statistics measurement.* Oxford University Press.

Demir, E. (2019). *R Diliyle İstatistik Uygulamaları [Statistics Applications with R Language].* Pegem Akademi.

Feinberg, R.A., & Rubright, J.D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice, 35,* 36-49. https://doi.org/10.1111/emip.12111

Finch, H., & Edwards, J.M. (2016). Rasch model parameter estimation in the presence of a non-normal latent trait using a nonparametric Bayesian approach. *Educational and Psychological Measurement, 76*(4), 662-684. https://doi.org/10.1177/0013164415608418

Fraenkel, J.R., & Wallen, E. (2009). *How to design and evaluate research in education.* McGraw-Hills Companies.

Gao, F., & Chen, L. (2005). Bayesian or non-Bayesian: A comparison study of item parameter estimation in the three-parameter logistic model. *Applied Measurement in Education, 18*(4), 351-380. https://psycnet.apa.org/doi/10.1207/s15324818ame1804_2

Goldman, S.H., & Raju, N.S. (1986). Recovery of one- and two-parameter logistic item parameters: An empirical study. *Educational and Psychological Measurement, 46*(1), 11-21. https://doi.org/10.1177/0013164486461002

Hambleton, R.K. (1989). Principles and selected applications of item response theory. In R.L. Linn (Ed.), *Educational Measurement,* (pp.147-200). American Council of Education.

Hambleton, R.K., & Cook, L.L. (1983). Robustness of ítem response models and effects of test length and sample size on the precision of ability estimates. In D.J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 31-49). Vancouver.

Hambleton, R.K., & Jones, R.W. (1993). An NCME instructional module on comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38-47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principals and applications.* Kluwer Academic Publishers.

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

Harwell, M., & Janosky, J. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement. 15*, 279-291. https://doi.org/10.1177/014662169101500308

Harwell, M., Stone, C.A., Hsu, T.C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101-125. https://doi.org/10.1177/014662169602000201

Hoaglin, D.C., & Andrews, D.F. (1975). The reporting of computation-based results in statistics. *The American Statistician, 29,* 122-126. https://doi.org/10.2307/2683438

Hulin, C.L., Lissak, R.I., & Drasgow, F. (1982). Recovery of two and three-parameter logistic item characteristic curves: A monte carlo study. *Applied Psychological Measurement, 6,* 249-260. https://psycnet.apa.org/doi/10.1177/014662168200600301

Karadavut, T. (2019). The uniform prior for bayesian estimation of ability in item response theory models. *International Journal of Assessment Tools in Education, 6*(4), 568-579. https://dx.doi.org/10.21449/ijate.581314

Kıbrıslıoğlu Uysal, N. (2020). *Parametrik ve Parametrik Olmayan Madde Tepki Modellerinin Kestirim Hatalarının Karşılaştırılması [Comparison of estimation errors in parametric and nonparametric item response theory models]* [Unpublished doctoral dissertation, Hacettepe University]. http://hdl.handle.net/11655/22495

Kirisci, L., Hsu, T.C., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25*(2), 146-162. https://doi.org/10.1177/01466210122031975

Kolen, M.J. (1985). Standard errors of Tucker equating. *Applied Psychological Measurement, 9*(2), 209-223. https://doi.org/10.1177/014662168500900209

Kothari, C.R. (2004). *Research methodology: methods and techniques* (2nd ed.). New Age International Publishers.

Köse, İ.A. (2010). *Madde Tepki Kuramına Dayalı Tek Boyutlu ve Çok Boyutlu Modellerin Test Uzunluğu ve Örneklem Büyüklüğü Açısından Karşılaştırılması [Comparison of Unidimensional and Multidimensional Models Based On Item Response Theory In Terms of Test Length and Sample Size]* [Unpublished doctoral dissertation]. Ankara University, Institute of Educational Sciences.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *frontiers in Psychology, 4*(863), 1-12. https://doi.org/10.3389/fpsyg.2013.00863

Lenth, R.V. (2022, December). emmeans: Estimated marginal means, aka Least-Squares Means. https://cran.r-project.org/web/packages/emmeans/index.html

Lim, H., & Wells, C.S. (2020). irtplay: An R package for online item calibration, scoring, evaluation of model fit, and useful functions for unidimensional IRT. *Applied psychological measurement, 44*(7-8), 563-565. https://doi.org/10.1177/0146621620921247

Linacre, J.M. (2008). *A user's guide to winsteps ministep: rasch-model computer programs.* https://www.winsteps.com/winman/copyright.htm

Lord, F.M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28,* 989-1020. https://doi.org/10.1177/001316446802800401

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Lawrence Erlbaum Associates.

Lord, F.M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel forms reliability. *Psychometrika, 48,* 233-245. https://doi.org/10.1007/BF02294018

Martin, A.D., & Quinn, K.M. (2006). Applied Bayesian inference in R using MCMCpack. *The Newsletter of the R Project, 6*(1), 2-7.

Martinez, J. (2017, December 1). bairt: Bayesian analysis of item response theory models. http://cran.nexr.com/web/packages/bairt/index.html

Maydeu-Ovivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrica, 71,* 713-732. https://doi.org/10.1007/s11336-005-1295-9

Meyer, D. (2022, December 1). e1071: Misc functions of the department of statistics, Probability Theory Group (Formerly: E1071), TU Wien. https://CRAN.R-project.org/package=e1071

Mislevy, R.J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*, 177-195. https://doi.org/10.1007/BF02293979

MoNE (2022). *Sınavla Öğrenci Alacak Ortaöğretim Kurumlarına İlişkin Merkezî Sınav Başvuru ve Uygulama Kılavuzu [Central Examination Application and Administration Guide for Secondary Education Schools to Admit Students by Examination]*. Ankara: MoNE [MEB]. https://www.meb.gov.tr/2022-lgs-kapsamindaki-merkez-sinav-kilavuzu-yayimlandi/haber/25705/tr

Morris, T.P., White, I.R., & Crowther, M.J. (2017). Using simulation studies to evaluate statistical methods. *Tutorial in Biostatistics, 38*(11), 2074-2102. https://doi.org/10.1002/sim.8086

Orlando, M. (2004, June). Critical issues to address when applying item response theory models. *Paper presented at the conference on improving health outcomes assessment*, National Cancer institute, Bethesda, MD, USA.

Pekmezci Barış, F. (2018). *İki Faktör Modelde (Bifactor) Diklik Varsayımının Farklı Koşullar Altında Sınanması [Investigation Of Orthogonality Assumption In Bifactor Model Under Different Conditions]* [Unpublished doctoral dissertation]. Ankara University, Institute of Educational Sciences, Ankara.

Ree, M.J., & Jensen, H.E. (1980). Effects of sample size on linear equating of item characteristic curve parameters. In D.J. Weiss (Ed.), *Proceedings of the 1979 computerized adaptive testing conference*. (pp. 218-228). Minneapolis: University of Minnesota. https://doi.org/10.1016/B978-0-12-742780-5.50017-2

Reise, S.P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*, 133-144. https://doi.org/10.1111/j.1745-3984.1990.tb00738.x

Revelle, W. (2022, October). psych: Procedures for psychological, psychometric, and personality research. https://cran.r-project.org/web/packages/psych/index.html

Robitzsch, A. (2022). sirt: Supplementary item response theory models. https://cran.r-project.org/web/packages/sirt/index.html

Samejima, F. (1993a). An approximation for the bias function of the maximum likelihood estimate of a latent variable for the general case where the item responses are discrete. *Psychometrika, 58*, 119-138. https://doi.org/10.1007/BF02294476

Samejima, F. (1993b). The bias function of the maximum likelihood estimate of ability for the dichotomous response level. *Psychometrika, 58*, 195-209. https://doi.org/10.1007/BF02294573

Sarkar, D. (2022, October). *lattice: Trellis graphics for R.* R package version 0.20-45, URL http://CRAN.R-project.org/package=lattice.

SAS Institute (2020). Introduction to Bayesian analysis procedures. In *User's Guide Introduction to Bayesian Analysis Procedures.* (pp. 127-161). SAS Institute Inc., Cary, (NC), USA.

Sass, D., Schmitt, T., & Walker, C. (2008). Estimating non-normal latent trait distributions within item response theory using true and estimated item parameters. *Applied Measurement in Education, 21*(1), 65-88. https://doi.org/10.1080/08957340701796415

Seong, T.J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement, 14*(3), 299-311. https://psycnet.apa.org/doi/10.1177/014662169001400307

Singmann, H. (2022, December). afex: Analysis of factorial experiments. https://cran.r-project.org/web/packages/afex/afex.pdf

Soysal, S. (2017). *Toplam Test Puanı ve Alt Test Puanlarının Kestiriminin Hiyerarşik Madde Tepki Kuramı Modelleri ile Karşılaştırılması [Comparison of Estimation of Total Score and Subscores with Hierarchical Item Response Theory Models]* [Unpublished doctoral dissertation]. Hacettepe University, Institute of Educational Sciences, Ankara.

Stone, C.A. (1992). Recovery of marginal maximum likelihood estimates in the two parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16*(1), 1-16. https://doi.org/10.1177/014662169201600101

Swaminathan, H., & Gifford, J.A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika*, *51*, 589-601. https://doi.org/10.1007/BF02295598

Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice*, *17*, 321-335. http://dx.doi.org/10.12738/estp.2017.1.0270

Tabachnick, B.G., & Fidell, L.S. (2014). *Using multivariate statistics* (6th ed.). Pearson New International Edition.

Thissen, D., & Wainer, H. (1983). Some standard errors in item response theory. *Psychometrika*, *47*, 397-412. https://doi.org/10.1007/BF02293705

Thorndike, L.R. (1982). *Applied Psychometrics.* Houghton Mifflin Co.

Van de Schoot, R., & Depaoli, S. (2014). Bayesian analyses: Where to start and what to report. *The European Health Psychologist*, *16*(2), 75-84.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis.* Springer-Verlag. https://doi.org/10.1007/978-0-387-98141-3

Wright, B.D., & Stone, M.H. (1979). *Best test design.* Mesa Press

Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three parameter logistic model. *Applied Psychological Measurement*, *8*, 125-145. https://doi.org/10.1177/014662168400800201

APPENDIX

```
#ÖNSEL (PRIOR) SCRIPT BLOCK*
# Generation of necessary prior distributions and data sets according to simulation conditions
library(psych)
library(e1071)
library(mirt)

#I: number of items
#N: number of individuals
#M: number of parameters
#D: distribution state

prior <- function (I, N, M=c("2PL", "3PL"), D=c("normal","left-skewed","right-
                                skewed","leptokurtic", "platykurtic"))
{
a <- rlnorm(I, meanlog = 0.3, sdlog = 0.2)
b <- rnorm(I, mean = 0, sd = 1)
c <- runif(I, min = 0.01, max = 0.25)

if (D=="normal") {k <- as.matrix(rnorm(N, mean = 0, sd = 1))}

else if (D==" left-skewed") {k <- as.matrix(c(rnorm(N*86/100, 2, 1)), runif(N*7/100, min = -
5, max = -4), runif(N*7/100, min = -4, max = -3)))}

else if (D=="right-skewed ") {k <- as.matrix(c(rnorm(N*86/100, -2, 1)), runif(N*7/100, min
= 3, max = 4), runif(N*7/100, min = 4, max = 5)))}

else if (D=="leptokurtic ") {k <- as.matrix(c(rnorm(N*3/100, -1, 100), rnorm(N*94/100, 0,
        0.00001), rnorm(N*3/100, 1, 100)) )}

else if (D=="platykurtic") {k <- as.matrix(c(rnorm(N*40/100, 0, 1)), runif(N*30/100, min = -
3, max = -1), runif(N*30/100, min = 1, max = 3)))}

if (M=="2PL")

{dat <- as.data.frame(simdata(a = a, d = b, N = N, itemtype = "dich", Theta = k))

model2pl <- mirt(data = dat, 1, itemtype = "2PL", SE = TRUE, verbose = FALSE, technical =
list(NCYCLES = 10000))

irt.parameters <- as.data.frame(coef(model2pl, simplify = TRUE)$items)
bias.a <- mean(irt.parameters[,1]-a)
bias.b <- mean(irt.parameters[,2]-b)
rmse.a <- sqrt(mean((irt.parameters[,1]-a)^2))
rmse.b <- sqrt(mean((irt.parameters[,2]-b)^2))

fit2pl <- M2(model2pl)
M2 <- fit2pl$M2
p <- fit2pl$p

data <- list(dat, bias.a, rmse.a, bias.b, rmse.b, M2, p, k)
```

```
print(data)}

else if (M=="3PL")

{
dat <- as.data.frame(simdata(a = a, d = b, guess = c, N = N, itemtype = "dich", Theta = k))

model3pl <- mirt(data = dat, 1, itemtype = "3PL", SE = TRUE, verbose = FALSE, technical =
list(NCYCLES = 10000))

parameters <- as.data.frame(coef(model3pl, simplify = TRUE)$item)
bias.a <- mean(parameters[,1]-a)
bias.b <- mean(parameters[,2]-b)
bias.c <- mean(parameters[,3]-c)
rmse.a <- sqrt(mean((parameters[,1]-a)^2))
rmse.b <- sqrt(mean((parameters[,2]-b)^2))
rmse.c <- sqrt(mean((parameters[,3]-c)^2))

fit3pl <- M2(model3pl)
M2 <- fit3pl$M2
p <- fit3pl$p

data <- list(dat, bias.a, rmse.a, bias.b, rmse.b, bias.c, rmse.c, M2, p, k)
  print(data)}}
```

**The codes of Bulut and Sünbül (2017) were used in some parts of this function.*

International Journal of Assessment Tools in Education
2024, Vol. 11, No. 2, 249–269

https://doi.org/10.21449/ijate.1387041

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

Detection of differential item functioning with latent class analysis: PISA 2018 mathematical literacy test

Selim Daşçıoğlu[1*], **Tuncay Öğretmen**[1]

[1]Ege University, Faculty of Education, Department of Educational Sciences, İzmir, Türkiye

ARTICLE HISTORY

Received: Nov. 06, 2023

Accepted: Feb. 20, 2024

Keywords:

Cross country DIF,
Finite mixture models,
MIMIC,
Latent class analysis,
PISA.

Abstract: The purpose of this research is to determine whether PISA 2018 mathematical literacy test items show a differential item functioning across countries. For this purpose, only the items in booklet number three were examined using the MIMIC method with Latent Class Analysis (LCA) approach. PISA 2018 tests are mostly developed in English. Therefore, in DIF analyses, the reference group is the UK, while the focal groups consist of the other countries examined in the research (Türkiye, Finland, Japan, and the USA). According to the results, of the 23 test items, statistically significant DIF was observed in eight items in the UK-Türkiye sample, in seven items in the UK-Finland sample, in eleven items in the UK-Japan sample, and in three items in the UK-USA sample. It is seen that the effect and size of DIF in non-homogeneous groups differ between groups and these effects can be examined in more detail with the LCA method.

1. INTRODUCTION

The emerging technological developments and globalization offer countries the opportunity to develop their educational policies in a way that can help them keep up with the changing world and direct those changes. Large-scale international exams and practices also provide an opportunity for countries to measure their own levels and compare the results with those of other countries. One of these applications, the Program for International Student Assessment (PISA), is a program implemented by the Organization for Economic Co-operation and Development (OECD) and aims to measure the ability of 15-year-old students to utilize their reading comprehension, mathematics, scientific knowledge, and skills to cope with real-life problems. International monitoring research in education enables countries to assess their situation, compare their level with that of other countries, and make social and political decisions accordingly (MEB, 2019).

Considering that such decisions would be taken based on the measurement results, the quality of the measurement tools becomes important. One of the most important features of a measurement tool is its validity. In its broadest sense, validity is the degree to which measurement results serve the purpose (Nunnally & Bernstein, 1994). For this, all test items are

*CONTACT: Selim Daşçıoğlu ✉ selimdascioglu@gmail.com Ege University, Faculty of Education, Department of Educational Sciences, İzmir, Türkiye

e-ISSN: 2148-7456

expected to distinguish individuals well. Zumbo (1999) stated that the concept, method, and process of validation are at the core of measurement, and in the absence of validity studies, the inferences to be made from the measurement results will be meaningless.

Validity is not related to measurement results but to inferences made from measurement results (Zumbo, 1999). From this point of view, based on the results of international tests, it is necessary to emphasize the validity of making valid comparisons and inferences between countries.

Sometimes, the results obtained from the tests may vary according to the subgroups of the individuals. Differential item functioning (DIF) occurs when test takers from different subgroups show different success probabilities on the item after matching the basic ability that the item aims to measure (Camilli & Shepard 1994; Clauser & Mazor, 1998; Zumbo, 1999). Contrarily, item bias occurs when the probability of answering an item correctly differs for individuals at the same ability level but from different subgroups. This is due to a factor other than the characteristic the test item is intended to measure (Camilli & Shepard 1994; Clauser & Mazor, 1998; Zumbo, 1999). Accordingly, biased items show DIF. However, not every item showing DIF may be biased. Therefore, bias is a systematic error that affects the inferences made from the measurement results (Zumbo, 1999). In comparisons between subgroups such as gender or countries according to test results, it is important for test developers and policymakers to determine whether test items show DIF in terms of the relevant variable to make more valid comparisons and more unbiased measurements.

Additionally, there are more complex structural equation models that include many latent or observed variables and covariates and aim to determine the relationships between these variables. In such models, if DIF or direct effects arising from the covariate are predetermined and not included in the established model, biased results may occur (Vermunt, 2010).

Individuals can be divided into observable subgroups like gender, religion, language, race, and socioeconomic level. Additionally, individuals can be divided into subgroups that cannot be directly observed according to some latent traits like intelligence, achievement, attitude, alcohol addiction, etc., that we are trying to measure. LCA is a statistical method that allows the categorization of individuals into meaningful latent classes for the measured latent trait (Lanza & Collins, 2010; McCutcheon, 1987). DIF can occur between observed groups and latent classes. Especially in cases where the observed groups are not homogeneous, ignoring latent classes may lead to biased results and biased decisions (Sawatzky et al., 2018). Therefore, finite mixture models have been developed that allow the DIF to be among the latent classes and the observed groups.

Most of the tests in PISA 2018 were developed in English and French, and cross-cultural and cross-linguistic adaptations were made by relevant stakeholders (OECD, 2016c). However, no matter how meticulously the cross-cultural adaptation is carried out, mistakes can be made that will cause psychometric bias. Considering that these tests aim to measure latent structures, it is better to realize how difficult this task is. It is a process that does not expire and must be repeated at regular intervals (Messick, 1989). Therefore, conducting all validity studies, such as bias studies, during test development and adaptation processes and at the end of the actual application will contribute to future applications and processes.

One of the constructs that PISA aims to measure is mathematical literacy. OECD (2019) defined mathematical literacy as:

> Mathematical literacy is an individual's capacity to reason mathematically and to formulate, employ, and interpret mathematics to solve problems in a variety of real-world contexts. It includes concepts, procedures, facts and tools to describe, explain and predict phenomena. It assists individuals to know the role that mathematics plays in the world and to make the well-founded judgments and decisions needed by constructive, engaged and reflective 21st century citizens.

As it can be understood from the definition, the measurement of high-level mathematical skills related to real life and the development of skills based on the measurement results are of great importance in terms of raising individuals with these skills.

Considering all this information, it becomes necessary to conduct a DIF research to make meaningful comparisons between countries based on the results of PISA 2018 literacy tests. In addition to the country variable, which is the observed group variable while conducting this research, LCA will be used in order to consider the subgroups of individuals according to the latent feature of mathematical literacy.

In this study, the primary reason for using the MIMIC model with the LCA approach, as suggested by Masyn (2017), is to test whether the items show DIF according to the covariate, stepwise. In other words, the three-step procedure is used. With the addition of the covariate to the latent class model, the item response probability of individuals changes, and therefore, the latent class membership of some individuals may also change (Vermunt, 2010). However, this is undesirable in the current research. Because it is thought that the covariate (country variable in this study) is not a predictor of the latent class variable (mathematical literacy in this study). The three-step procedure will enable controlling this undesirable situation in the second step (analysis steps will be explained later), in which it is determined whether the items show DIF (Vermunt, 2010; Masyn, 2017). Furthermore, there is a limited number of studies to determine DIF with the latent class MIMIC method in the literature (Masyn, 2017; Tsaousis et al., 2020). It is thought that this study, which is based on real data, will contribute to the literature.

In this study, whether the mathematical literacy subtest items in PISA 2018, in which Türkiye and many OECD countries participated, show DIF across countries will be examined with the latent class MIMIC method. In this cross-country research, the other countries within the scope of the research will be compared in pairs with the UK, which is the reference group, given that the OECD is Europe-based and the languages in which the test was developed are English and French. While choosing other countries, attention was paid to the fact that these countries were from different parts of the world and from different cultures, and therefore Türkiye, Finland, Japan and the USA were determined.

Adapting a test that aims to measure a latent construct for other cultures is a very complicated and difficult process. This process aims to keep the validity and reliability of the measurements high with many quantitative and qualitative research techniques (Hambleton, Merenda & Spielberger, 2005). This is also true for international applications such as PISA, the Progress in International Reading Literacy Study (PIRLS), and the Trends in International Mathematics and Science Study (TIMSS). It is thought that DIF determination and bias studies will shed light on test development and adaptation studies and will contribute to increasing the validity of the decisions to be taken according to the test results.

2. METHOD

The target population of PISA is students between 15 years and three months and 16 years and two months who are in seventh grade and above, attending educational institutions located in the participating countries (OECD, 2016a). Approximately 600,000 students from 79 countries, 37 of which are OECD members, participated in PISA 2018 application. This sample represents the target population of approximately 32 million students (MEB, 2019).

Only the UK, Türkiye, Japan, Finland and the USA samples were analyzed in this study. Additionally, it was limited to booklet number three of the mathematical literacy test. Here is a country-wise bifurcation of 1442 participating students: UK 516, Türkiye 281, Japan 243, Finland 217, and USA 185.

Research data were obtained from the official website of OECD (https://www.oecd.org/pisa/data/), which prepared the PISA 2018 application. The application, scoring, and coding of the mathematical literacy test examined in the research were carried out by the relevant stakeholders (OECD, 2016b).

Only the mathematical literacy subtest of booklet number three of the PISA 2018 was examined in a recent study. There are 23 items in the test. Item 22, with a partially correct answer (0-1-2), was divided into two categories (0-1), with fully correct answers as "1" and other answers as "0". All other items are in two categories. Among the students who took the test, those who could not answer at least one of the test items (156 response patterns) because they could not see the test period or for any other reason were excluded from the data set. Apart from these, students who saw the question and left it blank were coded as "0", assuming that they did not answer because they did not know the correct answer.

Of the 1286 students remaining at the end of these procedures, 451 were from the UK, 253 from Türkiye, 188 from Finland, 226 from Japan, and 168 from the USA. The descriptive statistics of the test are given in the table below. Analyses were made with the TAP (Test Analysis Program).

Table 1. *Descriptive statistics of the PISA 2018 mathematical literacy test.*

Test Statistics		Test Statistics	
Number of Students	1286	Variance	22.26
Number of Items	23	Skewness	0.10
Lowest Score	0	Kurtosis	-0.67
Highest Score	23	Mean of Item Difficulty	0.47
Median	11	Mean of Item Discrimination	0.49
Mean	10.74	Mean of Item Point Biserial Discrimination	0.39
Standard deviation	4.72	KR-20	0.84

When the values in Table 1 are examined, it is seen that the group is heterogeneous, the distribution is slightly flat (-0.67), and the skewness (0.10) is close to zero. Based on this information, it can be assumed that the distribution is normal (Fraenkel et al., 2011). In addition, it can be said that the test has medium difficulty according to the mean item difficulty index (0.47), and the test distinguishes the upper group and the lower group from each other well according to the mean point double series discrimination values (0.39). Furthermore, according to the alpha coefficient (0.84), the reliability of the test in terms of internal consistency is high (Kerlinger, 1999).

Confirmatory factor analysis was performed with the R package lavaan for the model in which all items of the mathematical literacy test were collected in a single factor (mathematical literacy) structure and used diagonally weighted least squares (DWLS) estimation. Because chi-square (χ^2 = 385.615, *sd* = 230, *p*<0.001) was affected by the sample size and tended to be statistically significant, other goodness-of-fit values were examined. According to the analysis results, the RMSEA (0.023), CFI (0.990), TLI (0.989), GFI (0.982), and AGFI (0.979) values indicated a good model fit; the SRMR (0.060) value gave an acceptable model fit value. Therefore, it can be accepted that the test measures a single-factor construct (Harrington, 2009).

2.1. Latent Class Analysis

In 1950, Lazarsfeld performed a cluster analysis with data consisting of dichotomous items. In 1974, Goodman developed this analysis using the method of maximum likelihood estimation with categorical variables and made it applicable in practice (Magidson & Vermunt, 2004).

LCA is a statistical method for detecting and describing homogeneous and not directly observable (latent) subgroups in which individuals are separated according to a certain latent characteristic. This method comprises only one subgroup in which each individual is included. These subgroups of individuals cannot be known precisely due to measurement error. Additionally, the responses of individuals in each latent class to the indicator variables are independent of each other. This is called the local independence assumption, which is the only assumption of this model. LCA is used in a wide range of fields, such as behavioral sciences, medicine, education and social sciences, and economics (Magidson & Vermunt, 2004).

Iterative methods such as expectation-maximization or the Newton-Raphson algorithm are used in parameter estimation in LCA (Lanza & Collins, 2010; Magidson & Vermunt, 2004; McCutcheon, 1987).

We can divide the selection of the most suitable model in LCA into two: absolute model fit and comparative model fit (Lanza & Collins, 2009). If there are a certain number of latent classes expected for the latent class variable according to the theoretical background, absolute model fit can be used. In absolute model fit, the likelihood ratio chi-square goodness-of-fit statistic, the G2 test (shown as L2 in Latent Gold software) is used (Magidson & Vermunt, 2004).

H_0 tested here is, "There is no statistically significant difference between the selected model and the population distribution." In order for H_0 to be accepted and selected as the appropriate model, $p>\alpha$ is expected in the determined K-class model. Otherwise, if ($p<\alpha$), H_0 is rejected, and the K-class model determined according to this statistic cannot be used, or other model fit methods can be used (Lanza &Collins, 2010; McCutcheon, 1987). However, as the number of indicator variables in the model and the number of categories of these variables increase, and as gaps occur in the cells in the contingency table, sparseness will occur, and G^2 will tend to be higher (Lanza & Collins, 2010; McCutcheon, 1987). In this case, this method, which is desired to be used for model selection, can be misleading. In such a case, the use of comparative model selection may be healthier.

One of the statistics used in the comparative model selection is the G^2 difference (ΔG^2) statistic. In this method, the G^2 differences of two models with class K and class (K+1) are tested. However, we cannot directly test two models with different latent class numbers in this way because we cannot know the correct reference distribution. Therefore, the bootstrap method is used for both models, and then ΔG^2 is tested. Here, H_0 means that there is no significant difference between the K-class model and the (K+1) class model, therefore, if $p>\alpha$, the K-class model, with a lower number of parameters and a simpler one, is chosen based on the principle of parsimony. Otherwise, if ($p<\alpha$), it is seen that there is a significant difference in the (K+1) class model; the (K+1) class model is selected (Magidson & Vermunt, 2004).

Other statistics most frequently used in comparative model fit are information criteria. These are information criteria such as BIC (Bayesian Information Criterion), AIC (Akaike Information Criterion), and CAIC (Consistent Akaike Information Criterion). When comparing models with these information criteria, the model with the lower information criterion value is preferred to the model with the higher value (Lanza & Collins, 2009; McCutcheon, 1987; Magidson & Vermunt, 2004).

It should also be added that whether the appropriate model is chosen by one of the absolute or comparative model fit methods when the latent classes are examined (the responses of the individuals in the latent classes to the indicator variables and the predicted item-response probabilities), the classes should be well separated from each other and well defined (Lanza & Collins, 2010; McCutcheon, 1987; Vermunt & Magidson, 2002). If the latent classes in the selected model are not homogeneous enough and cannot be separated from each other in a meaningful way, i.e., they cannot be defined well, it will not make sense for the applied statistics to point to the selected model.

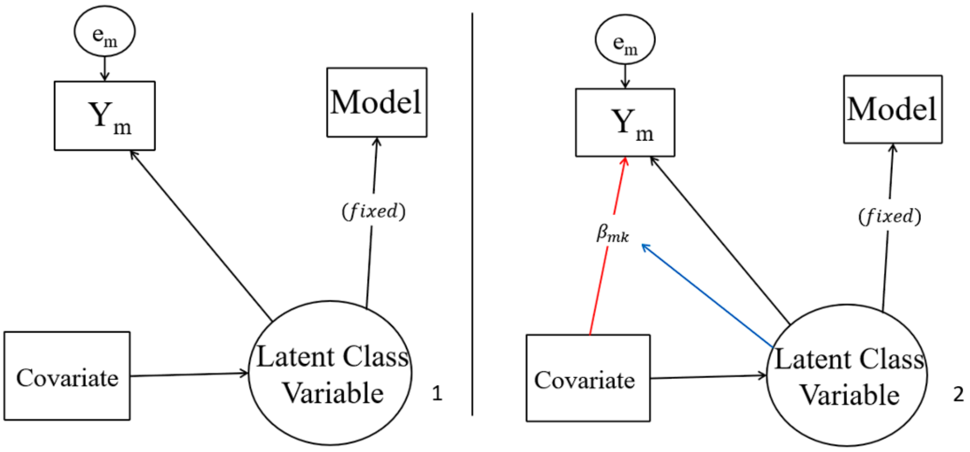
2.2. Latent Class MIMIC Model Steps

In this study, the steps of the latent class MIMIC method proposed by Masyn (2017) will be used. The steps of the analysis are as follows:

Initial Stage (Step 0): At this stage, LCA is performed with indicator variables (test items in this study) without a covariate, and the most suitable K-class model is determined. Individuals assigned to classes according to the selected K-class model are then numbered according to these classes. This is the first step of the three-step approach proposed by Vermunt (2010). The reason for class enumeration before including the covariate in the model is that when the covariate is included, the changes that may occur in the item response probabilities and latent class memberships of some individuals cannot be ignored (Nylund-Gibson et al., 2014).

Step 1: In this step, two different models are estimated. In the first model (M1.0), the group variable is included as a covariate in addition to the initial model (K-class). Here, the group variable has a direct effect only on the latent class variable. In other words, this model can be called the No-DIF model. In the second model (M1.1), the covariate included in the model has a direct effect on both items (non-uniform DIF where the effects of the covariate on the items are released to vary between classes). So, this model can also be called All-DIF. Then, the two models are compared using the likelihood ratio test (LRT). In this comparison, twice the difference of the loglikelihood (Δ-2LL) values of the models is tested with the chi-square test, which considers the difference in the number of parameters (ΔNpar) of the models as degrees of freedom. If H_0 cannot be statistically rejected (p>α), there is no evidence that the covariate is a source of DIF, and the analysis ends there. However, if H_0 is rejected (p≤α), there is sufficient evidence that the covariate can be a source of DIF for at least one of the indicator variables, and the second step is taken.

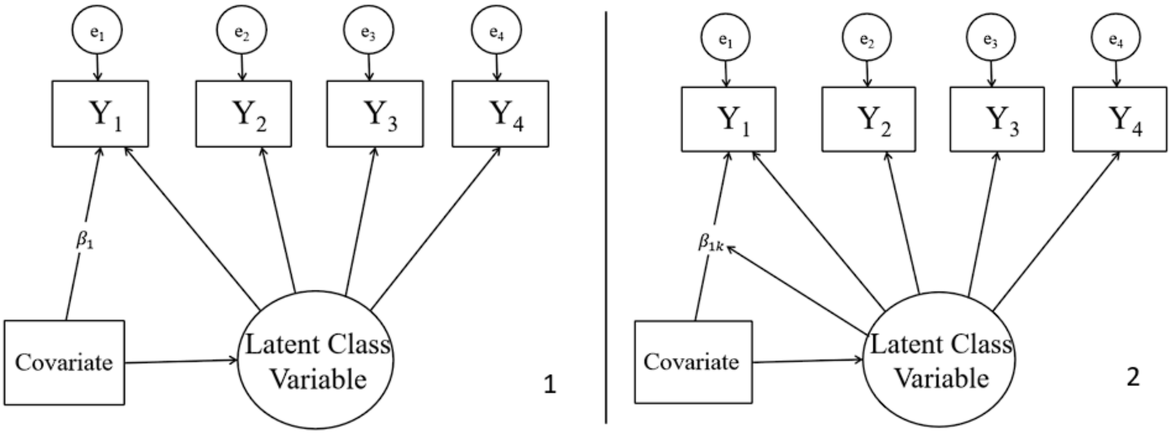
Figure 1. *In step 2, with a three-step approach, the M2.0.m model (1) in which the item Y_m covariate has no DIF effect, and the M2.1.m model (2) in which the covariate has a non-uniform DIF effect (β_{mk} is log odds ratio of endorsing item Y_m given membership latent class k for one-unit positive difference of covariate).*



Step 2: In this step, a non-uniform DIF test will be performed for each indicator variable (test item) separately. The three-step method is used for this (Magidson & Vermunt, 2004). In the first model, their membership in the initially obtained K-class model is fixed. The first item (m_1) and the covariate are included in the model. Here, the covariate has no direct effect on the item (Figure 1 left (1)), and this model is shown as M2.0.1. In the second model (M2.1.1), the covariate has a direct effect on the item, and the effects of the covariate on the items were left free to change between classes (Figure 1 right (2)). Then, the two models are compared with LRT. These model comparisons are made separately for each item (For example, M2.0.2 and M2.1.2 models for item 2). For items with statistically significant pairwise comparisons, there is sufficient evidence for DIF resulting from the covariate.

Step 3: In this step, a new latent class MIMIC model is estimated in line with the findings from the second step. In the model (M3.0), there is a non-uniform DIF effect for the items whose DIF was determined in step 2. For items for which no evidence of DIF can be obtained, the covariate has no direct effect. M3.0 is compared in pairs with M1.0 and M1.1. As a result of this comparison, it is expected that M3.0 is statistically better (p<α) than M1.0 and not worse than M1.1 (p>α).

Figure 2. *Models where (1) in the latent class model, the covariate is the source of uniform DIF for the indicator variable Y_1 and (2) the source of non-uniform DIF for the same variable.*

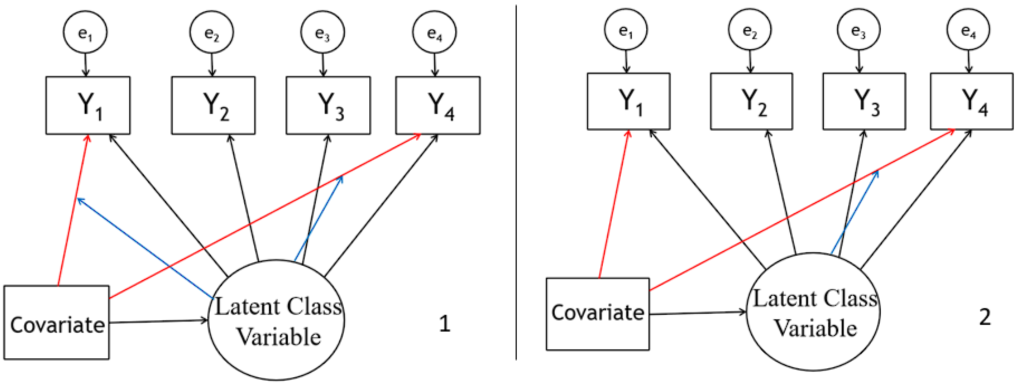


Step 4: In this step, a uniform DIF test will be performed for items showing DIF. In the estimated model (M4.1), unlike the M3.0 model, in one of the items showing non-uniform DIF, the variation of the common effect between the latent classes is fixed (Figure 2). Thus, only the covariate and the latent variable have a direct effect on that item. These models are set up separately for each item. In each model, the uniform DIF effect of only one item is tested (Figure 3). Each model is then individually compared (LRT) to the M3.0. If new models (such as M4.1 and M4.2) are not statistically worse than M3.0 (p>α), there is uniform evidence of DIF. Conversely, it can be said that there is evidence of non-uniform DIF.

Step 5: If there are items with uniform DIF detected in Step 4, a new model is estimated (M5.0) that these items show uniform DIF, unlike M3.0. Then, M3.0 LRT is compared with this model, and it is expected that M5.0 is not statistically worse (p> α).

Step 6: The direction and effect size of DIF in items with DIF will be determined. For this, the estimated coefficient (*β*) of the direct effect from the country variable to the item in items showing uniform DIF will be examined. As the value for the UK is coded as 0 and that for the other country is coded as 1, a positive coefficient will indicate DIF in favor of the other country (focal group), and a negative will indicate DIF in favor of the UK (reference group). For items with non-uniform DIF, the latent class or classes and the direction of the DIF will be measured. When evaluating the effect size, according to the ETS (Educational Testing Service) criteria, those equal to or less than 0.44 will be considered negligible (small) DIFs, those greater than 0.64 will be considered large DIFs, and those between these values will be considered medium-sized DIFs (Masyn, 2017; Tsaousis et al., 2020). Analyses were made with the Latent Gold 5.1 program (Vermunt & Magidson, 2016).

Figure 3. *In step 4 for Y_1; (1) non-uniform DIF effect for Y_1 and Y_4 (like M3.0), (2) uniform DIF for Y_1 and non-uniform DIF for Y_4 (like M4.1).*



3. FINDINGS

In the initial stage (Step 0), LCA was performed with mathematical literacy items in each sample. The following table shows the results of the LCA.

Table 2. *LCA results by samples.*

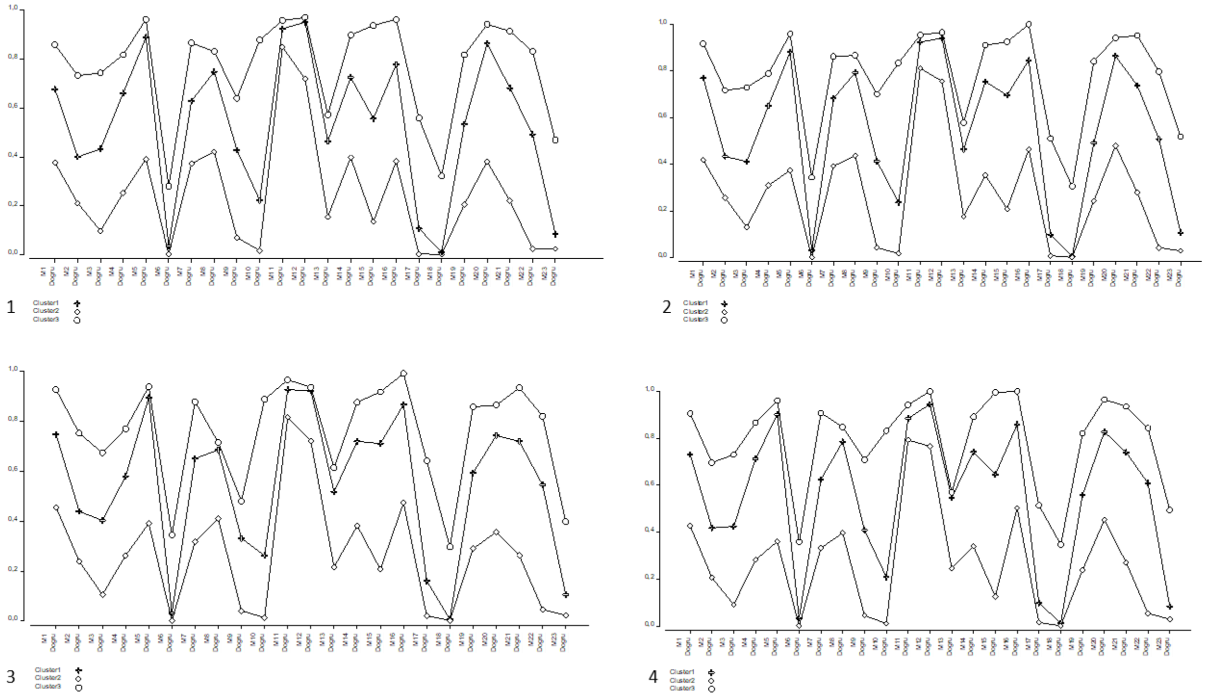
Sample	Model	LL	BIC(LL)	AIC(LL)	AIC3(LL)	CAIC(LL)	Npar	L^2	*df*	*p*
UK - Türkiye	1 Class	-9193.03	18536.86	18432.06	18455.06	18559.86	23	9207.16	681	0.00
	2 Class	-8164.76	16637.68	16423.51	16470.51	16684.68	47	7150.61	657	0.00
	3 Class	-7988.66	16442.85	16119.32	16190.32	16513.85	71	6798.43	633	0.00
	4 Class	-7920.36	16463.61	16030,72	16125.72	16558.61	95	6661.82	609	0.00
UK - Finland	1 Class	-8243.20	16634.97	16532.40	16555.40	16657.97	23	8282.64	616	0.00
	2 Class	-7415,20	15134.01	14924.40	14971.40	15181.01	47	6626.64	592	0.00
	3 Class	-7269.18	14997.00	14680.35	14751.35	15068,00	71	6334.59	568	0.00
	4 Class	-7199.01	15011.71	14588.02	14683.02	15106.71	95	6194.26	544	0.00
UK - Japan	1 Class	-8930.37	18010.64	17906.73	17929.73	18033.64	23	9070.12	654	0.00
	2 Class	-8040.66	16387.65	16175.32	16222.32	16434.65	47	7290.71	630	0.00
	3 Class	-7890.95	16244.66	15923,90	15994,90	16315.66	71	6991.29	606	0.00
	4 Class	-7833,30	16285.77	15856.59	15951.59	16380.77	95	6875.98	582	0.00
UK - USA	1 Class	-7968.92	16085.69	15983.84	16006.84	16108.69	23	8015.21	596	0.00
	2 Class	-7139.89	14581.91	14373.79	14420.79	14628.91	47	6357.16	572	0.00
	3 Class	-7005.27	14466.93	14152.53	14223.53	14537.93	71	6087,90	548	0.00
	4 Class	-6940.65	14491.98	14071.31	14166.31	14586.98	95	5958.68	524	0.00

Npar: number of parameters, df: degrees of freedom

Models with more than four latent classes are not given in Table 2 because they are not well defined. Table 2 shows that the p values of all latent class models are statistically significant. However, as stated under the heading "Parameter Estimation and Model Selection", because this figure tends to be statistically significant due to the number of variables and sparseness, other information criteria will be used in the model selection. Considering the models with the lowest information criterion values in all samples, the three-class model, according to BIC and CAIC; according to AIC and AIC3, the four-class model is more suitable for the data. Güngör Culha (2012) concluded in his research that “BIC and CAIC criteria give better results than other criteria in making the right decision while choosing the most suitable model as the sample grows.” Additionally, when the latent classes in three-class and four-class models are examined, it is seen that the classes are more homogeneous in the former model. Based on this information, it was concluded that the most suitable model for the data in all samples was the three-class model. Figure 4 shows the item-response probabilities of the latent classes in three-class models.

Figure 4 shows that the latent classes are separated from each other for all samples. The class with the highest probability of rendering a correct answer for all items was named as “High Achiever Class (HAC)”, the lowest class as “Low Achiever Class (LAC)” and the other class as “Moderate Achiever Class (MAC)”. The sizes of the latent classes are as follows: 15.4% of the UK–Türkiye sample is in HAC, 43.5% in MAC, and 41.1% in LAC. Of the UK–Finland sample, 14.7% are in HAC, 47.9% in MAC and 37.4% in LAC. Of the UK-Finland sample, 14.7% are in HAC, 47.9% in MAC, and 37.4% in LAC. Of the UK-Japan sample, 18.1% were in HAC, 46.1% in MAC, and 34.8% in LAC. In the UK-USA sample, 11.2% are in HAC, 49.6% in MAC and 39.2% in LAC. From here, we move on the next step of the analysis.

Figure 4. *Item-response probabilities of latent classes in the three-class model: (1) UK-Türkiye, (2) UK-Finland, (3) UK-Japan, (4) UK-USA.*



MIMIC analysis results of the UK-Türkiye, UK-Finland, UK-Japan, and UK-USA samples are given in Appendices (Table 5, 6, 7 and 8), respectively. In the first step, the M1.0 No-DIF model was compared with the M1.1 All-DIF model using LRT. When the tables were examined, a statistically significant difference was observed between the models in all samples (*Δ-2LL*=263.21, *ΔNpar*=69, *p*<0.001 for the UK-Türkiye; *Δ-2LL*=246.06, *ΔNpar* = 69, *p*<0.001 for the UK-Finland; *Δ-2LL*=428.34, *ΔNpar*=69, *p*<0.001 for the UK-Japan, and *Δ-2LL*=108.35, *ΔNpar*=69, *p*<0.001 for the UK-USA). This is sufficient proof that the country variable is a source of DIF for at least one of the indicator variables in at least one of the latent classes. From this point of view, the second step was started.

In step 2, the no DIF model (M2.0.m) established for each item and the non-uniform DIF model (M2.1.m) were compared with the LRT. In comparisons with statistically significant difference between them, it was concluded that the relevant item contained DIF originating from the country variable. According to the results in the tables in Appendices: In the UK-Türkiye sample, in items 5, 7, 11, 13, 14, 15, 16, 20, 21 and 22; in the UK-Finland sample, items 1, 7, 8, 11, 13, 15, 19 and 22; DIF originating from the country variable was found in items 1, 4, 8, 9, 11, 12, 13, 15, 17, 20 and 23 in the UK-Japan sample and in items 10, 16 and 23 in the UK-USA sample.

In step 3, a new model was estimated (M3.0), in which there was a non-uniform effect of DIF on the items in which DIF was detected in the previous step, and there was no direct effect on the other items from the country variable (M3.0), and this model was compared with M1.0 and M1.1. As expected in the UK-Finland and the UK-USA samples, M3.0 was statistically better (*p*<0.05) than M1.0 and not statistically worse than M1.1 (*p*>0.05). However, there was a statistically significant difference between the M3.0 and M1.0 and M1.1 models in the UK-Türkiye and the UK-Japan samples. Then, the BIC values of the models were examined. In both samples, the BIC of M3.0 was considerably lower than the BIC of M1.1 (in the UK-Türkiye, BIC= 16442.78 in M3.0, BIC=16627.33 in M1.1; in the UK-Japan, BIC= 16083.68 at M3.0, BIC=16259.45 at M1.1). Based on this information, it was decided that the most appropriate latent class MIMIC model up to this stage was M3.0 in all samples (Masyn, 2017; Tsaousis et al., 2020).

In step 4, the DIF type of the items for which DIF was detected in previous steps will be determined. For this, the variation of the direct effect between latent classes in one of these items at a time was consistent across classes (uniform DIF model for the item). The estimated models were compared with the M3.0. A statistically significant difference was accepted as evidence that the relevant item contained non-uniform DIF, and otherwise, it contained uniform DIF. Accordingly, in the UK-Türkiye sample, uniform DIF caused by the country variable was found in items 5, 7, 11, 14, 21, and 22, and non-uniform DIF in items 13, 15, 16, and 20. In the UK-Finland sample, items 1, 7, 8, 19, and 22 are uniform caused by the country variable, non-uniform in items 11, 13, and 15; In UK-Japan sample, items 1, 4, 9, 11, 13, 17, 20, and 23 are uniform caused by the country variable, and non-uniform for items 8, 12, and 15; In the UK-USA sample, uniform DIF was detected in items 10 and 16 caused by the country variable, and non-uniform DIF in item 23.

In step 5, a new latent class MIMIC model (M5.0) was estimated with the information obtained in the previous step, in which the items showing DIF had a direct effect from the country variable according to the type of DIF and the other items had no direct effect from the country variable. M5.0 and M3.0 were compared with LRT, and there was no statistically significant difference between models in all samples. In other words, M5.0 can be considered the most suitable model for all samples.

In step 6, the direction and magnitude of the DIF effects arising from the country variable in the items were examined. The results are shown in Table 3. According to the results in Table 3, a statistically significant, uniform, and negligible DIF effect caused by the country variable was observed in items 7, 11, 14, 21, and 22 in the UK-Türkiye sample. The DIF effect in items 7, 11, and 14 is in favor of Türkiye, but the DIF effect in items 21 and 22 is in favor of the UK. When the items showing non-uniform DIF caused by the country variable were examined, it was observed that for item 13, DIF was small in LAC and medium in MAC, a statistically significant, and DIF in favor of the UK. In items 15 and 16, the DIF effect, which is statistically significant only in MAC, is in favor of the UK and of negligible magnitude. In item 20, the DIF effect, which is statistically significant only in LAC, is in favor of the UK and is of negligible magnitude.

In the UK-Finland sample, items 1, 7, 8, 19, and 22 showed a statistically significant, uniform, and negligible DIF effect caused by the country variable. While the DIF effect in items 1, 7, and 8 is in favor of Türkiye, the DIF effect in items 19 and 22 is in favor of the UK. In item 13, the statistically significant non-uniform DIF effect caused by the country variable in favor of the UK is moderate in LAC and negligible in MAC. The statistically significant DIF effect in item 15 is in favor of Finland and negligible in LAC and MAC.

In the UK-Japan sample, items 1, 4, 9, 11, 13, 17, 20, and 23 showed statistically significant uniform DIF caused by the country variable. While the effect size of DIF in items 9 and 20 was medium, it was observed that DIF was negligible in other items. In addition, while the DIF effect in items 1, 11, and 17 is in favor of Japan, it is in favor of the UK in items 4, 9, 13, 20, and 23. The non-uniform DIF effect, which is statistically significant in item 8, is in favor of the UK and negligible in MAC and HAC. In item 12, the DIF effect, which is statistically significant only in MAC, is in favor of the UK and is negligible. In item 15, the statistically significant DIF effect is in favor of Japan and negligible in LAC and MAC. Another issue seen in Table 3 is that although item 5 in the UK-Türkiye sample and item 11 in the UK-Finland sample showed DIF in the previous steps, the DIF effects are not statistically significant in M5.0.

Table 3. *DIF effects from country variable in M5.0.*

		C1 (LAC)			C2 (MAC)			C3 (HAC)		
Samples	Item	β	SE	*p*	β	SE	*p*	β	SE	*p*
UK - Türkiye	5	0.11	0.06	0.06	0.11	0.06	0.06	0.11	0.06	0.06
	7	0.09	0.04	0.04	0.09	0.04	0.04	0.09	0.04	0.04
	11	0.38	0.09	0.00	0.38	0.09	0.00	0.38	0.09	0.00
	13	-0.28	0.11	0.01	-0.46	0.09	0.00	-0.04	0.10	0.70
	14	0.11	0.05	0.02	0.11	0.05	0.02	0.11	0.05	0.02
	15	-0.02	0.10	0.85	-0.32	0.07	0.00	-0.31	0.25	0.22
	16	-0.10	0.07	0.15	-0.33	0.08	0.00	-1.16	1.03	0.26
	20	-0.25	0.08	0.00	0.07	0.11	0.53	0.03	0.26	0.92
	21	-0.25	0.05	0.00	-0.25	0.05	0.00	-0.25	0.05	0.00
	22	-0.28	0.06	0.00	-0.28	0.06	0.00	-0.28	0.06	0.00
UK - Finland	1	0.19	0.06	0.00	0.19	0.06	0.00	0.19	0.06	0.00
	7	0.26	0.05	0.00	0.26	0.05	0.00	0.26	0.05	0.00
	8	0.17	0.05	0.00	0.17	0.05	0.00	0.17	0.05	0.00
	11	0.16	0.10	0.13	1.49	1.83	0.42	1.08	1.84	0.56
	13	-0.53	0.20	0.01	-0.41	0.08	0.00	0.07	0.14	0.60
	15	0.31	0.09	0.00	0.20	0.09	0.02	-0.94	0.93	0.31
	19	-0.16	0.05	0.00	-0.16	0.05	0.00	-0.16	0.05	0.00
	22	-0.22	0.06	0.00	-0.22	0.06	0.00	-0.22	0.06	0.00
UK - Japan	1	0.18	0.05	0.00	0.18	0.05	0.00	0.18	0.05	0.00
	4	-0.28	0.05	0.00	-0.28	0.05	0.00	-0.28	0.05	0.00
	8	0.10	0.08	0.19	-0.33	0.07	0.00	-0.33	0.13	0.01
	9	-0.52	0.07	0.00	-0.52	0.07	0.00	-0.52	0.07	0.00
	11	0.31	0.09	0.00	0.31	0.09	0.00	0.31	0.09	0.00
	12	-0.11	0.08	0.17	-0.37	0.11	0.00	-1.24	1.60	0.44
	13	-0.10	0.05	0.02	-0.10	0.05	0.02	-0.10	0.05	0.02
	15	0.23	0.10	0.02	0.30	0.10	0.00	-1.09	0.94	0.25
	17	0.34	0.06	0.00	0.34	0.06	0.00	0.34	0.06	0.00
	20	-0.49	0.06	0.00	-0.49	0.06	0.00	-0.49	0.06	0.00
	23	-0.15	0.07	0.04	-0.15	0.07	0.04	-0.15	0.07	0.04
UK - USA	10	-0.32	0.09	0.00	-0.32	0.09	0.00	-0.32	0.09	0.00
	16	0.19	0.06	0.00	0.19	0.06	0.00	0.19	0.06	0.00
	23	0.33	0.21	0.11	-0.31	0.17	0.07	-0.47	0.21	0.02

SE: standard error

In the UK-USA sample, on the other hand, a statistically significant and uniform DIF with a small effect size was detected in the direction of the UK in item 10 and in the direction of the USA in item 16. In item 23, however, the non-uniform DIF effect, which is statistically significant only in HAC, is moderately large in the direction of the UK.

In Table 4, DIF effects with an effect size below 0.45 are shown as A, above 0.64 are shown as C, and between these two values are shown as B. Additionally, DIF effects in favor of the UK (reference group) are shown with “-” and in favor of the other country (focal group) “+”.

Table 4. *Direction and magnitude of DIF effects.*

UK-Türkiye				UK-Finland				UK-Japan				UK-USA			
Item	LAC	MAC	HAC	Item	LAC	MAC	HAC	Item	LAC	MAC	HAC	Item	LAC	MAC	HAC
5	A+*	A+*	A+*	1	A+	A+	A+	1	A+	A+	A+	10	A-	A-	A-
7	A+	A+	A+	7	A+	A+	A+	4	A-	A-	A-	16	A+	A+	A+
11	A+	A+	A+	8	A+	A+	A+	8	A+*	A-	A-	23	A+*	A-*	B-
13	A-	B-	A-*	11	A+*	C+*	C+*	9	B-	B-	B-				
14	A+	A+	A+	13	B-	A-	A+*	11	A+	A+	A+				
15	A-*	A-	A-*	15	A+	A+	C-*	12	A-*	A-	C-*				
16	A-*	A-	C-*	19	A-	A-	A-	13	A-	A-	A-				
20	A-	A+*	A+*	22	A-	A-	A-	15	A+	A+	C-*				
21	A-	A-	A-					17	A+	A+	A+				
22	A-	A-	A-					20	B-	B-	B-				
								23	A-	A-	A-				

*p>0.05

According to this information, the uniform DIF coefficients can be interpreted as follows. In all latent classes, the probability of answering item 11 correctly for students in the Türkiye sample is approximately 1.46 times that of students in the UK sample ($e^{0.38} = 1.46$). The probability of students in the Finland sample answering item 7 correctly is approximately 1.30 times the probability of answering correctly for students in the UK sample ($e^{0.26} = 1.30$). The probability of students in the UK sample answering item 9 correctly is approximately 1.68 times that of students in the Japan sample ($e^{0.52} = 1.68$). The probability of students in the UK sample answering item 10 correctly is approximately 1.38 times the probability of answering item 10 correctly than the students in the US sample ($e^{0.32} = 1.38$).

However, according to non-uniform DIF coefficients, the probability of answering item 13 correctly for students in the UK sample in MAC is approximately 1.58 times the probability of answering correctly for students in the Türkiye sample ($e^{0.46} = 1.58$). In LAC, the probability of students in the UK sample answering item 13 correctly is approximately 1.70 times the probability of answering correctly for students in the Finland sample ($e^{0.53} = 1.70$). In MAC, the probability of students in the Japan sample answering item 15 correctly is approximately 1.35 times the probability of answering item 15 correctly than the students in the UK sample ($e^{0.30} = 1.35$). In HAC, the probability of students in the UK sample answering item 23 correctly is approximately 1.60 times the probability of answering item 23 correctly than the students in the US sample ($e^{0.47} = 1.60$). Other DIF effects can be interpreted similarly.

4. DISCUSSION and CONCLUSION

In this study, whether the PISA 2018 application mathematical literacy test items in booklet number three show DIF across countries was examined with the latent class MIMIC approach. The UK was chosen as the reference group, and Türkiye, Finland, Japan and the USA as the focal group.

Considering the number of items with DIF detected according to the country variable in the paired comparisons examined, it was seen that fewer items showed DIF in the UK-USA sample (three items) compared to other samples (UK-Türkiye nine items, UK-Finland seven items, UK-Japan 11 items). There is one item with a statistically significant B level DIF in the UK-Türkiye sample, one in the UK-Finland sample, two in the UK-Japan sample, and one item in the UK-USA sample. No statistically significant C level DIF effect was observed in any of the samples caused by the country variable. The fact that the number of items with DIF observed in the UK-USA sample and their effect sizes are considerably less than in other samples strengthen the opinion that the source of DIF in other samples is significantly related to test language. In addition, more DIF items were observed in the UK-Japan sample in terms of number and effect size compared to other samples. This again showed the importance of translation between languages and differences between cultures in adaptation studies. However, the items should be analyzed qualitatively to determine whether the DIF in the related items is due to the real difference between the groups or bias.

In the four sample examinations, different items showed DIF in different samples. However, it was also observed that some items showed DIF in the same direction in both samples. For example, in both the UK-Türkiye and the UK-Finland samples, item 7 showed DIF at level A in favor of the focal group, and item 22 showed DIF at level A in favor of the reference group. A similar situation can be said for items 1 and 15 in the UK-Finland and the UK-Japan samples. In addition, only item 13 showed DIF in favor of the reference group in the other three samples except the UK-USA, but with different effect sizes in different latent classes. Item 13 shows DIF at level A for all latent classes in the UK-Japan sample. But in the UK-Türkiye sample, level A DIF for LAC and level B DIF for MAC; and in the UK-Finland sample level B DIF for LAC and level A DIF for MAC was observed. Similarly, Saatçioğlu (2022) examined the DIF of PISA 2018 financial literacy items resulting from the gender variable using the latent class MIMIC method. As a result, it was determined that the DIF effect differed (non-uniformly) in latent classes in 5 out of 16 test items.

As mentioned before and as seen in this study, the LCA approach allows the examination of test items in terms of DIF not only according to the observed variables but also for the latent classes. As Zumbo et al. (2015) and Elkonca (2020) stated, it is thought that this will enable the DIF sources to be determined in more detail and accurately. However, it is seen that the effect and size of DIF in non-homogeneous groups differ between groups, and these effects can be examined in more detail with the LCA method. This is in line with the results of Oliveri et al. (2016), Sawatzky et al. (2018), and Uyar (2020).

4.1. Suggestions

1. At the end of the analysis, it was seen that some of the items whose DIF effect was detected in the second and fourth steps were not statistically significant in the final model (M5.0) (item 5 in the UK-Türkiye sample and item 11 in the UK-Finland sample). In future studies, as Masyn (2017) and Tsaousis et al. (2020) suggested, sequential procedures according to p values in DIF determination steps or simultaneous procedures in terms of DIF type can be tried in terms of reviewing and improving the latent class MIMIC procedures used in this research, and simulation and real data studies can be done to investigate Type I and Type II errors.

2. Different DIF determination methods can be compared with the method used in the research and the conditions under which the methods are strong or weak relative to each other can be investigated.

3. In this research, we examined only mathematical literacy test in PISA 2018 and only booklet number three. Other tests or booklets in the application can be examined in terms of different observed variables (such as gender, region of residence of the student, and socioeconomic structure).

4. Test developers should better consider the characteristics of countries, such as curriculum, language, and culture, in both test development and adaptation studies and should do their part more carefully to avoid situations that may cause bias.

Acknowledgments

This study has been produced from Master's Thesis that was conducted under the supervision of the Prof. Dr. Tuncay Öğretmen and prepared by Selim Daşçıoğlu.

Declaration of Conflicting Interests and Ethics

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

Contribution of Authors

Selim Daşçıoğlu: Investigation, Visualization, Conception, Methodology, Analysis, and Writing-original draft. **Tuncay Öğretmen:** Conception, Supervision, Software, Critical Review, and Validation.

Orcid

Selim Daşçıoğlu https://orcid.org/0000-0001-6820-4585
Tuncay Öğretmen https://orcid.org/0000-0001-7783-1409

REFERENCES

Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items*. SAGE Publications.

Clauser, B.E., & Mazor, K.M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31 - 44.

Elkonca, F. (2020). *ABİDE öz yeterlilik ölçeği DMF kaynaklarının gizil sınıf yaklaşımıyla incelenmesi [An analysis of the DIF sources of ABİDE self-efficacy scale by means of a latent class approach]* [Unpublished doctoral dissertation]. Gazi University.

Fraenkel, J.R., Wallen, N.E., & Hyun, H.H. (2011). *How to design and evaluate research in education*. McGraw-Hill Education.

Güngör Culha, D. (2012). *Örtük sınıf analizlerinde ölçme eşdeğerliğinin incelenmesi [Investigating measurement equivalence with latent class analysis]* [Unpublished doctoral dissertation]. Ege University.

Hambleton, R.K., Merenda, P.F., & Spielberger, C.D. (Eds.). (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Lawrence Erlbaum Associates.

Harrington, D. (2009). *Confirmatory factor analysis*. Oxford University Press.

Kerlinger, F.N. (1999). *Foundations of behavioral research*. Wadsworth Publishing.

Lanza, S., & Collins, L. (2009). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. John Wiley & Sons, Inc.

Magidson, J., & Vermunt, J.K. (2004). Latent class models. D. Kaplan (Eds.), *The sage handbook of quantitative methodology for the social sciences* (s. 175-198). Sage Publications.

Masyn, K.E. (2017). Measurement invariance and differential item functioning in latent class analysis with stepwise multiple indicator multiple cause modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(2) 180-197.

McCutcheon, A.L. (1987). *Latent class analysis.* Sage Publication.

MEB (2019). *PISA 2018 Türkiye ön raporu [PISA 2018 Results]*. T.C. Milli Eğitim Bakanlığı.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Research Article, 18*(2), 5-11.

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory third edition*. McGraw-Hill.

Nylund-Gibson, K., Grimm, R., Quirk, M., & Furlong, M. (2014). A latent transition mixture model using the three-step specification. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(3), 439-454.

OECD. (2016a). *Sampling in PISA*. OECD Publishing.

OECD. (2016b). *PISA 2018 technical report*. OECD Publishing.

OECD. (2016c). *PISA 2018 translation and adaptation guidelines*. OECD Publishing.

OECD. (2019). PISA 2018 mathematics framework. *PISA 2018 assessment and analytical framework* (s. 73-95). OECD Publishing. https://doi.org/10.1787/13c8a22c-en

Oliveri, M.E., Ercikan, K., Lyons-Thomas, J., & Holtzman, S. (2016). Analyzing fairness among linguistic minority populations using a latent class differential item functioning approach. *Applied Measurement in Education, 29*(1), 17-29. https://doi.org/10.1080/08957347.2015.1102913

Saatçioğlu, F.M. (2022). Differential item functioning across gender with MIMIC modeling: PISA 2018 financial literacy items. *International Journal of Assessment Tools in Education*, *9*(3), 631-653. https://doi.org/10.21449/ijate.1076464

Sawatzky, R., Russell, L.B., Sajobi, T.T., Lix, L.M., Kopec, J., & Zumbo, B.D. (2018). The use of latent mixture models to identify Invariant Items in test construction. *Qual Life Res, 27*(7), 1745-1755. https://doi.org/10.1007/s11136-017-1680-8

Tsaousis, I., Sideridis, G.D., & AlGhamdi, H.M. (2020). Measurement invariance and differential item functioning across gender within a latent class analysis framework: Evidence from a high-stakes test for university admission in Saudi Arabia. *Frontiers in Psychology*, *11*(622). https://doi.org/10.3389/fpsyg.2020.00622

Uyar, Ş. (2020). Latent class approach to detect differential item functioning: PISA 2015. *Eurasian Journal of Educational Research*, *20*(88), 179-198. https://doi.org/10.14689

Vermunt, J.K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*(4), 450-469.

Zumbo, B.D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): logistic regression modeling as a unitary framework for binary and Likert-Type (Ordinal) item scores*. ON: Directorate of Human Resources Research and Evaluation.

Zumbo, B.D., Liu, Y., Wu, A.D., Shear, B.R., Olvera Astivia, O.L., & Ark, T.K. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly*, *12*(1), 136-151. https://doi.org/10.1080/15434303.2014.972559

APPENDIX

Table 5. *Latent class MIMIC analysis results in the UK-Türkiye sample.*

Step	Model	LL	Npar	Comparison	Δ-2LL	ΔNpar	*p*
1	M1.0	-7979.74	73	M1.0 - M1.1	263.21	69	0.00
	M1.1	-7848.13	142				
2	M2.0.1	-1844,19	7	M2.0.1 - M2.1.1	3.68	3	0.30
	M2.1.1	-1842.35	10				
	M2.0.2	-1832.64	7	M2.0.2 - M2.1.2	1.73	3	0.63
	M2.1.2	-1831.77	10				
	M2.0.3	-1779.28	7	M2.0.3 - M2.1.3	6.08	3	0.11
	M2.1.3	-1776.24	10				
	M2.0.4	-1825.93	7	M2.0.4 - M2.1.4	1.57	3	0.67
	M2.1.4	-1825,14	10				
	M2.0.5	-1737.37	7	M2.0.5 - M2.1.5	12.30	3	0.01
	M2.1.5	-1731.23	10				
	M2.0.6	-1526.47	7	M2.0.6 - M2.1.6	0.90	3	0.83
	M2.1.6	-1526.01	10				
	M2.0.7	-1851.96	7	M2.0.7 - M2.1.7	8.44	3	0.04
	M2.1.7	-1847.74	10				
	M2.0.8	-1835.26	7	M2.0.8 - M2.1.8	3.90	3	0.27
	M2.1.8	-1833.31	10				
	M2.0.9	-1769.20	7	M2.0.9 - M2.1.9	7.42	3	0.06
	M2.1.9	-1765.49	10				
	M2.0.10	-1640.81	7	M2.0.10 - M2.1.10	1.00	3	0.80
	M2.1.10	-1640.31	10				
	M2.0.11	-1645.25	7	M2.0.11 - M2.1.11	27.90	3	0.00
	M2.1.11	-1631.30	10				
	M2.0.12	-1666.29	7	M2.0.12 - M2.1.12	1.46	3	0.69
	M2.1.12	-1665.56	10				
	M2.0.13	-1822.90	7	M2.0.13 - M2.1.13	35.71	3	0.00
	M2.1.13	-1805.04	10				
	M2.0.14	-1827,19	7	M2.0.14 - M2.1.14	11.81	3	0.01
	M2.1.14	-1821.29	10				
	M2.0.15	-1765.14	7	M2.0.15 - M2.1.15	13.38	3	0.00
	M2.1.15	-1758.45	10				
	M2.0.16	-1786.84	7	M2.0.16 - M2.1.16	16.09	3	0.00
	M2.1.16	-1778.79	10				
	M2.0.17	-1598.18	7	M2.0.17 - M2.1.17	6.03	3	0.11
	M2.1.17	-1595.17	10				
	M2.0.18	-1501.41	7	M2.0.18 - M2.1.18	1.68	3	0.64
	M2.1.18	-1500.57	10				
	M2.0.19	-1825.37	7	M2.0.19 - M2.1.19	3.01	3	0.39
	M2.1.19	-1823.87	10				
	M2.0.20	-1754.93	7	M2.0.20 - M2.1.20	8.37	3	0.04

Table 5. *(Continued)*

Step	Model	LL	Npar	Comparison	Δ-2LL	ΔNpar	*p*
2	M2.1.20	-1750.75	10				
	M2.0.21	-1789.39	7	M2.0.21 - M2.1.21	13.62	3	0.00
	M2.1.21	-1782.58	10				
	M2.0.22	-1707.47	7	M2.0.22 - M2.1.22	11.99	3	0.01
	M2.1.22	-1701.47	10				
	M2.0.23	-1610.24	7	M2.0.23 - M2.1.23	3.65	3	0.30
	M2.1.23	-1608.42	10				
3	M3.0	-7883.71	103	M1.0 - M3.0	192.05	30	0.00
				M3.0 - M1.1	71.16	39	0.00
4	M4.1	-7883.94	101	M4.1 - M3.0	0.44	2	0.80
	M4.2	-7883.73	101	M4.2 - M3.0	0.04	2	0.98
	M4.3	-7884.49	101	M4.3 - M3.0	1.55	2	0.46
	M4.4	-7888.26	101	M4.4 - M3.0	9.10	2	0.01
	M4.5	-7883.72	101	M4.5 - M3.0	0.02	2	0.99
	M4.6	-7886.75	101	M4.6 - M3.0	6.08	2	0.05
	M4.7	-7888.33	101	M4.7 - M3.0	9.22	2	0.01
	M4.8	-7887.52	101	M4.8 - M3.0	7.61	2	0.02
	M4.9	-7884.43	101	M4.9 - M3.0	1.44	2	0.49
	M4.10	-7885.33	101	M4.10 - M3.0	3.24	2	0.20
5	M5.0	-7887,18	91	M5.0 - M3.0	6.93	12	0.86

Table 6. *Latent class MIMIC analysis results in the UK-Finland sample.*

Step	Model	LL	Npar	Comparison	Δ-2LL	ΔNpar	*p*
1	M1.0	-7268.68	73	M1.0 - M1.1	246.06	69	0.00
	M1.1	-7145.65	142				
2	M2.0.1	-1633.39	7	M2.0.1 - M2.1.1	14.16	3	0.00
	M2.1.1	-1626.31	10				
	M2.0.2	-1680.58	7	M2.0.2 - M2.1.2	5.44	3	0.14
	M2.1.2	-1677.86	10				
	M2.0.3	-1634.01	7	M2.0.3 - M2.1.3	1.15	3	0.76
	M2.1.3	-1633.43	10				
	M2.0.4	-1674.09	7	M2.0.4 - M2.1.4	5.69	3	0.13
	M2.1.4	-1671.25	10				
	M2.0.5	-1566.33	7	M2.0.5 - M2.1.5	1.41	3	0.70
	M2.1.5	-1565.62	10				
	M2.0.6	-1383,12	7	M2.0.6 - M2.1.6	0.32	3	0.96
	M2.1.6	-1382.95	10				
	M2.0.7	-1667.79	7	M2.0.7 - M2.1.7	26.40	3	0.00
	M2.1.7	-1654.59	10				
	M2.0.8	-1635.72	7	M2.0.8 - M2.1.8	9.04	3	0.03
	M2.1.8	-1631.20	10				
	M2.0.9	-1585.90	7	M2.0.9 - M2.1.9	2.49	3	0.48

Table 6. *(Continued)*

Step	Model	LL	Npar	Comparison	Δ-2LL	ΔNpar	*p*
	M2.1.9	-1584.66	10				
	M2.0.10	-1507.85	7	M2.0.10 - M2.1.10	0.20	3	0.98
	M2.1.10	-1507.75	10				
	M2.0.11	-1495.75	7	M2.0.11 - M2.1.11	16.21	3	0.00
	M2.1.11	-1487.65	10				
	M2.0.12	-1498.40	7	M2.0.12 - M2.1.12	5.03	3	0.17
	M2.1.12	-1495.88	10				
	M2.0.13	-1665.78	7	M2.0.13 - M2.1.13	50.38	3	0.00
	M2.1.13	-1640.58	10				
	M2.0.14	-1634.23	7	M2.0.14 - M2.1.14	7.39	3	0.06
	M2.1.14	-1630.53	10				
	M2.0.15	-1613.60	7	M2.0.15 - M2.1.15	15.83	3	0.00
	M2.1.15	-1605.68	10				
	M2.0.16	-1579.34	7	M2.0.16 - M2.1.16	1.23	3	0.75
	M2.1.16	-1578.72	10				
	M2.0.17	-1454.75	7	M2.0.17 - M2.1.17	0.54	3	0.91
	M2.1.17	-1454.48	10				
	M2.0.18	-1349.08	7	M2.0.18 - M2.1.18	0.01	3	1.00
	M2.1.18	-1349.07	10				
2	M2.0.19	-1665.61	7	M2.0.19 - M2.1.19	14.83	3	0.00
	M2.1.19	-1658.20	10				
	M2.0.20	-1588.13	7	M2.0.20 - M2.1.20	0.54	3	0.91
	M2.1.20	-1587.86	10				
	M2.0.21	-1615.46	7	M2.0.21 - M2.1.21	2.09	3	0.55
	M2.1.21	-1614.41	10				
	M2.0.22	-1581.46	7	M2.0.22 - M2.1.22	19.09	3	0.00
	M2.1.22	-1571.91	10				
	M2.0.23	-1479.37	7	M2.0.23 - M2.1.23	2.87	3	0.41
	M2.1.23	-1477.93	10				
3	M3.0	-7175.34	97	M1.0 - M3.0	186.68	24	0.00
				M3.0 - M1.1	59.37	45	0.07
4	M4.1	-7177.62	95	M4.1 - M3.0	4.58	2	0.10
	M4.2	-7176.73	95	M4.2 - M3.0	2.79	2	0.25
	M4.3	-7175.50	95	M4.3 - M3.0	0.32	2	0.85
	M4.4	-7178.88	95	M4.4 - M3.0	7.10	2	0.03
	M4.5	-7179.90	95	M4.5 - M3.0	9.14	2	0.01
	M4.6	-7179.88	95	M4.6 - M3.0	9.09	2	0.01
	M4.7	-7177.54	95	M4.7 - M3.0	4.40	2	0.11
	M4.8	-7176.95	95	M4.8 - M3.0	3.24	2	0.20
5	M5.0	-7182.77	87	M5.0 - M3.0	14.86	10	0.14

Table 7. *Latent class MIMIC analysis results in the UK-Japan sample.*

Step	Model	LL	Npar	Comparison	Δ-2LL	ΔNpar	*p*
1	M1.0	-7881.14	73	M1.0 - M1.1	428.34	69	0.00
	M1.1	-7666.97	142				
2	M2.0.1	-1763.18	7	M2.0.1 - M2.1.1	13.77	3	0.00
	M2.1.1	-1756.29	10				
	M2.0.2	-1801.82	7	M2.0.2 - M2.1.2	3.88	3	0.27
	M2.1.2	-1799.88	10				
	M2.0.3	-1757.99	7	M2.0.3 - M2.1.3	2.69	3	0.44
	M2.1.3	-1756.65	10				
	M2.0.4	-1807.98	7	M2.0.4 - M2.1.4	38.34	3	0.00
	M2.1.4	-1788.81	10				
	M2.0.5	-1684.26	7	M2.0.5 - M2.1.5	6.05	3	0.11
	M2.1.5	-1681.23	10				
	M2.0.6	-1511.89	7	M2.0.6 - M2.1.6	0.92	3	0.82
	M2.1.6	-1511.44	10				
	M2.0.7	-1786.41	7	M2.0.7 - M2.1.7	4.54	3	0.21
	M2.1.7	-1784.14	10				
	M2.0.8	-1820,20	7	M2.0.8 - M2.1.8	29.20	3	0.00
	M2.1.8	-1805.60	10				
	M2.0.9	-1712.57	7	M2.0.9 - M2.1.9	69.78	3	0.00
	M2.1.9	-1677.68	10				
	M2.0.10	-1625.72	7	M2.0.10 - M2.1.10	0.44	3	0.93
	M2.1.10	-1625.49	10				
	M2.0.11	-1606.48	7	M2.0.11 - M2.1.11	19.54	3	0.00
	M2.1.11	-1596.71	10				
	M2.0.12	-1648.43	7	M2.0.12 - M2.1.12	17.28	3	0.00
	M2.1.12	-1639.79	10				
	M2.0.13	-1812.92	7	M2.0.13 - M2.1.13	8.26	3	0.04
	M2.1.13	-1808.79	10				
	M2.0.14	-1781.11	7	M2.0.14 - M2.1.14	3,58	3	0.31
	M2.1.14	-1779.32	10				
	M2.0.15	-1734.47	7	M2.0.15 - M2.1.15	25.06	3	0.00
	M2.1.15	-1721.94	10				
	M2.0.16	-1686.26	7	M2.0.16 - M2.1.16	6.49	3	0.09
	M2.1.16	-1683.02	10				
	M2.0.17	-1619.53	7	M2.0.17 - M2.1.17	23.48	3	0.00
	M2.1.17	-1607.79	10				
	M2.0.18	-1477.03	7	M2.0.18 - M2.1.18	0.51	3	0.92
	M2.1.18	-1476.78	10				
	M2.0.19	-1794.78	7	M2.0.19 - M2.1.19	6.84	3	0.08
	M2.1.19	-1791.36	10				
	M2.0.20	-1773.69	7	M2.0.20 - M2.1.20	85.47	3	0.00
	M2.1.20	-1730.96	10				
	M2.0.21	-1743.57	7	M2.0.21 - M2.1.21	6.29	3	0.10

Table 7. *(Continued)*

Step	Model	LL	Npar	Comparison	Δ-2LL	ΔNpar	*p*
2	M2.1.21	-1740.43	10				
	M2.0.22	-1705.71	7	M2.0.22 - M2.1.22	4.89	3	0.18
	M2.1.22	-1703.26	10				
	M2.0.23	-1600,48	7	M2.0.23 - M2.1.23	9.71	3	0.02
	M2.1.23	-1595.62	10				
3	M3.0	-7696.40	106	M1.0 - M3.0	369.47	33	0.00
				M3.0 - M1.1	58.87	36	0.01
4	M4.1	-7697.10	104	M4.1 - M3.0	1.39	2	0.50
	M4.2	-7698.16	104	M4.2 - M3.0	3,51	2	0.17
	M4.3	-7706.11	104	M4.3 - M3.0	19.41	2	0.00
	M4.4	-7696.54	104	M4.4 - M3.0	0.28	2	0.87
	M4.5	-7698.47	104	M4.5 - M3.0	4.13	2	0.13
	M4.6	-7699.63	104	M4.6 - M3.0	6.46	2	0.04
	M4.7	-7699.26	104	M4.7 - M3.0	5.72	2	0.06
	M4.8	-7702.96	104	M4.8 - M3.0	13.12	2	0.00
	M4.9	-7696.60	104	M4.9 - M3.0	0.38	2	0.83
	M4.10	-7696.97	104	M4.10 - M3.0	1.13	2	0.57
	M4.11	-7699,381	104	M4.11 - M3.0	5.95	2	0.05
5	M5.0	-7708.26	90	M5.0 - M3.0	23.70	16	0.10

Table 8. *Latent class MIMIC analysis results in the UK-USA sample.*

Step	Model	LL	Npar	Comparison	Δ-2LL	ΔNpar	*p*
1	M1.0	-7002.28	73	M1.0 - M1.1	108.35	69	0.00
	M1.1	-6948.11	142				
2	M2.0.1	-1548.69	7	M2.0.1 - M2.1.1	3.94	3	0.27
	M2.1.1	-1546.72	10				
	M2.0.2	-1556.39	7	M2.0.2 - M2.1.2	2.71	3	0.44
	M2.1.2	-1555.04	10				
	M2.0.3	-1505.45	7	M2.0.3 - M2.1.3	2.88	3	0.41
	M2.1.3	-1504.01	10				
	M2.0.4	-1538,00	7	M2.0.4 - M2.1.4	3,57	3	0.31
	M2.1.4	-1536.21	10				
	M2.0.5	-1452.23	7	M2.0.5 - M2.1.5	3.11	3	0.37
	M2.1.5	-1450.68	10				
	M2.0.6	-1272.42	7	M2.0.6 - M2.1.6	0.14	3	0.99
	M2.1.6	-1272.35	10				
	M2.0.7	-1561.46	7	M2.0.7 - M2.1.7	2.48	3	0.48
	M2.1.7	-1560.22	10				
	M2.0.8	-1534.03	7	M2.0.8 - M2.1.8	1.58	3	0.66
	M2.1.8	-1533.24	10				
	M2.0.9	-1474.56	7	M2.0.9 - M2.1.9	2.99	3	0.39
	M2.1.9	-1473.07	10				

Table 8. *(Continued)*

Step	Model	LL	Npar	Comparison	Δ-2LL	ΔNpar	*p*
2	M2.0.10	-1382.50	7	M2.0.10 - M2.1.10	10.68	3	0.01
	M2.1.10	-1377.16	10				
	M2.0.11	-1431.26	7	M2.0.11 - M2.1.11	0.62	3	0.89
	M2.1.11	-1430.95	10				
	M2.0.12	-1382.67	7	M2.0.12 - M2.1.12	1.55	3	0.67
	M2.1.12	-1381.90	10				
	M2.0.13	-1575.71	7	M2.0.13 - M2.1.13	3,54	3	0.32
	M2.1.13	-1573.94	10				
	M2.0.14	-1536.97	7	M2.0.14 - M2.1.14	3.61	3	0.31
	M2.1.14	-1535.16	10				
	M2.0.15	-1477.89	7	M2.0.15 - M2.1.15	2.15	3	0.54
	M2.1.15	-1476.82	10				
	M2.0.16	-1476.56	7	M2.0.16 - M2.1.16	11.40	3	0.01
	M2.1.16	-1470.86	10				
	M2.0.17	-1345.95	7	M2.0.17 - M2.1.17	1.96	3	0.58
	M2.1.17	-1344.98	10				
	M2.0.18	-1249.37	7	M2.0.18 - M2.1.18	2.74	3	0.43
	M2.1.18	-1248,00	10				
	M2.0.19	-1557.72	7	M2.0.19 - M2.1.19	1.70	3	0.64
	M2.1.19	-1556.87	10				
	M2.0.20	-1501.45	7	M2.0.20 - M2.1.20	2.93	3	0.40
	M2.1.20	-1499.98	10				
	M2.0.21	-1516.76	7	M2.0.21 - M2.1.21	2.25	3	0.52
	M2.1.21	-1515.64	10				
	M2.0.22	-1469.22	7	M2.0.22 - M2.1.22	2.81	3	0.42
	M2.1.22	-1467.82	10				
	M2.0.23	-1348.14	7	M2.0.23 - M2.1.23	8.27	3	0.04
	M2.1.23	-1344.01	10				
3	M3.0	-6982.99	82	M1.0 - M3.0	38.59	9	0.00
				M3.0 - M1.1	69.76	60	0.18
4	M4.1	-6983.53	80	M4.1 - M3.0	1.08	2	0.58
	M4.2	-6983.54	80	M4.2 - M3.0	1.11	2	0.57
	M4.3	-6987.32	80	M4.3 - M3.0	8.66	2	0.01
5	M5.0	-6984.10	78	M5.0 - M3.0	2.21	4	0.70

International Journal of Assessment Tools in Education
2024, Vol. 11, No. 2, 270–287

https://doi.org/10.21449/ijate.1268614

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

Research Article

Investigation of a multistage adaptive test based on test assembly methods

Ebru Doğruöz [1,*], **Hülya Kelecioğlu** [2]

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

ARTICLE HISTORY

Received: Mar. 21, 2023

Accepted: Feb. 13, 2024

Keywords:

Adaptive multistage test,
Test assembly,
PISA.

Abstract: In this research, multistage adaptive tests (MST) were compared according to sample size, panel pattern and module length for top-down and bottom-up test assembly methods. Within the scope of the research, data from PISA 2015 were used and simulation studies were conducted according to the parameters estimated from these data. Analysis results for each condition were compared in terms of mean RMSE and bias. According to the results obtained from the MST simulation based on the top-down test assembly method, mean RMSE values reduced when the module length increased and when the panel pattern changed from 1-2 to 1-2-2 and 1-2-3 for MST applied to small and large samples. Within the scope of the research, data from PISA 2015 were used and simulation studies were conducted using the parameters estimated from these data. Analysis results for each condition were compared in terms of mean RMSE and bias.

1. INTRODUCTION

The combination of computer technology and test implementations with item response theory (IRT) led to the emergence of computer adaptive tests (CAT). While these tests involve the use of a computer and are tailored to the examinee, IRT allows the opportunity to develop, apply and evaluate a test by considering the abilities of the examinee. Due to these advantages, CAT was used instead of paper and pencil tests. The first application of an adaptive test in the computer environment was completed by Reckase in 1974 (Wise & Kingsbury, 2000). In addition, the emergence and development of item response theory has enabled the realization of adaptive tests through the parameterization of examinee's abilities and item characteristics (Linden & Glas, 2000). Through computers, the examinees' ability can be estimated instantly after each response to an item. Thus, the next item is selected according to the examinee's ability. Accordingly, CAT has been adopted and used in many national and international exams around the world (Khorramdel et al., 2020; Kirsch & Lennon, 2017). Today, some of these exams prefer MST instead of CAT. For example, GRE (Graduate Record Examinations), PIAAC (Program for International Assessment of Adult Competencies), AICPA (American Institute of Certified Public Accountants') and MAPT (Massachusetts Adult Proficiency Test) use MST instead of CAT because of its advantages (American Institute of Certified Public Accountants, 2019; Educational Testing Service, 2018; Hogan et al., 2016; Zenisky et al.,

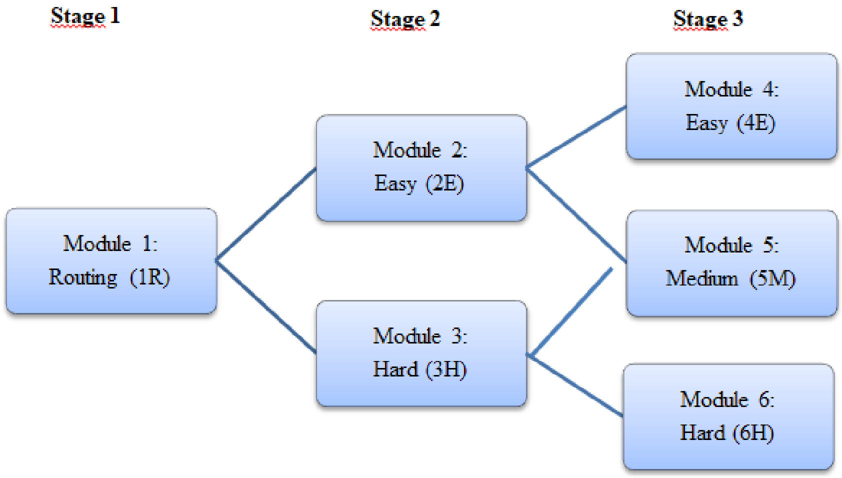
*CONTACT: Ebru Doğruöz ✉ ebrudemircioglu@karatekin.edu.tr Çankırı Karatekin University, Faculty of Humanities and Social Sciences, Department of Educational Measurement and Evaluation, Çankırı, Türkiye

e-ISSN: 2148-7456

2009). One of the reasons behind this trend is that MST acts as a bridge between linear test forms of paper and pencil testing and computer-based tests and computer-based test forms that are adaptable at item level. MST is both an adaptive test and also allows the opportunity for the test developer to investigate the test form ahead of time and check examinee's responses (Yan et al., 2014).

MST is defined as a a type of computerized adaptive testing allowing adaptation of the difficulty of the test according to the ability level of the examinee being tested. This assessment type comprises clustered components called *modules*, *stages*, *panels* and *pathways*. The smallest element of this cluster is the *module*. A module is a group of items formed by bringing items together. The level of module or modules is called the *stage*. A *panel* is a pattern formed by combining stages. The panel is the largest component of MST. For example, a panel formed with 1 module in the first stage, 2 modules in the second stage and 3 modules in the third stage is called the '1-2-3' MST panel pattern. The route taken by an examinee between stages and modules in the panel is called the *pathway*. Each examinee only follows one pathway during the test (Zenisky & Hambleton, 2014). The schematic appearance of the MST components is presented in Figure 1.

Figure 1. *An example of 3-stage MST panel.*



1.1. Test Assembly

Based on a variety of statistical features, the combination of items chosen from the item pool on test form is called test assembly. The assembly of the forms is formulated as a combinatorial optimization (CO) problem, referred to as the test assembly problem (Papadimitriou & Steiglitz, 1982; Theunissen, 1985; van der Linden & Boekkooi-Timminga, 1989). CO is the research of an element in a finite cluster optimized to a certain function. The CO problem may be formulated as in Equation 1.1:

$$\text{To maximize } \mathbf{F}(\mathbf{x}) \qquad (1.1)$$
$$\text{Subject to } \mathbf{x} \in X$$

$\mathbf{x} = (x_1, x_2, \ldots\ldots, x_n)^T$ is a binary decision vector describing a test. When $x_i = 1$, the item i is included in the test; when $x_i = 0$ the item i is not included on the test.

n is the number of items in the item pool.

X includes all binary vectors each describing a feasible test. For this reason, this set is called the *feasible set*. In practice, the feasible set is not given explicitly; however, it is implicitly indicated by an equation constraining the decision vector and a list of inclusions. This list directly comprises the test properties. For example, the applicable set containing items from 5 to 10 is presented in Equation 1.2:

$$5 \leq \sum_{i=1}^{n} x_i \leq 10 \tag{1.2}$$

$$x_i \in \{0,1\}$$

For this feasible set, the second restriction does not involve any CO problem. For example, for each appropriate solution **x** = $(x_1, x_2, \ldots\ldots, x_n)^T$ there should be a binary vector.

F(x) is a vector function; in other words, the target function (Veldkamp, 1999). For example, the Maximum Fisher Information of an adaptive with θ' ability estimation is calculated with the function in Equation 1.3:

$$\text{To maximize } \sum_{i=1}^{n} \mathrm{I_i}(\theta')x_i \tag{1.3}$$

$\mathrm{I_i}(\theta')$ is the Fisher information for item *I* at θ' ability level (Lord, 1980).

Accordingly, estimating the maximum number of non-overlapping tests that can be obtained from an item pool given the test characteristics is very important in the construction of the item pool. It should be noted that test pooling for MST is a very complex process. This is because test combination in MST is realized by simultaneously creating many panels that are parallel in terms of both coverage and psychometric properties. This combination is performed in two steps: (1) assembling modules from the item pool and (2) assembling panels of modules. These panels should also consist of modules that fulfill certain statistical requirements, such as target test information functions (TIFs) (Luecht & Nungester, 1998). In this context, limitations related to content balancing, exposure control, coverage effects, cognitive knowledge levels of test takers, item and test item overlap, item format, and word count must also be met (Hendrickson, 2007). For this reason, test combining in MSTs is usually performed through automatic test assembly (ATA) algorithms and computer programs (Breithaupt & Hare, 2007; Breithaupt et al., 2005; Luecht, 2000; Luecht, 2006; Luecht et al., 2006; Luecht & Nungester, 1998; van der Linden, 2005).

1.1.1. *Automated test assembly method*

Automated test assembly (ATA) is a modern approach to test assembly that applies advanced optimization algorithms on computers to automatically generate test forms. The most important feature of ATA is that it greatly improves the efficiency and accuracy of test assembly. This is because ATA enables computer-based selection of a suitable set of items from a large pool of pre-calibrated items (Theunissen, 1985; van der Linden, 2005; van der Linden & Boekkooi-Timminga, 1989; Veldkamp et al., 2013). The automated test assembly method may be applied with ATA computer software (e.g., CASTISEL, ConTEST) making calculation processes easier for test developers. The aim is to create test panels by choosing items from the item pool in modules taking into account the constraints such as content area, word count and item type. In this way, the process of choosing items from the item pool for modules is more convenient. This situation allowed the module development process to become more standardized.

1.1.2. *Test assembly methods: Top-down and bottom-up*

Luecht and Nungester (1998) recommended two strategies for the assembly of MST panels: top-down test assembly and bottom-up test assembly. Both strategies first require items to be assembled one by one into modules, then modules are assembled into panels. However, there are statistical differences in the stage of creating panels by combining modules between the strategies. The top-down test assembly strategy freely mixes and matches modules to create panels. The bottom-up test assembly strategy requires selective matching of modules to create panels. This is an indicator that the top-down test assembly method has a more complicated structure compared to the bottom-up test assembly method.

When combining modules to create panels in both test assembly strategies, the following steps are taken (Luecht & Nungester, 1998):

a) Production of statistical targets for test samples in different stages,

b) Determination of content features in the stages,
c) Creation of panels by combining modules abiding by the restrictions in the first and second steps.

Selection of statistical targets for modules is the most important decision in designing the MST pattern (Hendrickson, 2007). Zheng et al. (2012) created MST according to the top-down test assembly method based on the automatic approach. They compared this method with paper and pencil testing and CAT. According to the results of the study, MST utilized the item pool more effectively compared to pencil paper test and CAT, and the classification was performed more accurately. In a study discussing possible applications of adaptive or multistage tests for a Law Faculty Acceptance Test and considering the main approaches applied in the development of test assembly methods, a single-form test assembly approach was concluded to be an applicable method for testing in programs where the test is defined only by restrictions (Belov, 2016).

When research about test assembly methods is generally investigated, the common point appears to be that studies researched the top-down test assembly method proposed by Luecht and Nungester (1998) among test assembly methods and the test assembly method completed during exams. However, there are no experimental studies on how these test assembly methods give results under different conditions. Therefore, there is a question mark about whether the right decision is made in determining the test assembly method to be selected. The bottom-up test assembly method is chosen less often than the top-down test assembly method and is more advantageous for short test applications, which has made the top-down test assembly method a focal point for research. In the related literature, the bottom-up test assembly method was mostly used in the existing applications of MST (Hembry, 2014; Jodoin et al., 2006; Lu, 2010; Luecht et al., 2006; Wang, 2013; Wang, 2017; Yang, 2016; Zheng, 2014). There are a few studies in which the top-down test assembly method was used (Davis & Dodd, 2003; Lynn Chen, 2010; Zheng et al., 2016). For this reason, it is believed that this study can guide researchers on which of the 'top-down' or 'bottom-up' test assembly methods to prefer in the process of constructing the MST. Additionally, comparisons were made of the elements comprising MST like panel pattern, module length and stage number. In this framework, recommendations were developed regarding the module length, panel pattern and sample size required to make estimations with minimum error and bias in MST constructs. Because measurement precision in MSTs can be affected by module length and panel pattern (Zenisky & Hambleton, 2014). In addition, within the scope of the study, data from PISA 2015 were used and a simulation study was conducted using the parameters estimated from these data. PISA 2015 is an international, validated and reliable assessment, and this computer-based application is the basis for the MST to be used in the coming years, which is one of the reasons why PISA data were preferred in the research. Thus, a post-hoc simulation study was conducted based on real data. This is one of the important features that make the research strong. The results obtained in the study are expected to contribute to the applicability of MST. In line with this, within the scope of the present research, the aim was to compare test assembly methods and answer the following questions.

How do test assembly methods (top-down, bottom-up) impact the estimation of ability estimation conditional on module length, panel pattern, and sample size?

1.2. Subproblems

1. What changes occur in RMSE and bias values according to module length (6 and 12), panel pattern (‘1-2’, ‘1-2-2’ and ‘1-2-3’) and sample size (250 and 2000) for the top-down test assembly method in MST applications?
2. What changes occur in RMSE and bias values according to module length (6 and 12), panel pattern (‘1-2’, ‘1-2-2’ and ‘1-2-3’) and sample size (250 and 2000) for the bottom-up test assembly method in MST applications?

2. METHOD

2.1. Research Model

In this research, the aim was to compare the performance of test assembly methods for MST patterns with different features using IRT-based estimation and post-hoc simulation methods for the science literacy ability of examinees participating in the PISA implementation completed in 2015. For this purpose, real item data were used in the study. Therefore, this study is descriptive research based on post hoc simulation using real item parameters. Simulation studies consist of data generation and analysis processes appropriate to real-life situations (Burton et al., 2006; Ranganathan & Foster, 2003). Simulation data are often preferred because most MST applications have implementation problems, require a large sample size and a large item bank (Pihlainen et al., 2018; Xu et al., 2021; Zheng & Chang, 2015).

2.2. Participants

The participants in the research comprised examinees participating in PISA 2015 in the field of science literacy and answered booklet 91 because it is suitable for the structural features of the MST. The reason for choosing science literacy is that it constitutes the predominant area of the PISA 2015 application. Nearly 540,000 students representing 29 million students in nearly 72 countries, including 35 OECD countries, participated in the PISA 2015 implementation (OECD, 2015). Booklet number 91 was chosen as the data collection tool for the study, since the number of science literacy items and examinees who received the booklet were higher than for the other booklets, out of a total of 66 booklets (Forms 31-96) created according to the computer-based test. This booklet contained a total of 501 items in a variety of categories (two categories, multiple categories, open-ended) in the science literacy field. The item pool for the study comprised 159 items with two categories among the total of 501 items in booklet number 91. Analyses were completed on the dataset related to 15,059 students who answered these items.

2.3. Analysis

Analysis of data in the study was completed in two stages. In the first stage, data obtained from the study group comprising students participating in the PISA exam in 2015 were analyzed according to the 2 PL model based on IRT and an item pool was created for MST. In line with this, first, the data set obtained from the PISA implementation in 2015 was tested for a single dimension, local independence, model-data fit, item and ability parameter invariance assumptions. The suitability of the data set for factor analysis was tested with Bartlett's test and Kaiser-Meyer-Olkin (KMO) criteria (Bartlett's = 1584902.1, *sd* = 12561, *p* = 0.00; KMO = 0.98) and the data set was suitable. Item parameters and ability parameters related to examinees were estimated with the BILOG-MG (Zimowski et al., 1996) program. As the items had low correlation in the limited ability interval and the single dimension assumption was met, the local independence assumption was accepted. With the aim of investigating which logistic model was suitable for the data set, the data set was analyzed for suitability to 1 PL, 2 PL and 3 PL models. Accordingly, considering the difference between the –2 log (probability) values for 3 PL and 2 PL models was not much, the 2 PL model was chosen (–2 log (probability) $_{(1\text{ PL})}$ = 2125726.00 –2 log (probability) $_{(2\text{ PL})}$ = 2017798.00, (–2 log (probability) $_{(3\text{ PL})}$ = 1977773.91). These results are consistent with the technical report released by OECD (OECD, 2017). Thus, the 2 PL model was estimated to be suitable for calibration of the two-category data set identified to have a single dimension. According to descriptive statistics related to item and ability parameters, the data set had an item discrimination parameter value mean 1.16 and a standard deviation of 0.06 and a difficult parameter value mean 0.07 and a standard deviation of 0.30. The smallest ability parameter of individuals was -2.85, while the highest ability parameter was calculated as 2.97. In order to determine the invariance of item parameters, individuals were randomly divided into 11 groups. The item parameters were estimated according to the 2 PL model in different groups and the item parameters were compared between groups with the

Pearson moment multiplication correlation technique. Finally, significant and high levels of correlation (p<0.01) were identified between item parameters estimated in 11 groups comprising 1.369 individuals each and the invariance of item parameters assumption was met. The invariance of ability parameters was identified with significant positive, high-level correlations between ability parameters estimated in three randomly assigned subgroups comprising 53 items for all 15.059 individuals.

In the second stage of data analysis, an MST simulation was developed for each subproblem. Analyses were completed with item parameters chosen in accordance with 24 simulation conditions from the item pool and according to individual ability chosen in accordance with 24 simulation conditions. To create the MST, the 'xxIRT' (Luo, 2017) program using R (R Development Core Team, 2011) software was used. With the aim of increasing the generalizability of the results, 30 repeats were performed for each condition (Tian, 2018). The MST variables used in the MST simulations were test assembly (top-down and bottom-up), module length (6 and 12), panel pattern ('1-2', '1-2-2' and '1-2-3') and sample size (250 and 2000).

2.3.1. *Panel pattern*

In the research, MSTs with two ('1-2') and three ('1-2-2' and '1-2-3') stage panel patterns were created. These three-panel patterns were used in the research as they are included among the most researched MST panel patterns (Jodoin et al., 2006; Luecht et al., 2006; Wang, 2017; Zenisky, 2004). The '1-2' panel pattern comprises two stages and one panel. In the first stage, there is Module-1 (M) with a moderate difficulty level and in the second stage there is Module-2 (E) with an easy difficulty level and Module-2 (H) with a high difficulty level. The '1-2-2' panel pattern comprises three stages and two panels. The first stage includes Module-1 (M) with moderate difficulty, the second stage includes Module-2 (E) with easy difficulty and Module-2 (H) with high difficulty level and the third stage includes Module-3 (E) with easy difficulty and Module-3 (H) with high difficulty level. The '1-2-3' panel pattern comprises three stages and two panels. The first stage includes Module-1 (M) with a moderate difficulty level, the second stage comprises Module-2 (E) with easy difficulty and Module-2 (H) with high difficulty and the third stage includes Module-3 (E) with easy difficulty, Module-3 (M) with moderate difficulty and Module-3 (H) with high difficulty level.

2.3.1.1. Module length. The test length in MST studies was identified to vary between 33 and 60 items (Hambleton & Xing, 2006; Jodoin et al., 2006; Patsula, 1999; Zenisky, 2004). In this research, the number of modules representing short test length was chosen as 6, while the number of modules representing moderate test length was determined to be 12, twice that of the short test length. MSTs were designed so that when module length was 6, with the '1-2' panel pattern, individuals answered a total of 12 items, and with the '1-2-2' and '1-2-3' panel patterns they answered a total of 18 items. When the module length was 12, with the '1-2' panel pattern, individuals answered a total of 24 items, and with the '1-2-2' and '1-2-3' panel patterns they answered a total of 36 items.

2.3.1.2. Item Pool. There was a total of 159 items in the two-category data set calibrated according to the 2 PL model obtained from the PISA data administered in 2015.

2.3.1.3. Sample Size. The research sample comprised 250 and 2000 individuals chosen at random from among 15,059 individuals participating in the PISA test in 2015. When the literature is investigated, it appears sample sizes from 250 (Yan et al., 2014) to 5000 studies in MST research (Dallas, 2014; Sari, 2016; Wang, 2017; Xing & Hambleton, 2004; Yang, 2016). In this research, as the target was to assess the applicability to small samples in addition to large samples for the MST pattern, in the research 250 individuals represented the small sample and 2000 individuals represented the large sample.

2.3.1.4. Test Assembly. Most commonly used top-down and bottom-up automated test assembly methods used in MST studies were chosen. For both methods, the target test information function (TIF) value was determined with the mean maximum information (MMI) (Luecht, 2000; Luecht et al., 2006) strategy.

2.3.1.5 Referral Strategy and Scoring. In this study, the referral strategy was chosen as the commonly-used mean maximum information (MMI) strategy and scoring was done according to maximum likelihood estimation (MLE) (Luecht et al., 2006; Zenisky et al., 2010).

2.3.2. *Test administration*

The steps followed during test administration of the '1-2', '1-2-2' and '1-2-3' panel patterns investigated in the study were: (a) the individual was assigned one of two different panel patterns at random, (b) the individual responded the referral module (moderate difficulty) they were assigned, (c) after completing the referral module, the individual's ability was estimated using the maximum probability estimation (MPE), (d) after the first stage, the estimated ability of the individual (θ) and previously determined referral points were compared, and the individual was directed from the first stage to the second stage, (e) after the second stage, the individual's ability was again estimated with the MPE method, and (f) the test ended here for individuals tested with the '1-2' panel structure. For test administration of individuals tested with the '1-2-2' and '1-2-3' panel structures, their ability was predicted after the second stage (θ) and compared with the previously determined referral points and the individual was referred from the second stage to the third stage.

2.3.3. *Evaluation criteria*

The performance of the MST based on ability estimation was assessed according to mean RMSE and bias criteria that are frequently used in MST studies (Xiao & Bulut, 2022; Kim et al., 2015; Park, 2015; Sari & Raborn, 2018; Zheng, 2014). RMSE, and bias values for each simulation condition were calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{j=1}^{N}(\hat{\theta}_j - \theta_j)^2}{N}}, \text{ and}$$

$$Bias = \frac{\sum_{j=1}^{N}(\hat{\theta}_j - \theta_j)}{N},$$

where $\hat{\theta}_j$ and θ_j, j examinee predicted and true ability values; N is the total number of examinee. Multivariate analysis of variance (ANOVA) was used to test the significance of the MST variables in various conditions on the mean RMSE and bias values. 'Bonferroni' multiple comparison test was used to find out between which conditions the differences between the means were between. When interpreting the effect size, 0.00-0.19 was taken as very small, 0.20-0.49 as a small, 0.50-0.79 as medium, and 0.80 and larger as large effect sizes (Cohen, 1988).

3. FINDINGS

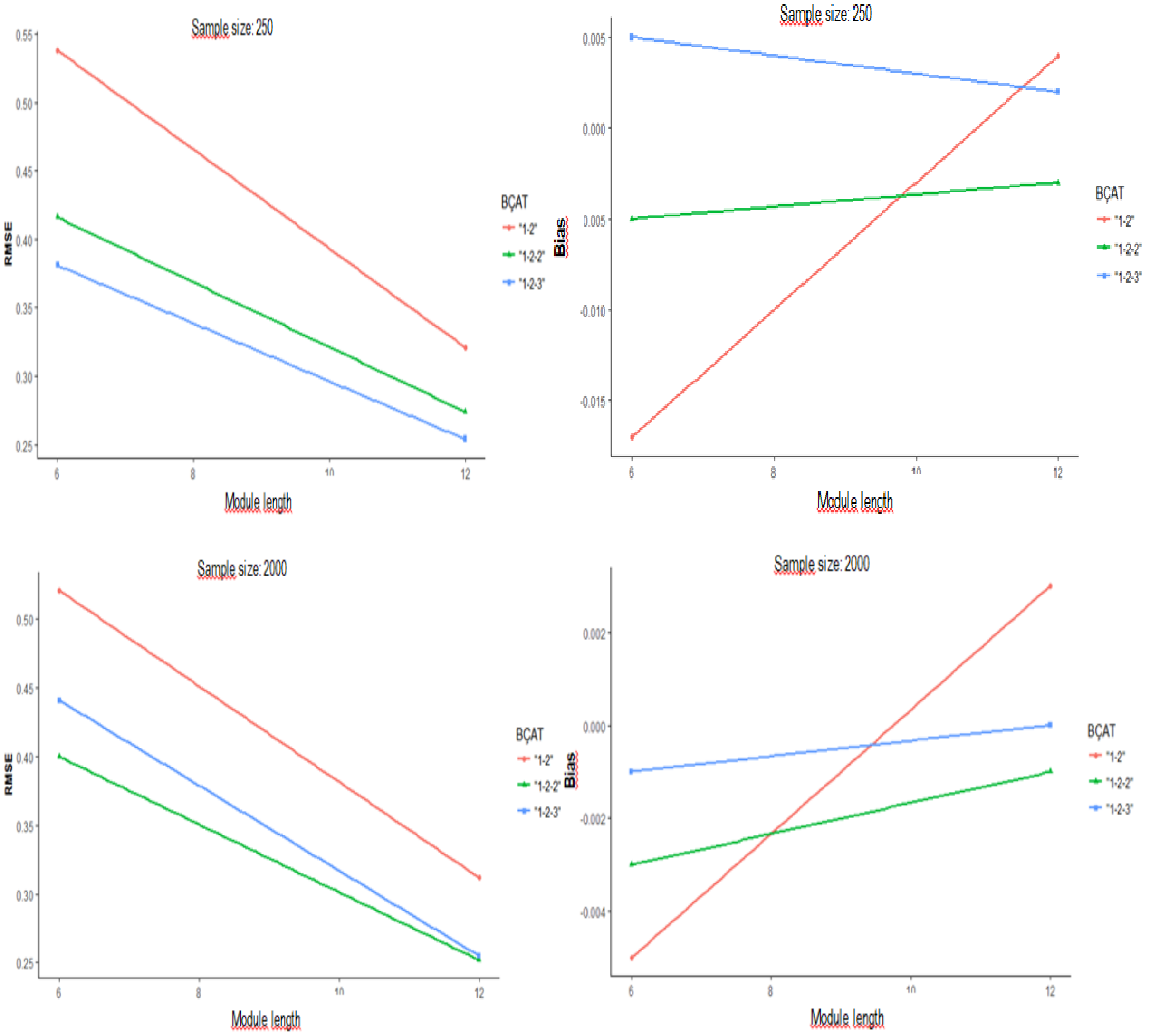
3.1. Findings on the Top-down Test Assembly Methods

We examined how the precision of ability estimation changes according to model lengths (6 and 12), panel patterns ("1-2", "1-2-2" and "1-2-3") and sample sizes (250 and 2000) in the top-down test assembly method in the MST application. In line with this, in order to interpret the findings, firstly MSTs were created in accordance with the simulation conditions in the problem. The findings related to ability estimation in MSTs created according to a variety of simulation conditions are presented in Table 1 and Figure 2.

Table 1. *Mean RMSE and bias values for MST created according to top-down test assembly method.*

Sample	Panel pattern	Module length	RMSE	Bias
250	“1-2”	6	0.538	-0.017
		12	0.321	0.004
	“1-2-2”	6	0.416	-0.005
		12	0.274	-0.003
	“1-2-3”	6	0.381	0.005
		12	0.254	0.002
2000	“1-2”	6	0.521	-0.005
		12	0.312	0.003
	“1-2-2”	6	0.400	-0.003
		12	0.252	-0.001
	“1-2-3”	6	0.441	-0.001
		12	0.255	0.000

Figure 2. *Plots of mean RMSE and bias values for MST created according to top-down test assembly method.*



As can be seen in Table 1, with the top-down test assembly method, the mean RMSE values obtained for different sample sizes, test lengths and panel patterns varied from 0.252 to 0.538. When the general lines of the results are investigated, the lowest error estimation was for the moderate length module applied to the large sample size with the '1-2-2' panel pattern, while the largest error estimation was for the short-length module applied to the small sample with the '1-2' panel pattern. When findings are investigated in terms of module length, for both sample sizes as module length increased, mean RMSE values appeared to reduce. When results are investigated in terms of panel pattern, the mean RMSE amount appeared to change for all test levels with the differentiation of panel patterns in small and large samples. In the transition from the '1-2' panel pattern to the '1-2-2' and '1-2-3' panel patterns, mean RMSE values fell. However, the mean RMSE values for both module lengths for '1-2-2' and '1-2-3' panel patterns applied to large samples were different with an increase for the transition from the '1-2-2' panel pattern to the '1-2-3' panel pattern. This increase was 0.041 for short module length and 0.003 for moderate module length in the transition from '1-2-2' panel pattern to the '1-2-3' panel pattern. When findings are investigated in terms of sample size, the increase in the sample size appeared to reduce mean RMSE values for both module lengths in all patterns, apart from the '1-2-3' panel pattern. For small samples, the lowest mean RMSE value was for the '1-2-3' panel pattern with a moderate length module, and for large samples, the lowest mean RMSE value was calculated for the '1-2-2' panel pattern with the moderate length module.

If the findings related to bias in Table 1 are investigated, mean bias values generally appear to be low. When the top-down test assembly method is chosen, the mean bias values vary from -0.017 to 0.005 for sample size, panel pattern and module length simulation conditions. The highest mean bias values were for '1-2' panel patterns applied to small samples with short module lengths. This was followed by short module length in small samples with '1-2-2' and '1-2-3' patterns, and the '1-2' panel pattern applied to large samples. The lowest mean bias value was calculated for the '1-2-3' panel pattern with moderate module length applied to large samples. This value was 0.000; in other words, this simulation condition had unbiased calculations. When findings were investigated in terms of module length, as module length increased, bias in panel patterns for both sample types was concluded to be reduced. When results are investigated in terms of panel pattern, for both module lengths, in small and large samples, the transition from the '1-2' panel pattern to '1-2-2' panel pattern and from '1-2-2' panel pattern to '1-2-3' panel pattern appeared to cause a fall in mean bias values. When findings are investigated in terms of sample size, as the sample size increased, the mean bias values were observed to fall by a small amount.

Within the scope of the subproblem in the research, whether the module length, panel pattern and sample size had statistically significant effects on the mean RMSE and bias findings obtained according to the top-down test assembly method was tested with the versatile ANOVA test. The F value and effect sizes (η^2) obtained from the ANOVA test are presented in Table 2.

Table 2. *Mean RMSE and ANOVA results for mean RMSE and bias values obtained when top-down test assembly method is chosen.*

	Evaluation Criteria					
		RMSE			Bias	
Study Conditions	*df*	*F*	η^2	*df*	*F*	η^2
Module length (M)	1	3379.332*	0.051	1	20.662*	0.049
Panel pattern (P)	2	404.320*	0.012	2	1.841	0.007
Sample (S)	1	0.034	0.052	1	3.753	0.007
P*M	2	50.648*	0.015	2	2.51	0.014
P*S	2	27.878*	0.008	2	8.395*	0.042
M*S	1	10.489*	0.001	1	0.686	0.014
P*M*S	2	12.019*	0.005	2	6.059*	0.028

*p<0.05

As observed in Table 2, the mean RMSE value obtained according to the top-down test assembly method significantly differed according to module length and panel pattern ($F_{1\text{-}358(module\ length)}$ = 3379.332, $p < 0.05$; $F_{2\text{-}357(Panel\ pattern)}$ = 404.320, $p < 0.05$). The eta-square values showed the efficacy of the module length and panel pattern on mean RMSE value was at moderate levels and the effect size was very small ($\eta^2_{(module\ length)}$ = 0.051, $\eta^2_{(Panel\ pattern)}$ = 0.012). To identify which panel patterns caused the difference among the panel patterns, the Bonferroni two-way comparison test was performed. According to the results of the test, the mean RMSE value was more affected by the '1-2-3' panel pattern ($\bar{X}$ = 0.423) compared to the '1-2-2' panel pattern ($\bar{X}$ = 0.335) and '1-2' panel pattern ($\bar{X}$ = 0.333). Additionally, the effects of the interactions of panel pattern-module length ($F_{4\text{-}355(P*M)}$ = 50.648, $p < 0.05$), panel pattern-sample size ($F_{4\text{-}355(P*S)}$ = 27.878, $p < 0.05$), module length-sample size ($F_{3\text{-}356(M*S)}$ = 10.489, $p < 0.05$) and panel pattern-module length-sample size ($F_{6\text{-}353(P*M*S)}$ = 12.019, $p < 0.05$) on mean RMSE values were significant. The panel pattern-module length ($\eta^2_{(P*M)}$ = 0.015), panel pattern-sample size ($\eta^2_{(P*S)}$ = 0.008), module length-sample size ($\eta^2_{(M*S)}$ = 0.001) and panel pattern-module length-sample size ($\eta^2_{(P*M*S)}$ = 0.005) had small levels of effect on mean RMSE value. However, the sample size did not significantly change the mean RMSE value.

As seen from Table 2, when the effects of module length, panel pattern and sample size on the mean bias values obtained according to the top-down test assembly method were examined, the mean bias values only appeared to significantly differ according to module length ($F_{1\text{-}358(module\ length)}$ = 20.662, $p < 0.05$). This finding is supported by the eta-square value ($\eta^2_{(module\ length)}$ = 0.049). Panel pattern and sample size did not cause a significant change in mean bias values. When significant effects of the interactions of these three variables on mean bias value were examined, panel pattern-sample size ($F_{4\text{-}355(P*S)}$ = 8.395, $p < 0.05$) and panel pattern-module length-sample size ($F_{6\text{-}353(P*M*S)}$ = 6.059, $p < 0.05$) interactions caused significant differences in mean bias value. Additionally, the effect of these variables on mean bias was at moderate levels ($\eta^2_{(P*S)}$ = 0.042, $\eta^2_{(P*M*S)}$ = 0.028).

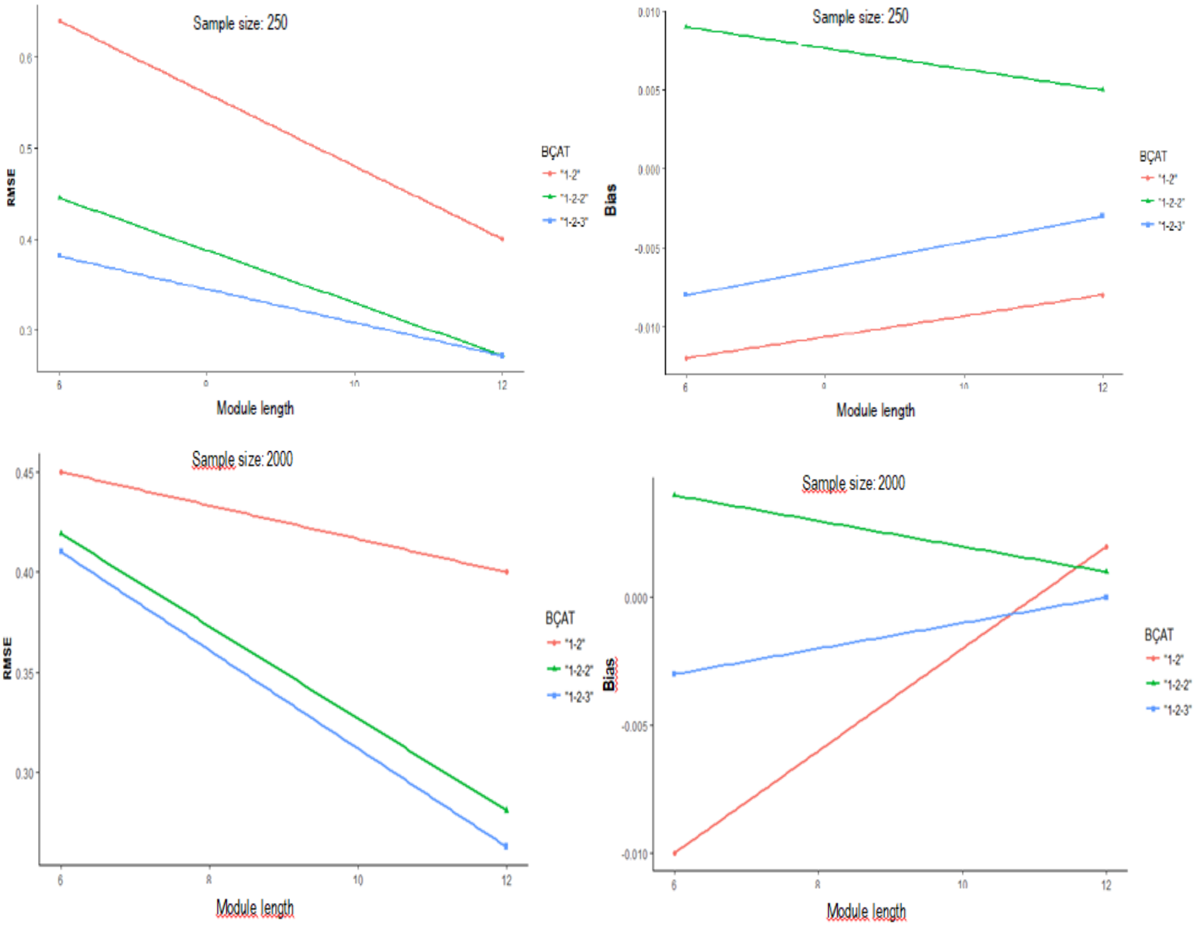
3.2. Findings on the Bottom-up Test Assembly Methods

Findings related to the change in the precision of ability estimations according to module lengths (6 and 12), panel patterns (1-2, 1-2-2 and 1-2-3) and sample sizes (250 and 2000) with the bottom-up test assembly method for MST applications are presented in Table 3 and Figure 3.

Table 3. *Mean RMSE and bias values for MST created According to bottom-up test assembly method.*

Sample	Panel pattern	Module length	RMSE	Bias
250	"1-2"	6	0.639	-0.012
		12	0.400	-0.008
	"1-2-2"	6	0.445	0.009
		12	0.272	0.005
	"1-2-3"	6	0.381	-0.008
		12	0.272	-0.003
2000	"1-2"	6	0.450	-0.010
		12	0.400	0.002
	"1-2-2"	6	0.419	0.004
		12	0.281	0.001
	"1-2-3"	6	0.410	-0.003
		12	0.263	0.000

Figure 3. *Plots of mean RMSE and bias values for MST created according to bottom-up test assembly method.*



As can be seen in Table 3, the mean RMSE values obtained for different module lengths, panel patterns and sample sizes according to the bottom-up test assembly method varied from 0.263 to 0.639. The lowest mean RMSEA estimation was for the '1-2-3' panel pattern with a moderate length module applied to a large sample, while the highest mean error estimation was for the '1-2' panel pattern with a small length module applied to a small sample. When the findings were investigated in terms of module length, for both sample sizes as the module length increased, the mean RMSE value appeared to reduce. When the findings were investigated according to panel pattern, for large and small samples, the differentiation of panel patterns changed the mean RMSE amount at all test levels. In the transition from the '1-2' panel pattern to '1-2-2' and '1-2-3' panel pattern, the mean RMSE values fell. When the findings were examined in terms of sample size, the increase in sample size appeared to reduce the mean RMSE values in many conditions. However, for the small sample, the moderate module length and '1-2-2' panel pattern, there was an increase of 0.09 when the same module length and panel pattern were applied to large samples. Additionally, when the '1-2-3' panel pattern was applied with short module length to small and large samples, a 0.29 increase was noticed. In small samples, the lowest mean RMSE value was calculated for the moderate length module with '1-2-2' and '1-2-3' panel patterns, while for large samples, the lowest mean RMSE value was obtained with the moderate length module applied in '1-2-3' panel pattern.

When the results relating to the bias obtained according to the bottom-up test assembly method are examined in Table 3, it appears that the mean values of the bias are generally very low. When the bottom-up test assembly method was chosen, the mean bias values according to module length, panel pattern and sample size simulation conditions varied from -0.012 to 0.009. The highest mean bias value belonged to the small sample with a short module length in the '1-

2' panel pattern. This pattern with short module length was followed by large samples with '1-2' patterns, then by short module length and small samples with '1-2-2' panel patterns. The lowest mean bias value for large samples was calculated for moderate module length with the '1-2-3' panel pattern. In these conditions, the calculated 0.000 mean bias value indicated bias-free calculations were performed. When the findings were examined in terms of module length, as module length increased, the bias for panel patterns in both sample types appeared to reduce. When the findings were examined in terms of panel patterns, for both module lengths with small and large samples, the transitions from the '1-2' panel pattern to '1-2-2' and from '1-2-2' panel pattern to '1-2-3' panel pattern reduced mean bias values. When the findings were examined in terms of sample size, as the sample size increased, the mean bias values appeared to reduce.

Within the scope of this subproblem, whether the effect of module size, panel pattern and sample size were statistically significant on mean RMSE and bias findings obtained according to the bottom-up test assembly method was tested with the versatile ANOVA test. The F value and effect sizes (η^2) obtained from the ANOVA test are presented in Table 4.

Table 4. *Mean RMSE and ANOVA results for mean RMSE and bias values obtained when bottom-up test assembly method is chosen.*

	Evaluation Criteria					
	RMSE			Bias		
Study Conditions	*df*	*F*	η^2	*df*	*F*	η^2
Module length (M)	1	2721.284*	0.032	1	6.400*	0.016
Panel pattern (P)	2	1000.355*	0.023	2	22.277*	0.105
Sample (S)	1	119.354*	0.001	1	1.786	0.005
P*M	2	10.654*	0.002	2	7.451*	0.033
P*S	2	140.741*	0.003	2	0.324	0.005
M*S	1	117.107*	0.001	1	0.592	0.005
P*M*S	2	149.044*	0.003	2	4.561*	0.022

*p<0.05

As can be seen in Table 4, the mean RMSE values obtained according to the bottom-up test assembly method differed significantly according to module length, panel pattern and sample size ($F_{1\text{-}358(module\ length)}$ = 2721.284, p < 0.05; $F_{2\text{-}357(Panel\ pattern)}$ = 1000.355, p < 0.05; $F_{1\text{-}358(sample)}$ = 119.354, p < 0.05). Module length and panel pattern had moderate effect on mean RMSE, while sample size had small effect ($\eta^2_{(module\ length)}$ = 0.032, $\eta^2_{(Panel\ pattern)}$ = 0.023, $\eta^2_{(sample)}$ = 0.001). To identify which panel patterns caused the difference, the Bonferroni two-way comparison test was performed. According to the test results, the '1-2-3' panel pattern ($\bar{X}$ = 0.472) had more effect on mean RMSE value compared to the '1-2-2' panel pattern ($\bar{X}$ = 0.356) and '1-2' panel pattern ($\bar{X}$ = 0.332). Additionally, the interactions of panel pattern-module length ($F_{4\text{-}355(P*M)}$ = 10.654, p < 0.05), panel pattern-sample ($F_{4\text{-}355(P*S)}$ = 140.741, p < 0.05), module length-sample ($F_{3\text{-}356(M*S)}$ = 117.107, p < 0.05) and panel pattern-module length-sample ($F_{6\text{-}353(P*M*S)}$ = 149.044, p < 0.05) had significant effects on mean RMSE value. The effect of these variables on mean RMSE was small ($\eta^2_{(P*M)}$ = 0.002, $\eta^2_{(P*S)}$ = 0.003, $\eta^2_{(M*S)}$ = 0.001, $\eta^2_{(P*M*S)}$ = 0.003).

As seen in Table 4, when the effects of module length, panel pattern and sample size on the mean bias values obtained according to the bottom-up test assembly method are examined, mean bias value differed significantly according to panel pattern and module length ($F_{1\text{-}358\ (module\ length)}$ = 6.400, p < 0.05; $F_{2\text{-}357(Panel\ pattern)}$ = 22.277, p < 0.05). The effect of module length on mean bias was small ($\eta^2_{(module\ length)}$ = 0.016), while the effect of panel pattern was at moderate levels ($\eta^2_{(Panel\ pattern)} = 0.105$). The Bonferroni two-way comparison test was performed to identify which panel pattern caused the difference. According to the test results, the '1-2-2' panel pattern ($\bar{X} = 0.008$) was more effective on mean RMSE value compared to the '1-2' panel pattern ($\bar{X} = 0.006$). However, sample size did not cause a significant difference in mean bias values. The interactions of panel pattern-module length and panel pattern-module length-sample size were observed to cause a significant difference in mean bias values ($F_{4\text{-}355(P*M)} = 7.451$, $p < 0.05$; $F_{6\text{-}353(P*M*S)} = 4.561$, $p < 0.05$). The effect of these variables on mean bias was at moderate levels ($\eta^2_{(P*M)} = 0.033$, $\eta^2_{(P*M*S)} = 0.022$).

4. DISCUSSION and CONCLUSION

Within the scope of the research, the performances of MSTs created according to top-down and bottom-up test assembly methods tested for module length, panel pattern and sample size using an item pool created from a real data set were compared. The MST components were module length, panel pattern and sample size. However, the study attempted to identify the correlation of these components with the test assembly method. For this reason, the focal point of the study was the top-down and bottom-up test assembly method recommended for combining MST panels introduced to the literature by Luecht and Nungester (1998). The research findings first showed that the module length affected mean RMSE and bias values with the top-down and bottom-up test assembly methods. For both test assembly methods, the moderate module length produced lower mean RMSE and bias values compared to the short module length. The probable reason for the difference in mean RMSE and bias values calculated for short and moderate module lengths may be the total item count. This situation may be interpreted as showing that as the total number of items in the test increases, the mean RMSE and bias values reduce. This finding is parallel to the findings of the study by Sari (2016) using the top-down test assembly method creating MST and CAT according to test management, content count and test length variables and comparing the performance of these two test types. In their study, they concluded that only test length had a significant effect on the mean RMSE value. This finding is also supported by the study of Yang (2016). In their study, the top-down test assembly method was used and as the test length increased, the RMSE and standard error values reduced. When the test length was 60, the bias was minimum, while it was maximum when the test length was 20. The mean bias values obtained according to the bottom-up test assembly method in this study significantly differed according to module length. For both samples, panel patterns with short module length had highest mean bias, while panel patterns with moderate module length had the smallest mean bias value. The study by Hembry (2014) studied the effect of two test lengths of short and moderate in MSTs created using the bottom-up test assembly method. This study had mean bias measures very close to zero and panel patterns with short test lengths had reduced mean RMSE and bias values. This finding is parallel to the findings in our research. Other similar findings were obtained in studies by Kim et al. (2013) using an OTB program as the test assembly method, Lynn Chen (2010) using the top-down test assembly method and Lu (2010) using the bottom-up test assembly method. A study by Zheng (2014) using the top-down test assembly method did not find a consistent difference between different module lengths.

However, in addition to the top-down and bottom-up test assembly methods, there are some MST studies, though few, using NAMSS, one of the automatic assembly methods. One of these studies by Dallas (2014) studied the directive and point effects of MSTs created by using 10 and 20 module lengths. The results of the study were similar to the results obtained for module lengths affecting MSTs investigated according to top-down and bottom-up test assembly methods completed in this study.

As supported by the studies mentioned above, the effect of module length on mean RMSE and bias values and the reason for the fall in mean RMSE and bias values as module length increases may be due to MSTs comprising short tests having lower measurement sensitivity. Longer tests ensure higher classification accuracy and consistency (Crocker & Algina, 1986; Luo, 2020).

Another finding in the research is the effect of panel patterns in top-down and bottom-up test assembly methods on mean RMSE and bias values. The change from the '1-2' panel pattern to '1-2-2' and '1-2-3' panel patterns according to the top-down test assembly method reduced mean RMSE and bias values in many conditions, while for the bottom-up test assembly method, it reduced mean RMSE and bias values in all conditions. This finding may be interpreted as showing that the increase in stage numbers in MST panels reduces the mean RMSE and bias. This finding is supported by the findings of a study comparing three-stage and two-stage MSTs by Patsula (1999), which found that three-stage MSTs produced less measurement error than two-stage MSTs. Additionally, the findings obtained from this study are consistent with the results of a study based on the 3 PL model by Zenisky (2004). Another similar finding was encountered in the study by Hembry (2014). In this study, MSTs created using the bottom-up test assembly method were investigated in four panel patterns of '1-3', '1-5'. '1-3-3' and '1-5-5'. Very small differences were obtained for estimated ability and mean bias values for the four panel patterns. Generally, mean bias measures very close to zero were obtained, as in this study. RMSE values were lower for the two-stage tests, different to the findings of this study. However, the difference between panel patterns was reported to be very low in this study. Additionally, there was a significant difference between '1-2', '1-2-2' and '1-2-3' panel patterns according to both methods in this research, with the '1-2-3' panel pattern concluded to have more effect on mean RMSE and bias values compared to other patterns. Sari (2016) obtained different findings in a study completed with the bottom-up test assembly method. In this study, the effects of the two-stage '1-3' and three-stage '1-3-3' panel patterns on RMSE were investigated, and it was reported that no significant difference was found. In research applying the bottom-up test assembly method, Yang (2016) obtained similar results to Sari (2016). In this study, four-panel structures were investigated ("1-3", "1-5", "1-3-3" and "1-5-5") and significant differences were not found. Another parallel finding to these studies was obtained in the study by Jodoin et al. (2006) and Luo and Kim (2018). Studies by Zheng et al. (2012) and Zheng (2014) used the top-down test assembly method and reported no significant differences were found between four-stage models and three-stage models. As can be seen, there are two different results about the effect of panel patterns on MST studies. The probable cause for the different results may be other variables that were fixed in both studies. In fact, it should not be ignored that increasing the number of stages in the panel structure may provide better measurement sensitivity as it is directly proportional to the individual's responses to higher numbers of items.

According to the research findings, for the top-down test assembly method, for '1-2' and '1-2-2' panel patterns with short and moderate module length, the increase in sample size lowered the mean RMSE and bias values. For the bottom-up test assembly methods, the increase in sample size with the '1-2' panel pattern for short and moderate module lengths, the '1-2-2' panel pattern with short module length and the '1-2-3' panel pattern with moderate module length lowered mean RMSE and bias values. Additionally, for the bottom-up test assembly method, the sample size had a statistically significant effect on mean RMSE and bias values, while for the top-down test assembly method, it was concluded there was no significant effect. In this context, no definite conclusion can be made about which test assembly method should be chosen for small or large sample sizes. In fact, in international studies of MST, the use of large-scale tests is an indicator of applicability for large samples. Based on the research findings, interpretations can be made about the applicability of the bottom-up test assembly method using '1-2' panel pattern with short and moderate module lengths, '1-2-2' panel pattern for short module lengths and '1-2-3' panel pattern with moderate module length for large samples. The reason for choosing sample size as a variable in the research is to ensure the ability to see possible outcomes when MST's, used for samples in large-scale international tests, are applied to institutional exams like for inspectors, specialists, and judges, completed with smaller samples in our country, or even in lesson selection exams applied in middle schools and high schools, in the future if appropriate computer infrastructure is developed. In a similar study investigating small sample sizes, Yan et al. (2014) investigated MSTs in accordance with the 'tree-based' approach. The study concluded that MSTs applied to small samples displayed good performance.

When assessed as a whole, the top-down and bottom-up test assembly methods produced similar findings and both methods are recommended for use in creating MSTs. Additionally, the research investigated the top-down and bottom-up test assembly methods among automatic test assembly methods. Later studies are recommended to study other methods like ASM, NAMSS and maximum priority index, in addition to these methods, linear programming methods and the test assembly method performed at the time of the exam called the 'on-the-fly' test assembly method in the literature. In the research, item and ability parameters suitable for only the 2 PL model were estimated as the item pool was created according to a real test set and the MST was created accordingly. In later studies, parameters may be estimated according to 2 PL and 3 PL models to research the effect of logistic models on MST performance.

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.

Contribution of Authors

Ebru Doğruöz: Investigation, Resources, Visualization, Software, Analysis, and Writing-original draft. **Hülya Kelecioğlu:** Methodology, Supervision, and Critical Review.

Orcid

Ebru Doğruöz https://orcid.org/0000-0001-6572-274X
Hülya Kelecioğlu https://orcid.org/0000-0002-0741-9934

REFERENCES

American Institute of Certified Public Accountants. (2019, February 18). *CPA exam structure*. https://www.aicpa.org/becomeacpa/cpaexam/examinationcontent.html

Belov, D.I. (2016). *Review of modern methods for automated test assembly and item pool analysis*. Law School Admission Council Research Report 16-01 March 2016, LSAC Research Report Series, 23 pages, https://www.lsac.org/docs/default-source/research-(lsac-resources)/rr-16-01.pdf

Breithaupt, K., Ariel, A., & Veldkamp, B. (2005). Automated simultaneous assembly for multistage testing. *International Journal of Testing, 5(*3), 319-330. https://doi.org/10.1207/s15327574ijt05038

Breithaupt, K., & Hare, D.R. (2007). Automated simultaneous assembly of multistage testlets for a high-stakes licensing examination. *Educational and Psychological Measurement*, *67*(1), 5-20. https://doi.org/10.1177/0013164406288162

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. CBS College Publishing.

Dallas, A. (2014). *The effects of routing and scoring within a computer adaptive multi-stage framework* [Unpublished Doctoral Dissertation]. The University of North Carolina.

Davis, L.L., & Dodd, B.G. (2003). Item exposure constraints for testlets in the verbal reasoning section of the MCAT. *Applied Psychological Measurement*, *27*(5), 335-356. https://doi.org/10.1177/0146621603256804

Educational Testing Service. (2018, February 18). *Computer-delivered GRE general test content and structure.* http://www.ets.org/gre/revised%5Cgeneral/about/content/computer/

Hambleton, R.K., & Xing, D. (2006). Optimal and nonoptimal computer-based test designs for making pass-fail decisions. *Applied Measurement in Education, 19*(3), 221-239. https://doi.org/10.1207/s15324818ame1903_4

Hembry, I.F. (2014). *Operational characteristics of mixed format multistage tests using the 3PL testlet response theory model* [Unpublished Doctoral Dissertation]. University of Texas at Austin.

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice, 26*(2), 44-52. https://doi.org/10.1111/j.1745-3992.2007.00093.x

Hogan, J., Thornton, N., Diaz-Hoffmann, L., Mohadjer, L., Krenzke, T., Li, J. & Khorramdel, L. (2016, July 5,). US program for the international assessment of adult competencies (PIAAC) 2012/2014: Main study and national supplement technical report (NCES 2016-036REV). U.S. Department of Education. National Center for Education Statistics. https://nces.ed.gov/pubs2016/2016036 rev.pdf

Jodoin, M.G., Zenisky, A., & Hambleton, R.K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education, 19*(3), 203-220. http://doi.org/10.1207/s15324818ame1903_3

Khorramdel, L., Pokropek, A., & van Rijn, P. (2020). Special Topic: Establishing comparability and measurement invariance in large-scale assessments, part I. *Psychological Test and Assessment Modeling*, *62*(1), 3-10. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam-2020-1/01_Khorramdel.pdf

Kim, S., Moses, T., & You, H. (2015). A comparison of IRT proficiency estimation methods under adaptive multistage testing. *Journal of Educational Measurement*, *52*(1), 70-79. https://doi.org/10.1111/jedm.12063

Kim, J., Chung, H., Dodd, B.G., & Park, R. (2012). Panel design variations in the multistage test using the mixed-format tests. *Educational and Psychological Measurement*, *72*(4), 574-588. https://doi.org/10.1177/0013164411428977

Kirsch, I., & Lennon, M.L. (2017). PIAAC: A new design for a new era. *Large-scale Assessments in Education*, *5*, 11. https://doi.org/10.1186/s40536-017-0046-6

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Lawrence Erlbaum Associates, Inc.

Luecht, R. (2000). *Implementing the CAST framework to mass produce high quality computer adaptive and mastery tests.* Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), New Orleans, LA.

Luecht, R.M. (2006). *Designing tests for pass-fail decisions using item response theory.* In S. Downing & T. Haladyna (Eds.), Handbook of test development, 575-596. Lawrence Erlbaum Associates.

Luecht, R., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education*, *19*(3), 189-202. https://doi.org/10.1207/s15324818ame1903_2

Luecht, R.M., & Nungester, R.J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, *35*(3), 229-249. https://www.learntechlib.org/p/87698/.

Luo, X. (2019). Automated test assembly with mixed-ınteger programming: The effects of modeling approaches and solvers. *Journal of Educational Measurement*, *57*(4), 547-565. https://doi.org/10.1111/jedm.12262

Luo, X., & Kim, D. (2018). A top-down approach to designing the computerized adaptive multistage test. *Journal of Educational Measurement*, *55*(2), 243-263. https://doi.org/10.1111/jedm.12174

Lynn Chen, L.Y. (2010). *An investigation of the optimal test design for multi-stage test using the generalized partial credit model* [Unpublished Doctoral Dissertation]. The University of Texas at Austin.

OECD (2015). *PISA 2015 technical report.* http://www.oecd.org/pisa/sitedocument/PISA-2015-Technical-Report-Chapter-1-Programme-for-International-Student-Assessment-an-Overview.pdf

OECD (2017). *PISA 2015 technical report.* http://www.oecd.org/pisa/sitedocument/PISA-2015-Technical-Report-Chapter-9-Scaling-PISA-Data.pdf

Papadimitriou, C.H., & Steiglitz, K. (1982). *Combinatorial optimization: Algorithms and complexity*. Prentice-Hall.

Park R. (2015). *Investigating the impact of a mixed-format item pool on optimal test designs for multistage testing* [Unpublished doctoral dissertation]. University of Texas, Austin.

Patsula, L.N. (1999). *A comparison of computerized-adaptive testing and multi-stage testing* [Unpublished doctoral dissertation]. University of Massachusetts at Amherst.

Pihlainen, K.A.I., Santtila, M., Häkkinen, K., & Kyröläinen, H. (2018). Associations of physical fitness and body composition characteristics with simulated military task performance. *The Journal of Strength & Conditioning Research*, *32*(4), 1089-1098. https://doi.org/10.1519/jsc.0000000000001921

Sari, H.İ. (2016). *Examining content control in adaptive tests: Computerized adaptive testing vs. computerized multistage testing* [Unpublished doctoral dissertation]. University of Florida.

Sari, H.I., & Raborn, A. (2018). What information works best? A comparison of routing methods. *Applied psychological measurement*, *42*(6), 499-515. https://doi.org/10.1177/0146621617752990

Şahin Kürşad, M., Çokluk-bökeoglu, Ö. & Çıkrıkçı, N. (2022). The study of the effect of item parameter drift on ability estimation obtained from adaptive testing under different conditions. *International Journal of Assessment Tools in Education*, *9*(3), 654-681. https://doi.org/10.21449/ijate.1070848

Theunissen, T.J.J.M. (1985). Binary programming and test design. *Psychometrika*, *50*(4), 411-420. https://link.springer.com/article/10.1007/BF02296260

Tian, C. (2018). Comparison of four stopping rules in computerized adaptive testing and examination of their application to on-the-fly multistage testing [Unpublished master dissertation]. University of Illinois.

Van der Linden, W.J., & Glas, C.A.W. (2000). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education*, *13*, 35-53. https://doi.org/10.1207/s15324818ame1301_2

van der Linden, W.J. (2005). *Linear models of optimal test design.* Springer.

van der Linden, W.J., & Boekkooi-Timminga, E. (1989). A maximin model for IRT-based test design with practical constraints. *Psychometrika*, *54*(2), 237-247. https://link.springer.com/article/10.1007/BF02294518

Veldkamp, B.P. (1999). Multiple-objective test assembly problems. *Journal of Educational Measurement, 36*, 253-66. http://www.jstor.org/stable/1435157

Veldkamp, B.P., Matteucci, M., & de Jong, M.G. (2013). Uncertainties in the item parameter estimates and robust automated test assembly. *Applied Psychological Measurement*, *37*, 123-139. https://doi.org/10.1177/0146621612469825

Wang, K. (2017). *Fair comparison of the performance of computerized adaptive testing and multistage adaptive testing* [Unpublished doctoral dissertation]. Michigan State University.

Wise, S.L., & Kingsbury, G.G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicologica*, *21*, 135-155. https://www.uv.es/revispsi/articulos1y2.00/wise.pdf

Xiao, J., & Bulut, O. (2022). Item selection with collaborative filtering in on-the-fly multistage adaptive testing. *Applied Psychological Measurement*, *46*(8), 690-704. https://doi.org/10.1177/01466216221124089

Xing, D., & Hambleton, R.K. (2004). Impact of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurement*, *64*, 5-21. https://doi.org/10.1177/0013164403258393

Xu, L., Wang, S., Cai, Y., & Tu, D. (2021). The automated test assembly and routing rule for multistage adaptive testing with multidimensional item response theory. *Journal of Educational Measurement*, *58*, 538-563. https://doi.org/10.1111/jedm.12305

Yan, D., Lewis, C., & von Davier, A. (2014). Overview of computerized multistage tests. In D. Yan, A.A. von Davier, & C. Lewis (Eds.). *Computerized Multistage Testing: Theory and Applications*, 3-20. Chapman & Hall.

Yan, D., von Davier, A.A., & Lewis, C. (Eds.). (2014). *Computerized Multistage Testing: Theory and Applications (1st ed.)*. Chapman and Hall/CRC. https://doi.org/10.1201/b16858

Yang, L. (2016). *Enhancing item pool utilization when designing multistage computerized adaptive tests* [Unpublished doctoral dissertation]. Michigan State University.

Zenisky, A. (2004). *Evaluating the effects of several multistage testing design variables on selected psychometric outcomes for certification and licensure assessment* [Unpublished doctoral dissertation]. University of Massachusetts at Amherst.

Zenisky, A., & Hambleton, R. (2014). Multistage test designs: Moving research results into practice. In Yan, D., Von Davier, A., & Lewis, C. (Eds.), *Computerized Multistage Testing: Theory and Applications*, 21-36. Chapman & Hall.

Zenisky, A., Hambleton, R.K. & Luecht, R.M. (2010). Multistage testing: Issues, designs and research. In: der Linden, W.J. & Glas, C.A.W. (Eds.). *Elements of Adaptive Testing*. 355-372. Springer.

Zenisky, A.L., Sireci, S.G., Martone, A., Baldwin, P., & Lam, W. (2009). Massachusetts adult proficiency tests technical manual supplement: 2008-2009. *Center for Educational Assessment Research*. http://www.umass.edu/remp/docs/MAPTTMSupp7-09 final.pdf

Zheng, Y. (2014). *New methods of online calibration for item bank replenishment* [Unpublished Doctoral Dissertation]. University of Illinois at Urbana-Champaign.

Zheng, Y., & Chang, H.-H. (2015). On-the-Fly assembled multistage adaptive testing. *Applied Psychological Measurement*, *39*(2), 104-118. https://doi.org/10.1177/0146621614544519

Zheng, Y., Nozawa, Y., Gao, X., & Chang, H. (2012). *Multistage adaptive testing for a large-scale classification test: the designs, heuristic assembly, and comparison with other testing modes*. Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME) (ACT Research Reports 2012-6). Vancouver, British Columbia, Canada.

Zheng, Y., Nozawa, Y., Zhu, R., & Gao, X. (2016). Automated top-down heuristic assembly of a classification multistage test. *Int. J. Quantitative Research in Education*, *3*(4), 242-265. https://doi.org/10.1504/IJQRE.2016.082387

Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*. [Computer software]. Scientific Software International.

International Journal of Assessment Tools in Education
2024, Vol. 11, No. 2, 288–302
https://doi.org/10.21449/ijate.1353393
journal homepage: https://dergipark.org.tr/en/pub/ijate
Research Article

Developing a quality assessment model (QAM) using logical prediction: Binary validation

Sameer Mohammed Majed Dandan [1*], Odai Falah Mohammad AL-Ghaswyneh [2]

[1]Northern Border University, Faculty of Business Administration, Department of Information Systems Management, Box: 1321, Arar, P.O. 91431 Saudi Arabia
[2]Northern Border University, Faculty of Business Administration, Department of Marketing, Box: 1321, Arar, P.O. 91431 Saudi Arabia

ARTICLE HISTORY
Received: Aug. 31, 2023
Accepted: Mar. 22, 2024

Keywords:
Assessment,
Binary system,
Competencies transfer,
Prediction,
Quality,
Boolean.

Abstract: This study focuses on evaluating the quality of competency transfer through various assessment methods and results, considering diverse stakeholder perspectives. The research aims to introduce an innovative approach for validating assessment outcomes, leveraging predicted sub-measurements, and transforming Boolean parameters' symbols into a binary coding system. This transformation simplifies the validation process by employing logical equations. The study's sample involves the adaptation of a competency transfer model, which combines internal parameters with the novel logical assessment method. The research findings indicate that the binary 2^x system effectively simplifies quantitative and qualitative data representation within the validation process. This system facilitates the early detection of potentially ambiguous results, enabling the creation of validation procedures grounded in organizational cultural dimensions, outcomes, reports, and assessments. The proposed Quality Assessment Model (QAM) serves as a powerful tool for prediction, enhancing the quality of both quantitative and qualitative data outcomes. This approach generates distinct values, precise predictive measurements, and valuable result quality suitable for informed decision-making in various contexts. Ultimately, the study contributes to the advancement of assessment methodologies, enabling stakeholders to make more accurate and reliable judgments based on the quality of competency transfer.

1. INTRODUCTION

In the dynamic realm of quality assessment, the pursuit of more precise and streamlined methods remains a constant endeavor. As the realm of technology continually reshapes our environment, the demand for assessment techniques that resonate with the intricacies of contemporary data and systems intensifies. In light of this backdrop, the call for a novel quality assessment approach emerges-one that leverages the potency of logical prediction rooted in binary validation. While traditional quality assessment techniques have proven their value, they often encounter challenges in effectively encompassing the intricacies of intricate data and systems. The emergence of binary validation, accompanied by the integration of predictive

*CONTACT: Sameer Mohammed Majed Dandan ✉ samotoom@hotmail.com Northern Border University University, Faculty of Business Administration, Department of Information Systems Management Box: 1321, Arar, P.O. 91431 Saudi Arabia, +962790209524, +966541558544

e-ISSN: 2148-7456

analytics, presents a promising avenue for surmounting these limitations. This innovative strategy not only holds the potential to elevate precision but also offers the prospect of streamlining the assessment process, thus conserving invaluable time and resources. This introduction will delve into the rationale behind the pursuit of a new quality assessment method, exploring the shortcomings of existing approaches and highlighting the potential advantages of a logical prediction-based framework. By delving into the intricacies of this topic, we aim to shed light on the critical role that such a method could play in a variety of fields, from software development to data analysis and beyond. As we embark on this exploration, we invite you to join us in considering the compelling case for embracing a new era of quality assessment through logical prediction and binary validation based on an algorithm that uses questionnaires at the same time. Algorithms or questionnaires are different tools to collect data and produce results. Both questionnaires and algorithms serve distinct purposes and have their advantages depending on the context. Meanwhile, questionnaires are a method of data collection that involves presenting a series of questions to individuals or groups to gather information. They are commonly used in surveys, research studies, and assessments. Questionnaires can be useful for collecting qualitative and quantitative data directly from participants. They offer the following advantages: rich data, flexibility, subjective information, and exploration. However, questionnaires also have limitations, such as being biased, consuming time and effort, and limited sample size. On the other hand, algorithms are sets of rules or instructions designed to solve specific problems or perform tasks. In the context of data analysis and decision-making, algorithms can automate processes, identify patterns, and make predictions based on data. They have several advantages, such as speed in processing and analyzing large datasets quickly, making them efficient for tasks that involve data crunching. Algorithms are also described with consistency in providing consistent results across different instances. They considered a complex pattern that can identify intricate patterns and relationships within data that may be difficult for humans to discern, and scalability to be applied to a wide range of data without much additional effort. The algorithms, however, have come with challenges that lie under the data quality, interpretability, results that are difficult to interpret or explain or a lack of context in understanding nuanced or contextual information. In conclusion, the choice between questionnaires and algorithms depends on the goals and requirements of the specific task. Questionnaires are valuable for gathering detailed qualitative data and capturing subjective experiences, while algorithms excel at processing large datasets and automating decision-making processes. Often, a combination of both approaches can provide a more comprehensive understanding and effective solutions.

1.1. Literature Review

The dimensions of organizational culture exhibit variations based on internal and external activities, as well as outcomes associated with processes, services, and products (Dandan, 2017). Furthermore, the assessment outcomes stemming from these activities demonstrate divergence due to the specific assessment type, methodologies employed, sample sizes, assessment dates, and underlying objectives (Alas et al., 2015; Schwartz, 1994). These assessment tools culminate in definitive results (Göckede et al., 2004; Graymore et al., 2008; Hawthorne et al., 2016) . In this context, Thireau (2002) emphasizes that each result holds valuable significance and meaning. These collective aspects coalesce to represent quality (Mitra, 2016; Shewfelt, 1999). Additionally, a quality validation model is utilized to ensure the coherence of results and facilitate comparisons between multiple stakeholders within the model. It may also involve comparisons with internal or external audits, as well as evaluation reports from public and/or private agencies (Arnold et al., 2012; Dias et al., 2014; East et al., 2016; Fox, 1981; Grönroos, 1984; Jabangwe et al., 2015; Pinson et al., 2013; Wittenberg et al., 2016). The process of competency transfer has been subject to examination over several decades to assess both individual and group competitiveness. This intricate process involves various participants (Brandt & Dimmitt, 2015; Gutierrez Gutierrez et al., 2016; Koskinen & Pihlanto, 2006). These authors bring diverse perspectives to the table, expressed through quantitative methods such as questionnaire responses and involvement in designing or taking exams, and it is essential to enhance these processes by predicting thinking patterns and neural activities, as highlighted by Fayaz et al. (2018). Furthermore, the landscape of assessments has evolved and remains variable due to the influence of numerous factors (McCallin & McCallin, 2009). Satisfaction levels serve as an example of an assessment approach, encompassing tacit knowledge that becomes formalized once combined with explicit knowledge. The measurement of tacit knowledge involves quantitative techniques to extract data and subsequently present statistics about the percentage, level, limitations, types, and values within each data segment (Nonaka & Konno, 1998; Nonaka & Teece, 2001; Purdy et al., 2018). It is important to note that satisfaction is influenced by perceptions and is affected by emotional, organizational, contextual, and policy-related elements inherent in the tested data, collectively constituting the organizational environment. Assessments are inherently influenced by actors' varying levels of satisfaction and perspectives on particular issues. Consequently, it becomes imperative to establish an evaluative model that accommodates these differing viewpoints. While numerous studies have employed diverse assessment approaches, these still encapsulate specific viewpoints. Recognizing this, the current study endeavors to introduce a novel methodology for validating result quality. This involves the utilization of triangulation procedures, as proposed by Guion (2002). Furthermore, the study incorporates sub-assessments to examine the coherence of responses. The approach draws inspiration from the conversion of actors and results into a binary system (Boole, 1854). This simplifies the assessment technique, aligning with multi-factor modeling series principles akin to those presented by Li and Yu (2020). Notably, this study stands as an original endeavor, setting it apart from previous research. It employs the competency transference model outlined by Dandan (2017) as a practical example of implementation. The suggested Quality Assessment Model (QAM) is a novel concept, while the prior studies lacking on presenting a predicting process to validate the data used for assessment of results in any research using binary system. It draws inspiration from the triangulation method, which validates data accuracy from various sources. Using an example model for evaluating competency transfer data among schools, graduates, and employers initiated by Dandan (2017), validation is conducted from three different perspectives within the model. Each part of the model offers a unique viewpoint on similar questions. The expected results are transformed into binary values (0, 1) representing true and false. These values facilitate usage of the arithmetic logical operator the “AND” gate or specific formulas to assess expected results. If the results are true among the three parties, the questions or survey results are accepted for further analysis. Otherwise, an alternative formulaic analysis is employed.

1.1.1. *Quality assessment using logical prediction algorithms*

In a study by Sharma et al. (2021), they examined a comprehensive survey that reviews various machine learning techniques employed in quality assessment. It covers traditional methods as well as emerging approaches such as logical prediction algorithms. The paper discusses the advantages of logical prediction in improving assessment accuracy and provides insights into its application across different domains. Meanwhile, Alas et al. (2015) focused specifically on logical predictive modeling; this paper explores the integration of logical reasoning into quality assessment processes. The authors present a novel algorithm that combines binary validation with logical inference to enhance assessment outcomes. Real-world case studies illustrate the effectiveness of this approach in areas like software testing and anomaly detection. In addition, a review paper by Burggräf et al. (2021) examined the applications of predictive analytics in quality assurance. It discusses the role of logical prediction algorithms in identifying potential defects or anomalies before they impact the system. The authors emphasize the importance of accurate prediction models for ensuring high-quality products and services. Earlier, Singh et al. (2017) focused on the data science domain; this study investigates the integration of logical inference techniques in quality assessment processes. The authors present a framework that leverages logical prediction algorithms to identify data inconsistencies, leading to improved data quality and more reliable analysis outcomes. (Jafarian et al., 2020) paper explored the application of logical prediction algorithms in software testing. It discusses how binary validation and logical reasoning can be used to identify anomalies and potential defects early in the development lifecycle. The authors highlight the benefits of this approach in reducing debugging efforts and improving software reliability. Prediction is also used in many health fields as protein detection, and a comparative study by Chen and Siu (2020) evaluated different machine learning techniques for quality assessment, including logical prediction algorithms. The authors compare the performance of logical prediction-based methods with traditional approaches and discuss the advantages of using logical inference in enhancing assessment accuracy.

1.1.2. *Quality assessment using logical prediction algorithms as binary code*

Hranisavljevic et al. (2020) delved into the practical application of logical prediction algorithms; this study focuses on anomaly detection in binary code. The authors propose a method that leverages logical inference to detect unusual patterns and behaviors in executable files. Real-world case studies demonstrate the effectiveness of the approach in identifying malicious code and software vulnerabilities. Later on, Tian et al. (2021) published a paper that introduced the concept of utilizing logical prediction algorithms for assessing the quality of binary code. It outlines the challenges associated with traditional methods and presents a framework that combines binary validation and logical reasoning to enhance the accuracy of identifying defects and vulnerabilities in compiled software. Meanwhile, Wang's (2023) paper explored the integration of logical prediction algorithms in the context of embedded systems. It discusses how logical reasoning can be used to assess the quality and reliability of binary code running on resource-constrained devices. The authors provide insights into the benefits of this approach in ensuring the robustness of embedded software. Being in the same marathon of developing novel prediction techniques, Zhang (2023) focused on code analysis; this research investigates the role of logical inference in improving the accuracy of identifying code defects and vulnerabilities. The authors propose a method that combines static analysis with logical prediction algorithms to achieve more reliable results. The study showcases the effectiveness of this approach in various software security scenarios. Croft et al. (2023) addressed the quality assessment of compiled software; this study presents a systematic approach that employs logical prediction algorithms. The authors highlight how binary validation and logical reasoning can be used to uncover hidden code flaws that may evade traditional analysis techniques. The paper emphasizes the importance of incorporating logical inference in modern software quality assurance practices. This is not so far from the comparative study by Bride et al. (2021) that evaluated the effectiveness of machine learning techniques, including logical prediction algorithms, for verifying the correctness and reliability of binary code. The authors analyze the performance of different methods in identifying bugs and vulnerabilities, shedding light on the advantages of logical inference in this context.

1.1.3. *Quality assessments using modelled algorithms*

Baqais and Alshayeb (2020) explored a systematic review of a comparative study that examined the efficiency and accuracy of automated quality assessment algorithms for software code, comparing them with manual reviews. It discusses the benefits of algorithms in terms of scalability, consistency, and reduced human bias. The paper also addresses challenges, such as algorithmic limitations in detecting certain code quality issues. In addition, Cetiner and Sahingoz (2020) sought to examine predictive algorithms by comparing their performance in quality assessment across various domains. The authors discuss the advantages of algorithms in predicting potential quality issues before they manifest, leading to proactive problem-solving. They also explore the need for continuous refinement of algorithms to adapt to evolving quality standards. This earlier was obtained in a study by Marchisio et al. (2018) that presented an approach that leverages algorithms to analyze user feedback and extract meaningful insights for quality assessment. It discusses how algorithms can identify patterns and trends in large datasets, offering a data-driven perspective on quality. The paper emphasizes the benefits of algorithmic analysis in processing and interpreting vast amounts of user-generated content.

1.1.4. *Competency transfer process in business school*

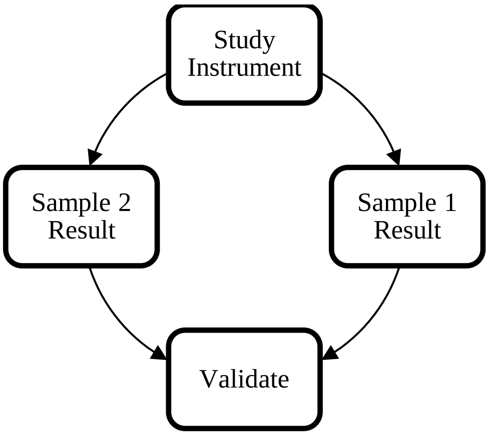
A comprehensive review paper examined the evolution of competency-based education (CBE) in business schools and explored how CBE aligns with the demands of the modern workforce and the changing nature of business. The authors discuss how CBE frameworks enable the transfer of relevant competencies to students, preparing them for real-world challenges. (Silitonga, 2021). Before that, Bratianu et al. (2020) suggested a design aspect to analyze various competency transfer models implemented in business schools. They reviewed how these models integrate theoretical knowledge with practical skills, emphasizing experiential learning and industry collaboration. The paper highlights the benefits of well-structured competency transfer processes in producing job-ready graduates. Meanwhile, Alnasib (2023) investigates the role of (DigComp) as a digital tool and technology in enhancing competency transfer within Teacher-business education. It reviews the utilization of online platforms, simulations, and virtual environments to simulate real-world scenarios. The authors explore how technology-driven learning experiences prepare teachers to meet students' demands in dynamic business environments. Moreover, Wohlfart et al. (2022) selected a CBT of industry relevance; this case study examines how business schools align their curriculum with industry demands. It discusses how competency transfer processes can bridge the gap between academic knowledge and practical skills. The paper showcases examples of collaborations between business schools and corporations to ensure graduates possess the required competencies.

2. METHOD

2.1. Improvement of Assessment

Mostly, it is known that to evaluate results between two different perspectives, you need to find a comparative tool (Shi, 2013). The following model expresses a sample of evaluation actions between different results (Figure 1).

Figure 1. *Validation of two different results.*



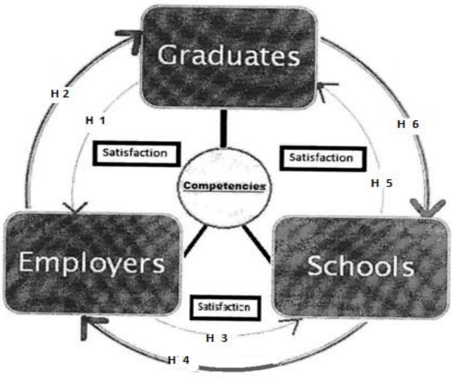
For example, if you use a questionnaire to ask two different samples about their perspective of a determined issue, you need a third sample that is qualified, expert, or similar to previous samples’ cultural dimensions to justify results. The Validation of Assessments' Results (VAR) expressed with the following formula:

$$(\boldsymbol{VAR})^{\boldsymbol{n}} = if \begin{cases} R\,n0 \quad \text{AND} \quad R\,n1 \;=\; VAR\,n\,(0\;AND\;1) & ,\; Substitute\; Hypothesis\; Accepted \\ R\,n0 \quad \text{AND} \quad R\,n1 \;\neq VAR\,n\,(0\;AND\;1) & ,\; Validation\; Procedure\; Accepted \\ R\,n0 \quad \text{AND} \quad R\,n1 = \;VAR\,n\,(0\;AND\;1) & ,Hypothesis\; Accepted \end{cases}$$

2.2. Participants

In this case, the Competency transference model (Dandan, 2017) as shown in Figure 2, is used as a case study to examine the validation assessment method.

Figure 2. *Competency transfer model (Dandan, 2017).*



2.3. Measurement

The study used a systematic review of assessment methodologies and techniques and abbreviated this result as an assessment result with ASSR. Each actor has a symbol. The actors and relations in our case are:

1- E: Employer.
2- S: School.
3- G: Graduate.
4- H_1, H_2,……H_6.: The hypothesis of relationships based on the level of satisfaction.

2.4. Proceedings

Validation of stakeholders' perspectives are based on organizational cultural dimensions, reports, outcomes, and at the same time, sub-assessments of these parts (Dami et al., 2018; Evans et al., 2018). These sub-assessments are engaged by or under independent authorities of evaluations and monitoring inside or outside the organization. Also, to simplify the validation process, the study assumed symbolic equations that were used based on the binary system to draw the validation map. The expected results are calculated based on the number of assessments, and at the same time, the number of assessments is calculated based on the number of actors as defined in a binary system in Table 1.

Table 1. *Binary table of actors, assessments, and results.*

No of Actors	Formula of Assessments	Expected Results
1	21	2
2	22	4
3	23	8
4	24	16
5	25	32
6	26	64

2.5. Model Assessment Validation Procedures

It is imperative to establish the significance of validation rules and procedures before initiating the assessment process to construct a validation framework grounded in arithmetic, logical, or algorithmic systems. This validation process can be effectively executed through the integration of sub-assessments (Woods, 2018). Particularly, organizational culture and activities are deemed as optimal avenues for quantitatively measuring data. Within the scope of this study, the following validation procedures have been posited:

- Conformity Check: Assessing the alignment of collected data with predefined validation rules. This step ensures that data adhere to expected standards. During this step, the collected data of any questionnaire must be aligned with the domain of the study, applicable for evaluation, and correlated between parties of the sample to help prediction succeed. i.e., questions of collecting data between social studies will not be accepted to predict the agricultural assessment studies.
- Consistency Examination: Scrutinizing data for logical coherence and internal consistency. This procedure ensures that data points within the assessment are harmonious. Here, any questionnaire's data will be tested by QAM model and must be evaluated early as biometric and consistency tests such as alpha Cronbach if they are available.
- Cross-validation: Employing multiple sources or approaches to validate data accuracy. This approach enhances the reliability of the collected information. As mentioned in previous steps, data accuracy is mandatory.
- Triangulation Verification: Utilizing multiple data collection methods to corroborate findings. This technique increases confidence in the accuracy of the gathered data. Triangulation of using many approaches will lead to selecting suitable evaluation methods for one type of data. This will give mirror results if the data are accurate and the results are correct. Meanwhile, QAM using prediction techniques will be a successful method to evaluate results earlier.
- External Validation: Comparing collected data with external benchmarks or reference sources to affirm its accuracy and validity. Meanwhile, statistical reports of different stakeholders and evaluation reports of many external parties are available, which will help compare the results from different perspectives. Therefore, this stage will assist to ensuring that QAM is suitable if the prediction is adequate with results.
- Expert Review: Involving subject-matter experts to review and validate the data collected, enhancing its credibility and quality. This point lies under the previous one, where experts and professionals help to examine the accuracy of methods and results.
- Time-Series Analysis: Examining data trends over time to ensure consistency and detect any anomalies or deviations. Time matters, and perspectives differ so prediction must be conducted within an acceptable time interval.
- Contextual Relevance Assessment: Evaluating the contextual relevance of collected data to ensure it accurately represents the intended information. In qualitative studies, the collected data must be correct to assess pure information that is accredited to present results accurately.

By instituting these validation procedures, the study aims to create a robust validation map that ensures the accuracy, consistency, and reliability of the assessment outcomes.

2.6. Validation of Graduate Perspective and Data Consistency

A comprehensive validation approach is proposed to ascertain the validity of graduate perspectives and ensure the consistency of collected data. This approach involves the integration of diverse assessments, including annual academic or national examinations, which assess skills and learning outcomes. This comprehensive strategy incorporates various measurement techniques, such as evaluating individual performance, gauging responses in both individual and group work settings, engaging in debates and discussions, conducting workshops, analyzing case studies, administering examinations, documenting students' scientific research achievements, assessing innovative products and inventions, as well as evaluating entrepreneurial endeavors. This amalgamation of assessment procedures forms a holistic framework termed "Validate Graduate Competencies and Satisfaction" (VG-CS), ensuring a thorough and accurate validation process for graduate competencies and satisfaction levels.

2.7. Validation of School Perspectives

To validate the perspectives held by educational institutions, an encompassing approach is proposed, focusing on the capability of academic staff to transfer knowledge and skills effectively. This validation process entails a thorough examination of essential components such as a robust curriculum, effective teaching methods, well-equipped infrastructures, adherence to legislations and policies, comprehensive plans, and the availability of sufficient funds. The impact of these factors on graduates' achievements is a pivotal aspect of this assessment. This validation process hinges on the utilization of key indicators to measure the effectiveness of the school environment. These indicators encompass annual reports that highlight tangible achievements by graduates, including scholarships earned, successful project funding, contributions to scientific research, and various accomplishments. Additionally, the assessment includes the perspective of students, gauging the quality of education from their standpoint through annual assessments. Moreover, corporate social responsibilities are considered, assessing the interaction between the school, employers, and the broader community. This engagement is further evidenced by records of training courses, job preparation programs, and social initiatives. This holistic approach, referred to as "Validate School Competencies and Satisfaction" (VS-CS), ensures the comprehensive evaluation of the school's abilities and the overall satisfaction of stakeholders, aligning the institution's efforts with the broader goals of education and skill development.

2.8. Validation of Employer Perspectives and Activities

The validation process for employers is facilitated by adherence to disclosure policies, enhancing the ease of assessment. This validation is achieved through the examination of various key indicators that provide insight into the engagement between employers and graduates. One of the core validation indicators is the analysis of the number of employed graduates. This data offers a tangible measure of the effectiveness of the educational institution in preparing students for the job market. Additionally, the number and nature of job advertisements by employers serve as valuable evidence of their engagement in the recruitment of graduates. A crucial aspect of this validation process involves corporate social responsibilities. These responsibilities are assessed based on the extent to which employers collaborate with educational institutions and graduates. This collaboration may encompass initiatives such as funding research and projects, actively participating in educational endeavors, and engaging in annual meetings with decision-makers to discuss career opportunities and growth prospects. Moreover, the organization of events like Job days and the establishment of memorandums of training and recruitment further underline employer engagement. This comprehensive validation approach, referred to as "Validate Employer Activities and Satisfaction" (VE-AS), ensures a thorough evaluation of employers' activities and their satisfaction with the quality of graduates entering the workforce. By aligning employers' perspectives with educational goals, this approach enhances the employability of graduates and reinforces the relationship between educational institutions and the professional world.

2.9. Validation Using Binary Equations

In this section, we have gathered the anticipated assessment outcomes from various approaches within the university as well as from employers. The objective is to delineate the framework of a binary test that encapsulates the interactions among the three stakeholders. For this purpose, we have employed the AND gate, which is characterized as follows:

$$x \wedge y = x \ \times y = \min\,(x, y)$$

- AND, present an expression of $x \wedge y$, while $x \wedge y = 1$ if $x = y = 1$ and $x \wedge y = 0$ otherwise.

2.10. Validation of Model Assessment Results

Utilizing the binary representation of the QAM (Quality Assessment Model) actors and their corresponding assessment outcomes, the study has established the validation map as outlined in Table 2. This comprehensive map not only encompasses the primary assessments but also incorporates recommended sub-assessments strategically designed to verify the accuracy and coherence of the major assessment procedures. Within the framework of this study, all hypotheses (H1, H2, … H6) have been considered pre-approved by default, under the presumption of a high level of satisfaction among all stakeholders, particularly within higher education (Vassiliadis & Schwarz, 1990). To empirically evaluate this assumption, the study proposes the utilization of questionnaires as a data collection instrument. These questionnaires are designed to elicit responses from the three key actors, capturing their distinct perspectives regarding the extent of their satisfaction. Through this methodological approach, the study seeks to systematically gather and analyze data, offering insights into the level of satisfaction within the relationships between the stakeholders.

This assessment aligns with the broader objective of the study, which is to substantiate the presumed high satisfaction levels and validate the proposed hypotheses within the context of higher education. See Table 2.

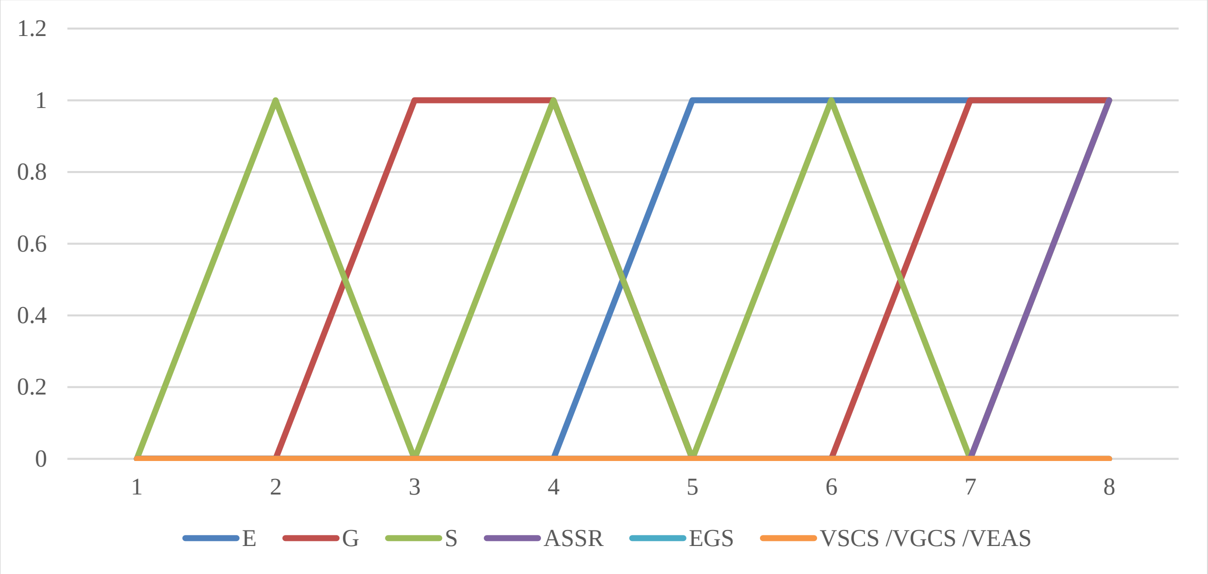
Table 2. *Validation map.*

Validation map of Competency Transfer Model					
AND Operation for all Operands (E, G, S)			Binary expression of assessment result	The opposite actor	Quality Validation Procedures using Suggested Sub- Assessment
Employer	Graduate	School			
E	G	S	ASSR	EGS	VSCS /VGCS /VEAS
0	0	0	0	----	Substitute hypothesis
0	0	1	0	S	VSCS
0	1	0	0	G	VGCS
0	1	1	0	E	VEAS
1	0	0	0	E	VEAS
1	0	1	0	G	VGCS
1	1	0	0	S	VSCS
1	1	1	1	-----	Hypothesis accepted

Within the Boolean AND gate framework, two significant outcomes emerge. The first outcome pertains to a scenario in which all stakeholders express dissatisfaction. In this case, the study recommends resorting to substitute hypotheses due to the rejection of the initially proposed hypotheses. Conversely, if all stakeholders express satisfaction, the proposed hypotheses are accepted.

As illustrated in Figure 3, the school consented to confer degrees on two occasions. The initial instance pertains to the endorsement from employers, signifying the school's confidence in the academic accomplishments' ability to effectively convey skills to graduating individuals who also validate this process. The subsequent instance involves the mutual agreement between the school and employers, ensuring satisfaction on both ends. Here, the school is successful in facilitating the optimal transfer of competencies, enabling graduates to acquire skills in alignment with the perspective of employers. See Figure 3.

Figure 3. *Validation map of competency transfer model.*



As shown in Figure 3, each cross point is considered as a probability result of prediction that reflects the data of Table 2. The first party 'S' of the school recorded four times on the line on 1, 3, 5, and 7 cross points of zero value, respectively, and value '1' on points 2, 4m 6 and synchronized with assessment result on point 8 to express that the hypothesis accepted. According to the table, the party 'G' of the graduate recorded a twice hit on a zero on cross points 2 and 6, and one on both trial cross points '3-4' and '6-8'. In addition, the third actor here, party 'E' the employer, confirmed zero value in the first four trials on the cross points 1, 2, 3, and 4 and raised to value '1'to meet both of school 'S' and graduate 'G' on points 5, 6, 7 and all of them "E, S, and G" confirmed value '1 on cross point trial 8 of prediction values. These values of the probability results are similar to the Boolean arithmetic probabilities of (0, 1) between three parties. These results will not be accurate or correct unless all the parts are '1' for each or '0' for all.

The critical role of validation becomes evident when any one of the three primary stakeholders deviates from the consensus level of satisfaction exhibited by the other two. The study designates the distinct stakeholders as represented by the letters E, G, and S, signifying Employer, Graduate, and School, respectively. If any or all of these stakeholders present a conflicting perspective, suggesting either lower or higher levels of satisfaction, it indicates a disparity.

In such instances, the study suggests the utilization of predefined validation procedures, namely VS-CS, VG-CS, and VE-AS, to rigorously assess the precision and coherence of the major assessment quality. Additionally, the selection of these validation procedures is influenced by the unique organizational environment, dimensions, outcomes, reports, and assessments inherent in the case study context.

Through this comprehensive validation approach, the study endeavors to ensure the reliability of the assessment results, effectively accounting for varying perspectives and potential disparities among the stakeholders. This approach aligns with the study's overarching goal of substantiating the hypotheses and validating the assessment framework within the specific organizational context.

3. FINDINGS

The Quality Assessment Validation Model (QAM) has uncovered that the validation procedures, namely VS-CS, VG-CS, and VE-AS, meticulously crafted to align with organizational cultural dimensions, outcomes, reports, and assessments and have played a pivotal role in shaping a virtual sub-assessment approach. This approach serves as a mechanism to scrutinize and ascertain the precision and cohesiveness of major assessment quality. By employing the binary system, the virtual sub-assessment method becomes adept at pinpointing potential weak points and anomalies within the assessment outcomes. It effectively identifies values that might appear inconsistent or suspicious when assessed from differing perspectives on the same matter. Furthermore, the binary system can be harnessed in various ways to enhance or invalidate multiple hypotheses. This approach underscores the versatility of the binary system in contributing to a comprehensive assessment validation process. Through these mechanisms, the model promotes accuracy, reliability, and robustness in assessing major quality evaluations, offering a nuanced and thorough understanding of the assessment outcomes from multiple angles. Moreover, the Quality Assessment Model (QAM) offers several additional benefits beyond its core functionality of validating assessment outcomes. Some of these benefits include:

- Comprehensive Understanding: QAM provides a holistic approach to assessment validation, taking into account various stakeholders' perspectives, organizational cultural dimensions, and contextual factors. This leads to a more comprehensive understanding of assessment quality.
- Enhanced Decision-Making: By identifying weaknesses, anomalies, and potential biases in assessment outcomes, QAM empowers decision-makers to make more informed and accurate decisions based on reliable data.
- Transparent Accountability: QAM promotes transparency and accountability in the assessment process. It allows stakeholders to understand the validation methods employed and the reasoning behind assessment results, fostering trust in the assessment outcomes.
- Continuous Improvement: The binary system and virtual sub-assessment approach of QAM can be used iteratively to identify areas for improvement in assessment methodologies. This supports a cycle of continuous enhancement in assessment practices.
- Effective Resource Allocation: By pinpointing weaknesses and areas of concern, QAM aids in directing resources to the right areas for improvement, optimizing the allocation of time, effort, and budget.
- Adaptability: QAM can be adapted to different contexts, industries, and assessment types. Its flexibility makes it a valuable tool for a wide range of quality assessment scenarios.
- Reduced Bias: QAM's systematic approach minimizes potential biases that may arise from relying solely on one stakeholder's perspective. It offers a balanced view of assessment outcomes.
- Strategic Alignment: QAM ensures that assessment objectives align with broader organizational goals and objectives. This strategic alignment enhances the relevance and impact of assessment results.
- Consistency: The use of predefined validation procedures and the binary system in QAM promotes consistency in assessment evaluation, leading to more reliable and comparable results over time.
- Sustainability: QAM promotes the sustainability of assessment practices by identifying areas of concern early on, allowing for timely adjustments and improvements to maintain assessment quality over the long term.

In summary, the Quality Assessment Validation Model (QAM) offers benefits that go beyond mere validation, providing a framework that enhances decision-making, accountability, transparency, and overall assessment quality while enabling organizations to refine their assessment practices continuously.

4. DISCUSSION and CONCLUSION

Utilizing the binary system to formulate validation procedures for assessing the quality of results derived from both quantitative and qualitative assessments holds significant potential. This approach can yield distinct studies, precise predictive measurements, and valuable outcome quality applicable across various scientific disciplines. Leveraging the competency model proposed by (DANDAN, 2017) enhances the ability to differentiate between hypotheses derived from systematic reviews and those resulting from triangulation methods, thereby facilitating hypothesis testing. The introduced Quality Assessment Model (QAM) effectively illustrates how researchers can strategically anticipate and identify potentially suspicious results. By preemptively devising sub-assessment strategies before these findings emerge, researchers can proactively validate the quality of their results. Additionally, the adoption of the AND gate as an expression of the validation process offers a robust and versatile approach for assessing both qualitative and quantitative values. This approach's applicability spans diverse domains, ranging from computer science to art, health, engineering, and even space science. Moreover, the categorization of validation based on sub-assessment within the organizational environment contributes to the meticulous examination of assessment result accuracy and consistency. This categorization strategy ensures a comprehensive and systematic approach to evaluating the quality of assessment outcomes, enhancing their reliability and applicability across various contexts. In essence, the binary system-driven Quality Assessment Validation Model presents a multidimensional approach that transcends disciplinary boundaries, providing researchers with a systematic and adaptable tool to enhance the accuracy, reliability, and overall quality of assessment results.

For further studies and recommendations based on results, the study recommended exploring additional validation procedures in various scientific disciplines that hold_immense potential for enhancing the quality and reliability of assessment outcomes. This can contribute to the development of standardized methods that can be applied across different domains, leading to more accurate predictions and informed decision-making. Moreover, extending research to different areas of science, such as systems, medicine, the environment, climate, and agriculture, presents exciting opportunities for proactive risk management and strategic planning. By identifying potential failures and risks early on, researchers can develop predictive procedures that aid in minimizing negative impacts and optimizing resource allocation. In summary, pursuing further studies that delve into diverse validation techniques and applying these methods to various scientific fields has the potential to advance both research methodologies and practical applications. It can lead to more robust assessments, better predictions, and improved strategies for managing risks and uncertainties in complex systems and environments.

Acknowledgments

The authors wish to acknowledge the approval and the support of this research study by the grant no: (BSAA-2023-12-2296) from the Deanship of Scientific Research in Northern Border University, Box: 1321, Arar, P.O. 91431 Saudi Arabia.

Declaration of Conflicting Interests and Ethics

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

Contribution of Authors

Sameer Dandan: Review previous studies, design study instrument and analysis implementation. **Odai AL-Ghaswyneh**: Review methodology, examining the data consistency, review results. **Both authors:** 1st draft of the paper, proofreading, final version.

Orcid

Sameer Mohammed Majed Dandan https://orcid.org/0000-0003-0140-312X
Odai Falah AL-Ghaswyneh https://orcid.org/0000-0002-9851-3407

REFERENCES

Alam, S.M.T. (2015). Factors affecting job satisfaction, motivation and turnover rate of medical promotion officer (MPO) in pharmaceutical industry: a study based in Khulna city. *Asian Business Review*, *1*(2), 126-131.

Alas, R., Gao, J., & Carneiro, J. (2015). Connections between ethics and cultural dimensions. *Engineering Economics*, *21*(3).

Alnasib, B.N. (2023). Digital Competencies: Are Pre-Service Teachers Qualified for Digital Education? *International Journal of Education in Mathematics, Science and Technology*, *11*(1), 96-114.

Arnold, J.G., Moriasi, D.N., Gassman, P.W., Abbaspour, K.C., White, M.J., Srinivasan, R., . . . Van Liew, M.W. (2012). SWAT: Model use, calibration, and validation. *Transactions of the ASABE*, *55*(4), 1491-1508.

Baqais, A.A.B., & Alshayeb, M. (2020). Automatic software refactoring: A systematic literature review. *Software Quality Journal*, *28*(2), 459-502.

Boole, G. (1854). *An investigation of the laws of thought: on which are founded the mathematical theories of logic and probabilities*. Dover Publications.

Brandt, C., & Dimmitt, N. (2015). Transfer of learning in the development of peer tutor competence. *Learning and Teaching in Higher Education: Gulf Perspectives*, *12*(2).

Bratianu, C., Hadad, S., & Bejinaru, R. (2020). Paradigm shift in business education: a competence-based approach. *Sustainability*, *12*(4), 1348.

Bride, H., Cai, C.-H., Dong, J., Dong, J.S., Hóu, Z., Mirjalili, S., & Sun, J. (2021). Silas: A high-performance machine learning foundation for logical reasoning and verification. *Expert Systems with Applications*, *176*, 114806.

Burggräf, P., Wagner, J., Heinbach, B., Steinberg, F., Schmallenbach, L., Garcke, J., . . . Wolter, M. (2021). Predictive analytics in quality assurance for assembly processes: Lessons learned from a case study at an industry 4.0 demonstration cell. *Procedia CIRP*, *104*, 641-646.

Cetiner, M., & Sahingoz, O.K. (2020). A comparative analysis for machine learning based software defect prediction systems. 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT).

Chen, J., & Siu, S. W. (2020). Machine learning approaches for quality assessment of protein structures. *Biomolecules*, *10*(4), 626.

Croft, R., Babar, M.A., & Kholoosi, M.M. (2023). Data quality for software vulnerability datasets. IEEE/ACM 45th International Conference on Software Engineering (ICSE).

Dami, S., Barforoush, A.A., & Shirazi, H. (2018). News events prediction using Markov logic networks. *Journal of Information Science*, *44*(1), 91-109.

Dandan, S.M. (2017). *Stakeholder Satisfication with Competencies Transfer in the Framework of Educational Policy Elements* [Book, Faculty of Organisation Studies, Ljubljana]. FOS Novi trg 5, 8000 Novo mesto.

Dias, J.M.P., Oliveira, C.M., & da Silva Cruz, L.A. (2014). Retinal image quality assessment using generic image quality indicators. *Information Fusion*, *19*, 73-90.

East, R., East, R., Uncles, M.D., Uncles, M.D., Romaniuk, J., Romaniuk, J., . . . Lomax, W. (2016). Validation and sufficiency. *European Journal of Marketing*, *50*(3/4), 661-666.

Evans, R., Saxton, D., Amos, D., Kohli, P., & Grefenstette, E. (2018). Can Neural Networks Understand Logical Entailment? *arXiv preprint arXiv:1802.08535*.

Fayaz, M., Ullah, I., & Kim, D.-H. (2018). Underground risk index assessment and prediction using a simplified hierarchical fuzzy logic model and kalman filter. *Processes*, *6*(8), 103.

Fox, D.G. (1981). Judging air quality model performance. *Bulletin of the American Meteorological Society*, *62*(5), 599-609.

Göckede, M., Rebmann, C., & Foken, T. (2004). A combination of quality assessment tools for eddy covariance measurements with footprint modelling for the characterisation of complex sites. *Agricultural and Forest Meteorology*, *127*(3), 175-188.

Graymore, M.L., Sipe, N.G., & Rickson, R.E. (2008). Regional sustainability: How useful are current tools of sustainability assessment at the regional scale?. *Ecological Economics*, *67*(3), 362-372.

Grönroos, C. (1984). A service quality model and its marketing implications. *European Journal of Marketing*, *18*(4), 36-44.

Guion, L.A. (2002). *Triangulation: Establishing the validity of qualitative studies*. University of Florida Cooperative Extension Service, Institute of Food and Agricultural Sciences, EDIS.

Gutierrez Gutierrez, L., Barrales-Molina, V., & Tamayo-Torres, J. (2016). The knowledge transfer process in Six Sigma subsidiary firms. *Total Quality Management & Business Excellence*, *27*(5-6), 613-627.

Hawthorne, G., Saggar, M., Quintin, E.-M., Bott, N., Keinitz, E., Liu, N., . . . Reiss, A.L. (2016). Designing a creativity assessment tool for the twenty-first century: Preliminary Results and insights from developing a design-thinking based assessment of creative capacity. In *Design Thinking Research* (pp. 111-123). Springer.

Hossain, M. (2015). Dimensions of satisfaction factors: Road to successful & sustainable organization. *IOSR Journal of Business and Management*, *17*(8), 94-106.

Hranisavljevic, N., Niggemann, O., & Maier, A. (2020). A novel anomaly detection algorithm for hybrid production systems based on deep learning and timed automata. *arXiv preprint arXiv:2010.15415*.

Jabangwe, R., Börstler, J., Šmite, D., & Wohlin, C. (2015). Empirical evidence on the link between object-oriented measures and external quality attributes: a systematic literature review. *Empirical Software Engineering*, *20*(3), 640-693.

Jafarian, T., Masdari, M., Ghaffari, A., & Majidzadeh, K. (2020). Security anomaly detection in software ‐ defined networking based on a prediction technique. *International Journal of Communication Systems*, *33*(14), e4524.

Klein, S.M., & Maher, J. (1966). Education level and satisfaction with pay. *Personnel Psychology*, *19*(2), 195-208.

Koskinen, K.U., & Pihlanto, P. (2006). Competence transfer from old timers to newcomers analysed with the help of the holistic concept of man. *Knowledge and Process Management*, *13*(1), 3-12.

Li, F., & Yu, F. (2020). Multi-factor one-order cross-association fuzzy logical relationships based forecasting models of time series. *Information Sciences*, *508*, 309-328.

Marchisio, M., Barana, A., Fioravera, M., Rabellino, S., & Conte, A. (2018). A model of formative automatic assessment and interactive feedback for STEM. IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC).

McCallin, A., & McCallin, M. (2009). Professional Perspective-Factors influencing team working and strategies to facilitate successful collborative teamwork. *New Zealand Journal of Physiotherapy*, *37*(2), 61.

Mitra, A. (2016). *Fundamentals of quality control and improvement*. John Wiley & Sons.

Nonaka, I., & Konno, N. (1998). The concept of "ba": Building a foundation for knowledge creation. *California Management Review*, *40*(3), 40-54.

Nonaka, I., & Teece, D.J. (2001). *Managing industrial knowledge: creation, transfer and utilization*. Sage.

Pedraja-Rejas, L., Rodriguez-Ponce, E., Rodriguez Mardones, P., Ganga Contreras, F., & Villegas Villegas, F. (2016). Determinants of the level of satisfaction of students in their schools: An exploratory study in chile. *Interciencia*, *41*(6), 401-406.

Pinson, M.H., Staelens, N., & Webster, A. (2013). The history of video quality model validation. Multimedia Signal Processing (MMSP), IEEE 15th International Workshop on.

Purdy, C., Wang, X., He, L., & Riedl, M. (2018). Predicting generated story quality with quantitative measures. Fourteenth Artificial Intelligence and Interactive Digital Entertainment Conference.

Robertson, R. (2016). Globalization, cultural dimensions. *The Wiley Blackwell Encyclopedia of Race, Ethnicity, and Nationalism*.

Saleh, S.D., & Otis, J.L. (1964). Age and level of job satisfaction. *Personnel Psychology, 17*(4), 425-430.

Schwartz, S.H. (1994). *Beyond individualism/collectivism: New cultural dimensions of values*. Sage Publications, Inc.

Sergiovanni, T. (1967). Factors which affect satisfaction and dissatisfaction of teachers. *Journal of Educational Administration*, *5*(1), 66-82.

Sharma, T., Kechagia, M., Georgiou, S., Tiwari, R., Vats, I., Moazen, H., & Sarro, F. (2021). A survey on machine learning techniques for source code analysis. *arXiv preprint arXiv:2110.09610*.

Shewfelt, R.L. (1999). What is quality? *Postharvest Biology and Technology*, *15*(3), 197-200.

Shi, G. (2013). *Data mining and knowledge discovery for geoscientists*. Elsevier.

Silitonga, P. (2021). Competency-based education: A multi-variable study of tourism vocational high school graduates. *Journal of Teaching in Travel & Tourism*, *21*(1), 72-90.

Singh, M., Gupta, P.K., Tyagi, V., Sharma, A., Ören, T., & Grosky, W. (2017). *Advances in computing and data sciences: First international conference, ICACDS 2016, Ghaziabad, India, November 11-12, Revised Selected Papers* (Vol. 721). Springer.

Thireau, M. (2002). Does brain weight have meaning. *The bi-monthly Journal of The BWW Society*, *2*(2), 1-3.

Tian, D., Jia, X., Ma, R., Liu, S., Liu, W., & Hu, C. (2021). BinDeep: A deep learning approach to binary code similarity detection. *Expert Systems with Applications*, *168*, 114348.

Vassiliadis, S., & Schwarz, E.M. (1990). High speed parity prediction for binary adders using irregular grouping scheme. In: Google Patents.

Wang, A. (2023). Embedded system architecture-computer embedded software defect prediction based on genetic optimisation algorithms. *International Journal of Information Technology and Management*, *22*(3-4), 262-280.

Wittenberg, E., Kravits, K., Goldsmith, J., Ferrell, B., & Fujinami, R. (2016). Validation of a model of family caregiver communication types and related caregiver outcomes. *Palliative & Supportive Care*, 1-9.

Wohlfart, O., Adam, S., & Hovemann, G. (2022). Aligning competence-oriented qualifications in sport management higher education with industry requirements: An importance–performance analysis. *Industry and Higher Education*, *36*(2), 163-176.

Woods, J. (2018). Against reflective equilibrium for logical theorizing.

Zhang, Z. (2023). *Revamping Binary Analysis with Sampling and Probabilistic Inference* Purdue University Graduate School.

International Journal of Assessment Tools in Education

2024, Vol. 11, No. 2, 303–319

https://doi.org/10.21449/ijate.1369023

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

The general attitudes towards artificial intelligence (GAAIS): A meta-analytic reliability generalization study

Melek Gülşah Şahin [1,*], **Yıldız Yıldırım** [2]

[1]Gazi University, Gazi Faculty of Education, Department of Educational Sciences, Ankara, Türkiye
[2]Aydın Adnan Menderes University, Faculty of Education, Department of Educational Sciences, Aydın, Türkiye

ARTICLE HISTORY

Received: Sep. 30, 2023

Accepted: Mar. 22, 2024

Keywords:

General attitudes towards artificial intelligence scale,
Reliability generalization,
Cronbach's α,
Mc Donald's ω,
Moderator analysis.

Abstract: This study aims to generalize the reliability of the GAAIS, which is known to perform valid and reliable measurements, is frequently used in the literature, aims to measure one of today's popular topics, and is one of the first examples developed in the field. Within the meta-analytic reliability generalization study, moderator analyses were also conducted on some categorical and continuous variables. Cronbach's α values for the overall scale and the positive and negative subscales, and McDonald's ω coefficients for positive and negative subscales were generalized. Google Scholar, WOS, Taylor & Francis, Science Direct, and EBSCO databases were searched to obtain primary studies. As a result of the screening, 132 studies were found, and these studies were reviewed according to the inclusion criteria. Reliability coefficients obtained from 19 studies that met the criteria were included in the meta-analysis. While meta-analytic reliability generalization was performed according to the random effects model, moderator analyses were performed according to the mixed effect model based on both categorical variables and continuous variables. As a result of the research pooled, Cronbach's α was 0.881, 0.828, and 0.863 for total, the negative, and positive subscales respectively. Also, McDonald's ω was 0.873 and 0.923 for negative and positive subscales respectively. It was found that there were no significant differences between the reliability coefficients for all categorical variables. On the other hand, all continuous moderator variables (mean age, standard deviation age, and rate of female) had a significant effect.

1. INTRODUCTION

In everyday life, applications related to artificial intelligence are encountered or used almost daily. Reasons such as the fact that computers play an essential role in our lives and that their use increases due to the convenience they bring and the different experiences they offer every day and that they eliminate the problem of space and time in accessing information, provide an understanding of the popularity of artificial intelligence. There are many studies related to artificial intelligence in many fields. When the keyword "artificial intelligence" is searched in Google Scholar, 3.490.000 research studies are found. Especially ChatGPT, which is one of the most important AI applications recently, maintains its popularity in all fields.

*CONTACT: Melek Gülşah Şahin ✉ mgulsahsahin@gazi.edu.tr Gazi University, Gazi Faculty of Education, Department of Educational Sciences, Ankara, Türkiye.

e-ISSN: 2148-7456

Artificial intelligence is the combination of science and engineering to create intelligent computers or programs to perform tasks related to human intelligence (McCarthy, 2004). Here, human intelligence and machines interact with each other. In other words, artificial intelligence is defined as the ability of a computer to perform features such as reasoning, problem-solving, inferring, and generalizing, as in humans (Arslan, 2020). One of the most important examples of artificial intelligence is Cog and Kismet, developed in MIT laboratories in the 1990s. While Cog is an upper-body robot with visual, emotional, and kinesthetic features, Kismet is a robot head with active vision and facial expressions (Turkle et al., 2006). In addition, the most crucial feature of Kismet and Cog is that they are robots with humanoid behaviors that can communicate emotionally and socially with people. While traditional robots are equipped with applications for less communication with humans, Kismet and Cog are social robots open to sharing with humans (Breazeal, 2004). In the field of education, one of the first applications of artificial intelligence was Skinner's individualized teaching machines implemented in 1958 (Arslan, 2020). Today, artificial intelligence applications show themselves in all fields without slowing down. In addition, our accessibility to artificial intelligence is increasing day by day. SIRI, which is one of the smartphone applications, is also an AI application (Kaplan & Haenlein, 2019). In addition, many applications such as autonomous cars, virtual classrooms, face recognition, patient tracking systems, instant language translators, automation, investment tools, games, and language translations are AI applications that are constantly developing and updating themselves (Arslan, 2020; Wang et al., 2022)

Artificial intelligence applications both facilitate human life and help them gain new knowledge and experiences. Examining individuals' knowledge, experiences, attitudes, and opinions toward AI applications also contributes to the literature on the development of artificial intelligence technology. In the literature, there are studies examining individuals' attitudes toward AI applications that manifest themselves in almost all fields and that we benefit from their applications or results practically every day in our lives. For example, in the field of economics, the effect of the use of artificial intelligence in shopping on consumers' decision-making processes (Nica, 2022), the determination of attitudes toward the use of artificial intelligence in personal financial planning (Waliszewski, 2020) and in the field of health, the attitudes of dermatologists towards the use of artificial intelligence in dermatology (Polesie, 2020), the attitudes of medical students towards the use of artificial intelligence in radiology and general medicine (Pinto dos Santos et al., 2019), the attitudes and perceptions of dental students towards the use of artificial intelligence in various clinical tasks have been examined (Yuzbasioglu, 2021). It can be stated that artificial intelligence and computer technology have an important place in educational life from kindergarten to university (Kandlhofer et al., 2016). There are also studies in the field of education such as the effect of using artificial intelligence in learning environments on students' attitudes (Huang, 2018), investigating artificial intelligence anxiety and attitudes toward machine learning in pre-service teachers (Hopcan et al., 2023), and investigating university students' attitudes towards the use of SIRI in English as a foreign language (EFL) learning (Haryanto, 2019).

When the related literature is examined, it is seen that there are studies in which different measurement tools have been developed to determine attitudes toward artificial intelligence. Some of these are Attitude Towards Artificial Intelligence Scale (ATAI) (Sindermann et al., 2020), Threats of Artificial Intelligence Scale (TAI) (Kieslich et al., 2021), Negative Attitude towards Artificial Intelligence Scale (NAAIS) (Persson et al., 2021), General Attitudes Towards Artificial Intelligence Scale (GAAIS) (Schepman & Rodway, 2020; 2023), and AI Attitude Scale (AIAS-4) (Grassini, 2023). Within the scope of this research, the scales used for artificial intelligence were examined and it was determined that the most cited attitude scale was the General Attitudes Towards Artificial Intelligence Scale (GAAIS). There are 134 Google Scholar citations of the term, 113 in 2020 when GAAIS was developed, and 21 in 2023 while it has been cited a total of 53 times in the Web of Science. The other reason for selecting this scale is that validity and reliability studies have been conducted in different cultures, such as Türkiye (Kaya et al., 2022), Korea (Seo & Ahn, 2022), Finland (Bergahdl et al., 2023), and Germany (Carolus et al., 2023). The scale Schepman and Rodway (2020) developed includes 20 items and two sub-dimensions. While the positive subscale represents social and personal benefits, the negative subscale represents concerns. The developed scale was applied to 100 people (50 women and 50 men) over the age of 18 who were not students. The majority of the respondents worked in the service sector. They observed jobs from a variety of socioeconomic classes (such as cleaner, caretaker, linen assistant, sales assistant, etc.) and created 16 positive items (opportunities, benefits, and positive emotions) and 16 negative items (concerns and negative emotions) that mirrored the positive and negative themes discovered from the literature. Explanatory Factor Analysis (EFA) was conducted for the items developed. Seven items were removed from the scale based on the item correlation matrix because they showed a low correlation with other items. As a result of the EFA performed for the remaining 25 items, five items were removed because four items had factor loading values below 0.40 and 1 item had equal loading values in both dimensions, leaving 20 items. EFA was applied again to the remaining 20 items. Bartlett's Test of Sphericity $\chi^2 = 817$, $df = 190$, $p < .001$, and The Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO MSA) was 0.86, indicating adequate sample size. In the final model, eight items were loaded on the first factor (negative views of AI), and twelve were loaded onto the second (positive attitudes towards AI). In this way, the assumptions made when the items were created about their positivity and negativity were statistically justified, providing good construct validity for the factor structure. The first and second factors accounted for 25.6% and 15.5% of the variance, respectively. The model fit measures were RMSEA = 0.0573, 90% CI [0.007, 0.068], TLI = 0.94, and the model test $\chi^2 = 182$, $df = 151$, $p = .046$, which are acceptable, namely positive attitudes towards AI ($\alpha = 0.88$) and negative attitudes towards AI ($\alpha = 0.83$).

For the validity evidence of the scale, the Technology Readiness Index scale (TRI), consisting of 18 items and four sub-dimensions, was applied to the study group and correlation and regression analyses were performed with innovation, optimism, discomfort, and insecurity sub-dimensions. The sub-dimensions of the scale were taken as independent variables, positive attitudes toward AI and negative attitudes towards AI were taken as dependent variables, and regression analysis was performed. GAAIS was reported as a valid and reliable scale when all the data were evaluated together.

Schepman and Rodway (2022), in a two-stage study (Study 1 and Study 2) that they considered the second dimension of scale development, applied the previously developed GAAIS to a sample group of 304 people to conduct CFA in Study 1 and examined its construct validity. They examined various model fit indices: $\chi^2 = 223.08$, $df = 169$, $p = 0.003$, $\chi^2/df= 1.32$, CFI = 0.987, TLI = 0.986, SRMR = 0.065, RMSEA = 0.032, 90% CI [0.019, 0.044], $p = 0.997$, suggesting an imperfect fit. In Study 1; the researchers found the standard solutions as a range of 0.310 – 0.851 for 8-item negative attitudes towards the AI subscale (GAAIS-NA); they also found that the standard solutions were in the range the range of 0.464 – 0.803 for the 12-item positive attitudes towards AI subscale (GAAIS-PA). The factor covariance was 0.492, 95% CI [0.455,0.528], SE = 0.019, $z = 26.215$, $p < 0.001$, and the correlation between the two factors was $r = 0.397$, $p < 0.001$. In addition, correlation and regression analyses were conducted between TRI and GAAIS in Study 1. In Study 2, correlation and regression analyses were conducted in a sample group of 300 people with the scores obtained from the 30-item Big Five Inventory-2 Short Form (Soto & John, 2017) consisting of Extraversion, Agreeableness, Conscientiousness, Negative Emotionality, and Open-Mindedness dimensions, the 13-item Corporate Distrust Scale (Adams et al., 2010), and the 6-item General Trust Scale (Yamagishi & Yamagishi). They also found $\alpha = 0.85$ for the positive attitude dimension and $\alpha = 0.82$ for the negative attitude dimension. As a result, it was concluded that the GAAIS performed valid and reliable measurements in this study as well as in the previous scale development study.

Additionally, no studies on the meta-analytic reliability of the GAAIS have been found in the literature. Accordingly, this study aims to generalize the meta-analytic reliability of the GAAIS, which is known to perform valid and reliable measurements and is frequently used in the literature, aims to measure one of today's popular topics, and is one of the first examples developed in the field. Thus, it is thought that this scale will provide preliminary information and shed light on how the reliability of this scale will change according to the variables (different cultures, language, age, rate of females, etc.).

2. METHOD

This study follows the meta-analytic method of Vacha-Haase (1998) and aims to generalize the reliability of the General Attitudes Towards Artificial Intelligence Scale. Accordingly, Cronbach's α values for the overall scale and Cronbach's α values for the positive and negative sub-dimensions were generalized. In addition, McDonald's ω coefficients for positive and negative sub-dimensions were also generalized.

2.1. Data Collection and Coding Process

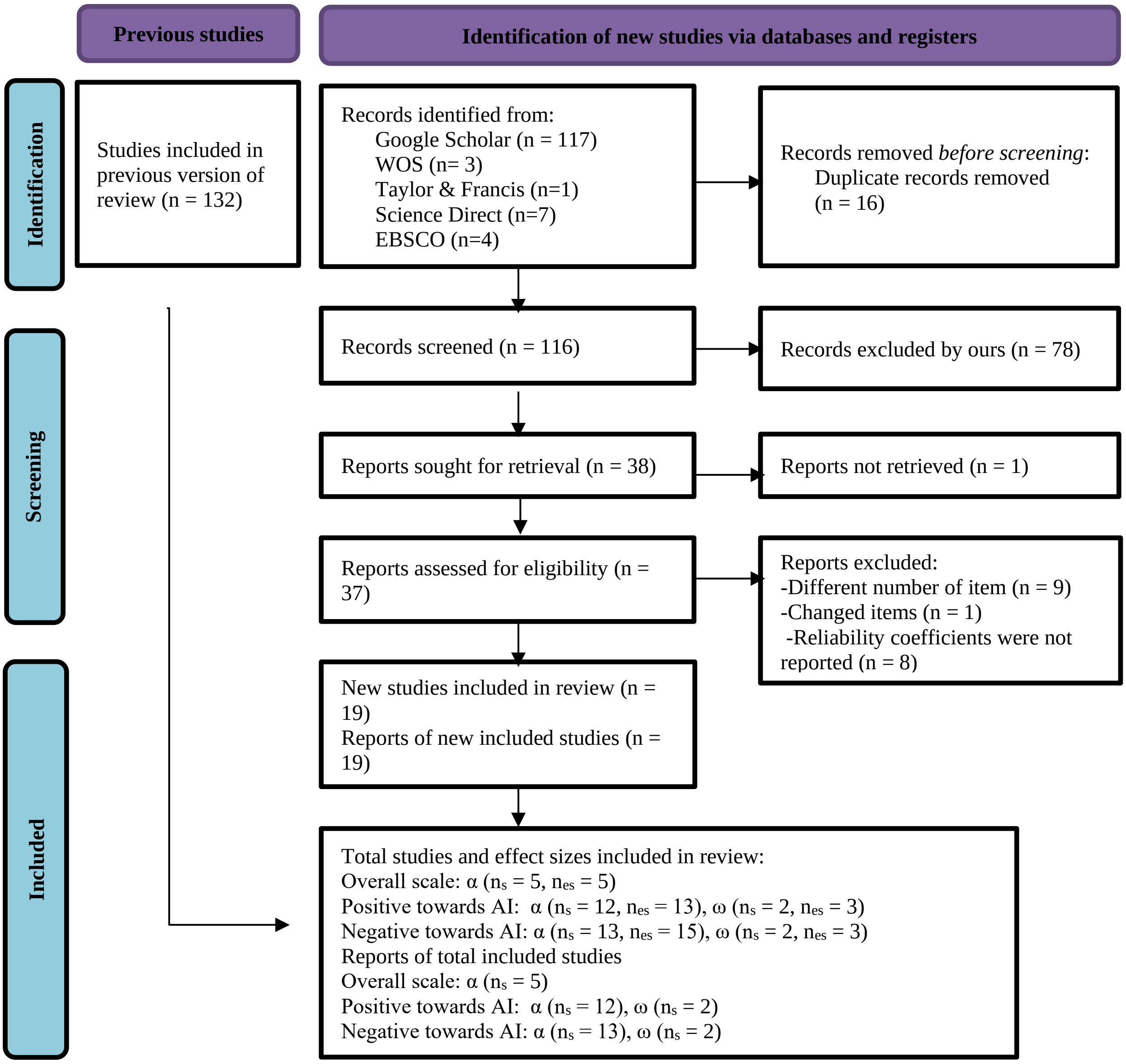
Google Scholar, WOS, Taylor & Francis, Science Direct, and EBSCO databases were searched with the keyword "Attitudes Towards Artificial Intelligence Scale" to access the studies in which the scale was used to perform meta-analytic reliability generalization. As a result of the searches, the full texts of all studies in the databases were examined and included in the meta-analytic reliability generalization if they met the specified criteria. The inclusion criteria can be listed as follows:

- The reporting language of the study can be any language.
- One of the original or adapted forms of the General Attitudes Towards Artificial Intelligence Scale (overall or subscales) must have been used.
- The overall and/or sub-dimensional reliability (Cronbach's α and/or McDonald's ω) of the scale must have been reported.
- The size of the sample to which the scale was applied must have been reported.

The PRISMA 2020 Statement flowchart for the identification, searching, and inclusion of full-text articles reviewed according to these criteria is given in Figure 1 (Page et al., 2021).

According to the PRISMA flowchart shown in Figure 1, 132 studies were identified by searching the determined databases. Of these 132 studies, 16 were eliminated because they were duplicates, 78 were eliminated because they did not use the scale and were related to the subject, and one was eliminated because it could not be accessed. When the remaining 37 studies were examined using the inclusion criteria, it was seen that different numbers of items were used in nine studies (6, 8, 9, 15, 16, and 32 items) and the items were changed in one study. The reliability coefficient was not reported in 8 of the remaining 27 studies. Thus, 19 studies were included in the analysis. In 5 of these 19 studies, 5 Cronbach's α coefficients for the overall scale were generalized. 15 Cronbach's α coefficients from 13 studies for negative attitudes toward AI and 13 Cronbach's α coefficients from 12 studies for positive attitudes towards AI were included in the analysis. Besides, for the other coefficient included in the study, McDonald's ω, three coefficients from 2 studies were generalized for positive and negative subscales.

Figure 1. *PRISMA flowchart.*


Previous studies
Identification of new studies via databases and registers
Identification
Screening
Included
Studies included in previous version of review (n = 132)
Records identified from:
Google Scholar (n = 117)
WOS (n= 3)
Taylor & Francis (n=1)
Science Direct (n=7)
EBSCO (n=4)
Records removed *before screening*:
Duplicate records removed
(n = 16)
Records screened (n = 116)
Records excluded by ours (n = 78)
Reports sought for retrieval (n = 38)
Reports not retrieved (n = 1)
Reports assessed for eligibility (n = 37)
Reports excluded:
-Different number of item (n = 9)
-Changed items (n = 1)
-Reliability coefficients were not reported (n = 8)
New studies included in review (n = 19)
Reports of new included studies (n = 19)
Total studies and effect sizes included in review:
Overall scale: α (n_s = 5, n_{es} = 5)
Positive towards AI: α (n_s = 12, n_{es} = 13), ω (n_s = 2, n_{es} = 3)
Negative towards AI: α (n_s = 13, n_{es} = 15), ω (n_s = 2, n_{es} = 3)
Reports of total included studies
Overall scale: α (n_s = 5)
Positive towards AI: α (n_s = 12), ω (n_s = 2)
Negative towards AI: α (n_s = 13), ω (n_s = 2)

*n_s = number of studies, n_{es} = number of effect sizes

After the search, the studies selected by the inclusion criteria were coded by two coders. In addition to coding reliability coefficients and sample sizes, the two coders coded some descriptive variables to perform a reliability generalization analysis. These variables were (i) publication citation, (ii) year of publication, (iii) language of publication, (iv) type of research, (v) overall reliability type, (vi) sub-dimension reliability type, (vii) the number of response categories, (viii) scale language, (ix) country, (x) mean age, (xi) standard deviation of age, (xii) study group, (xiii) rate of females, and (xiv) study field. The frequencies of the sub-categories for the categorical variables are given in Table 1.

Since all the research data were coded by two coders, the percentage of inter-coder agreement was calculated according to Miles and Huberman (1994), and the agreement was found to be 100%. After the agreement of the coded data was determined, the data were analyzed.

Table 1. *Frequency of studies and effect sizes for Cronbach's α.*

		GAAIS-Overall		GAAIS-Negative		GAAIS-Positive	
	Categories	n_s	n_{es}	n_s	n_{es}	n_s	n_{es}
Publish Type	Manuscript	4	4	13	15	12	13
	Proceeding	1	1	-	-	-	-
Publish Language	English	5	5	12	14	11	12
	Korean	-	-	1	1	1	1
Scale Language	English	4	4	9	11	8	9
	Korean	-	-	2	2	2	2
	German	-	-	1	1	1	1
	Turkish	-	-	1	1	1	1
	Arabic	1	1	-	-	-	-
Likert Type	3-point	1	1	-	-	-	-
	5-point	3	3	13	15	12	13
	7-point	1	1	-	-	-	-
Region	Asia	2	2	5	5	5	5
	Europe	1	1	5	6	5	6
	America	1	1	2	2	2	2
Study Group	Adult	4	4	7	9	7	9
	Student	1	1	3	3	3	3
	Adult and Student	-	-	2	3	1	1
Research Type	Correlational	3	3	8	9	7	7
	Scale Adaptation/Development	-	-	6	6	6	6
	Descriptive	1	1	-	-	-	-
	Experimental	1	1	-	-	-	-
Study Field	Psychology	1	1	7	8	7	8
	Health Science	1	1	4	4	4	4
	Management/Communication	3	3	2	3	1	1

GAAIS-Overall = General Attitude of Artificial Intelligence Scale, GAAIS-Positive = GAAIS Positive Subscale, GAAIS-Negative = GAAIS Negative Subscale

2.2. Data Analysis

In the studies handled within the scope of the research, meta-analytic reliability generalization regarding the reliability coefficients for both the overall scale and the subscales was carried out. Meta-analytic reliability generalization, which is an extension of the validity generalization, is used to determine the mean measurement error variance between studies and the sources of the variance (Vacha-Haase, 1998). Meta-analytic reliability generalization analysis was performed with the CMA v.2 program. The reliability coefficient, which is the correlation coefficient, is not suitable for meta-analysis, because the variance depends on correlation. Therefore, it can be combined by transforming and then transforming to a reliability coefficient again (Borenstein et al., 2009; Thompson & Vacha-Haase, 2000). The reliability coefficients obtained in the study were transformed into Fisher's z statistics before being included in the analysis. This transformation method has been suggested in the literature and is often used by meta-analysts (Beretvas et al., 2002). Heterogeneity was examined to determine the type of model to be used in the analyses (Borenstein et al., 2009). Q statistic and its significance (Cochran, 1954), I^2 statistic (Higgins & Thomson, 2002), and τ^2 values were analyzed to examine the heterogeneity of the distributions of the studies. The τ^2 estimates were made following the Der-

Simonian Laird (1986) method. Then, publication bias, a crucial issue in meta-analysis studies, was examined. The research also paid attention to scanning in different databases for publication bias. In addition, Rosenthal's (1979) fail and safe method, Begg and Mazumdar's (1994) rank correlation test, Egger's linear regression test (Egger et al., 1997), and Duval and Tweedie's trim and fill method based on funnel plot were used to examine publication bias in the data obtained.

In the study, the α coefficient for the overall reliability of the GAAIS and the reliability generalization of the α and ω coefficients for the subscales were analyzed (Vacha-Haase, 1998). These analyses were carried out according to the random effect model since heterogeneity exists statistically and theoretically (Borenstein et al., 2009). The reliability coefficient may vary depending on the applied group (Crocker & Algina, 1986). Heterogeneity in social sciences is a theoretically expected situation. Because the measures were obtained from individuals living in different regions, speaking different languages, and of different ages and characteristics, within the scope of the research, moderator analyses were performed according to the mixed effect model based on both categorical variables and continuous variables. In the selection of variables, situations where the reliability value may differ in the literature were determined (Aslan et al., 2022; Hess et al., 2014; Lopez-Pina et al., 2015; Ozdemir et al., 2020; Yin & Fan, 2000). Analog ANOVA analysis was performed for each subgroup based on region, study group, research type, and study field variables in Table 1. At this stage, the statistical significance of the reliability coefficients obtained for each subgroup was analyzed. Analog ANOVA is performed to test the significance of the difference in the dependent variable in the subcategories of the categorical independent variable. If there is heterogeneity, this variability may be due to subgroups, so the sources of heterogeneity can be determined by performing Analog to ANOVA. Also, meta-regression analyses were performed for continuous variables such as mean age, standard deviation of age, and rate of females (Caruso & Edwards, 2001; Hess et al., 2014; Youngstrom & Green, 2003). Thus, sample characteristics that could reveal differences in the homogeneity of the group were considered moderator variables in the study (Henson & Thompson, 2002). The significance of the models and explained variance values were reported (Hedges & Pigott, 2004).

3. RESULTS

Within the scope of the research, meta-analytic reliability generalization of the reliability coefficients for GAAIS, GAAIS-Negative, and GAAIS-Positive was conducted. For Cronbach's α, the overall scale, and its subscales were considered, while for McDonald's ω, only the subscales were considered because studies reporting McDonald's ω did not calculate this coefficient for the overall scale. For reliability generalizations, publication bias results were first examined and are presented in Table 2.

Table 2. *Publication bias.*

	Overall GAAIS (α)	GAAIS-Negative (α)	GAAIS-Positive (α)	GAAIS-Negative (ω)	GAAIS-Positive (ω)
Rosenthal Fail-safe	1782	15904	13545	5038	7215
Kendall's τ	0.000	-0.210	0.115	0.333	0.667
Intercept	-2.188	-2.218	-3.579	8.628	7.074
Adjusted studies	0	0	0	0	0

When the failsafe-N results regarding publication bias given in Table 2 were examined, it was an indication that publication bias did not exist since the number of missing studies that should be added for the overall reliability coefficient to be non-significant was higher than the criterion value (5k+10) for the overall scale and subscales (Rosenthal, 1979). *k* is the number of studies used in calculating this criterion value. Kendall's τ and Egger regression intercept values were not significant. These tests showed that there was no evidence of funnel plot asymmetry (Begg & Mazumdar, 1994; Egger et al., 1997; Rothstein et al., 2005). Finally, when the results of Duval and Tweedie's trim and fill method were analyzed, it was observed that the number of adjusted studies for the funnel plot to be symmetric was 0 in all results. Accordingly, it could be said that there was no publication bias according to the method by Duval and Tweedie. When all the evidence was analyzed together, it was concluded that there was no publication bias for all coefficients in the overall scale and subscales. In addition, the induction rate was calculated in the study and this rate was found to be 29.63% ((8/27) ×100)) (Vacha-Haase et al., 2000; Sanchez-Meca et al., 2021). The pooled Cronbach's α and McDonald's ω values for the overall scale and subscales and also heterogeneity statistics are presented in Table 3.

Table 3. *Results for overall effect sizes and heterogeneity.*

Overall Coefficients						
Coefficients		*k*	RC [LL_{RC}- UL_{RC}]	Q	I^2	τ^2
Cronbach's α	Overall	5	0.881* [0.849-0.907]	12.002*	66.671	0.014
	Negative	15	0.828* [0.807-0.846]	65.502*	78.627	0.011
	Positive	13	0.863* [0.840-0.883]	90.917*	86.801	0.020
Mc Donald's ω	Negative	3	0.873* [0.859-0.886]	6.074*	67.075	0.002
	Positive	3	0.923* [0.916-0.929]	3.068	34.820	0.000

*p<0.05, RC: Reliability Coefficient, LL_{RC}: Lower Limit, UL_{RC}: Upper Limit, *k:* Number of studies

When analyzing the significance of the reliability values for the total scale and the subscales in Table 3, it was found that all coefficients obtained in both types of reliability coefficients were statistically significant. When Cronbach's α was analyzed, the highest value was obtained in the overall scale, while the lowest value was obtained in the negative subscale. When McDonald's ω values were analyzed, the highest value was obtained in the positive subscale. When both reliability types were analyzed for the subscales, it was observed that McDonald's ω reliability values were higher than Cronbach's α values.

In the heterogeneity values given in Table 3, the Q value was found to be significant for all scales where Cronbach's α was generalized, while for the ω coefficient, it was found to be significant for GAAIS-Negative and not significant for GAAIS-Positive. For I^2, another evidence of heterogeneity, GAAIS-overall α, and GAAIS-Negative ω could be considered as moderate heterogeneity indicators. For GAAIS-Negative α and GAAIS-Positive α, I^2 could be said to be a high-level heterogeneity indicator. For GAAIS-Positive ω, a low level of heterogeneity was determined (Higgins et al., 2003). When the variance between studies (τ^2) was analyzed, it was seen that all of them except GAAIS-Positive ω were different from 0 and there was a variance between studies. For the ω coefficient of the GAAIS-Positive subscale, it could be said that there was no variance between the studies. In general, heterogeneity existed for both α coefficients and ω coefficients. Forest plots for Cronbach's α coefficient are given for the negative subscale and positive subscale in Appendix 1 and Appendix 2. When the forest plots were examined, it was seen that the Cronbach's alpha (standard error) of the primary studies for both the negative and positive subscales were distributed heterogeneously. The results of moderator analysis with categorical and continuous variables are presented in Table 4.

Table 4. *Results for categorical/continuous moderator analysis.*

			GAAIS-Negative		
Categorical Moderator	Categories	*k*	α [LLα-ULα]	Q (*df*)	*p*
Region	Asia	5	0.811 [0.763-0.850]	1.101(2)	0.577
	Europe	6	0.827 [0.788-0.859]		
	America	2	0.848 [0.785-0.894]		
Study Group	Adult	9	0.826 [0.799-0.850]	0.923(2)	0.630
	Student	3	0.813 [0.756-0.858]		
	Adult and Student	3	0.843 [0.800-0.878]		
Research Type	Correlational	9	0.831 [0.803-0.855]	0.127(1)	0.722
	Scale Devel. /Adapt.	6	0.823 [0.788-0.853]		
Study Field	Psychology	8	0.834 [0.808-0.857]	2.449(2)	0.294
	Health Science	4	0.800 [0.752-0.839]		
	Communication	3	0.840 [0.799-0.873]		
		k	β [SE]	Q_M	Q_R
Continuous Moderator	Mean Age	14	0.005 [0.002]	10.098*	46.870*
	Standard Deviation of Age	14	0.007 [0.003]	4.705*	52.263*
	Rate of Female	15	-0.419 [0.105]	15.795*	49.707*
			GAAIS-Positive		
		k	α [LLα-ULα]	Q(*df*)	*p*
Region	Asia	5	0.847 [0.814-0.876]	3.010(2)	0.222
	Europe	6	0.878 [0.855-0.898]		
	America	2	0.852 [0.800-0.892]		
Study Group	Adult	9	0.870 [0.844-0.892]	0.253(1)	0.615
	Student	3	0.857 [0.802-0.897]		
Research Type	Correlational	7	0.852 [0.823-0.877]	1.638(1)	0.201
	Scale Devel./Adapt.	6	0.875 [0.849-0.897]		
Study Field	Psychology	8	0.872 [0.846-0.893]	0.529(1)	0.467
	Health Science	4	0.855 [0.811-0.889]		
		k	β [SE]	Q_M	Q_R
Continuous Moderator	Mean Age	12	0.008 [0.002]	25.393*	63.679*
	Standard Deviation of Age	12	0.014 [0.004]	15.857*	73.216*
	Rate of Female	13	-0.411 [0.106]	15,112*	75.805*

$^*p<0.05$, $LL_α$: Lower Limit, $UL_α$: Upper Limit, *k:* Number of studies, β: Slope, Q_M: Q values for model, Q_E: Q values for residual

The moderator analysis handled the categorical variables (region, study group, research type, and study field). When analog ANOVA results for the negative subscale and the positive subscale were examined, it was observed that Cronbach's α did not differ significantly in the sub-categories of the variables. In the negative sub-dimension, it was observed that the Cronbach α value obtained in the region-based analyses was the highest in the American region and the lowest in the Asia region. In the analysis based on the study group, the highest reliability value was obtained in the "adult and student" subgroup and the lowest in the student subgroup. In the analysis based on research type, higher reliability values were found in correlational studies. In the study field, the highest reliability value was obtained in communication and the lowest in health science.

In the positive subscale, the highest reliability value was obtained in Europe and the lowest in Asia as a result of the region-based analog to ANOVA. However, this difference was not significant. There was no significant difference between Cronbach's α results based on study groups, but a higher α coefficient was obtained in the analyses conducted with adults. In the research type, the reliability value obtained from correlational studies was relatively higher and the difference was not significant. Finally, it was determined that the reliability value obtained in the field of psychology (based on the field of study) was higher and not statistically significant.

In the moderator analysis, the mean age, standard deviation of age, and rate of females were considered continuous variables. In the negative subscale, the model based on mean age was found to be statistically significant. There was a positive relationship between the mean age and the α coefficient. When the extent to which the mean age explained the variability in the α coefficient was examined by (QM/(OM+QE)) × 100 (Borenstein et al., 2009; Card, 2012), it was seen that the variance explained was 17.73%. It was concluded that the model established by considering the standard deviation of age as a continuous variable was also statistically significant. There was a positive relationship between the standard deviation of age and α coefficient. The variability of the standard deviation of age in α coefficient was 8.26%. In the moderator analysis based on the rate of females, the model was significant and there was a negative relationship between the model and the α coefficient, and the α coefficient decreases as the rate of female participants increases. The rate of the female variable explained 24.11% of the variability in the α coefficient.

When the results of continuous moderator analysis for positive attitudes toward AI were examined, it was seen that mean age, standard deviation of age, and rate of female variables all significantly predicted Cronbach's α. Among these variables, mean age and standard deviation of age positively predicted Cronbach's α, while the rate of females predicted it negatively. It could be said that the α coefficient increased with the rise in mean age and standard deviation of age, and the α coefficient decreased with the increase in the rate of females. When the extent to which mean age explained the variability in the α coefficient was analyzed, it was found that the variance explained was 28.508%. The variance explained by the standard deviation of age was 17.802%. Finally, the variance explained by the rate of females was 16.621%. In both positive and negative attitudes towards AI subscales, it was observed that mean age explained the most variability in the α coefficient.

4. DISCUSSION and CONCLUSION

This study aimed to generalize the reliability of the overall GAAIS scale and its subscales. Cronbach's α coefficient was examined for the overall scale, and Cronbach's α and McDonald's ω reliability coefficients were examined for the subscales. When the reliability coefficients were reviewed, it was seen that Cronbach's α coefficient was mainly examined in primary studies. Cronbach's α coefficient is a reliability coefficient that is frequently calculated in the literature (Osburn, 2002; Warrens, 2014). In addition, all reliability coefficients estimated for the overall scale and subscales are above .70 (Nunnally, 1978). Accordingly, it can be said that the overall reliability of the scale is high. McDonald's ω coefficients calculated for the subscales were higher than Cronbach's α. In general, the α coefficient is also lower than the other coefficients. In other words, Cronbach's α is defined as the lower limit of reliability (Kristoff, 1974; Novick & Lewis, 1967).

In our study, moderator analyses were conducted by selecting variables that were frequently examined in the literature in reliability generalization studies. In the region-based analyses, it was determined that the reliability value of the GAAIS scale did not change significantly in the studies conducted in Asian, European, and American regions. However, the reliability values obtained varied according to the regions. This difference was determined as 0.037 in the negative subscale and 0.031 in the positive subscale, but it was not significant. Based on this, it can be stated that the error rates of the responses of people living in different regions to the scale were also different. Obtaining different results in different regions could also be explained by the differentiation in terms of the homogeneity of the distribution of individuals' views on AI practices. In this study, it was observed that the overall reliability values of Europe and America were higher than those of Asia in both the negative subscale and the positive subscale. Similar to the results of this study, there are reliability generalization studies in the literature that calculate lower overall reliability coefficients in Asia (Alcorer-Bruno et al., 2020; Vassar, 2008)

In the categorical moderator analysis based on the study group, reliability estimates were calculated in both subscales in different subgroups, and it was concluded that the difference was not significant. Other studies in the literature conclude that there is no significant difference between the study groups (Thompson & Cook, 2002; Wallace & Weller, 2002). The fact that the lowest reliability value for both subscales was obtained in the student group can be explained by the fact that the students' responses to the scale were more inconsistent or that their views on the AI application were more homogeneous compared to the other group. In the positive subscale, a higher reliability value was obtained in the category of both adults and students. It can be stated that the AI applications that the students encounter in their educational life are also similar compared to those of the other adult groups. This result is related to the heterogeneity of the group and can be explained by the higher value of Cronbach's α. When the reliability for the group of students is generalized, there are also studies in the literature where lower reliability values were obtained compared to more heterogeneous adult or adult-student groups (Eser & Dogan, 2023; Yoruk & Sen, 2022).

Another variable type handled in the study was research type. The reliability values obtained also differed whether the research type was correlational or scale development/adaptation. Although this difference was higher, especially in the positive subscale, it was insignificant in both subscales. In the moderator analysis based on the study field, different coefficients were obtained in different study areas where the research was conducted and this difference was not found to be significant, which is similar to the studies in the literature (Ozdemir et al., 2020). However, in both subscales, the primary studies were conducted mainly in the field of psychology. In the negative and positive attitude subscales, Cronbach's α reliability value obtained from the studies conducted in psychology was higher compared to health science, which may be due to the higher number of studies in the field of psychology. As the number of studies increases, the heterogeneity of the sample may increase. In addition, this result can also be explained by the fact that the groups studying in the field of health are more homogeneous. The characteristics of the sample groups selected in psychology research (occupational status, age groups, family status, education levels, etc.) may differ. In the negative subscale, unlike the positive subscale, data were also obtained in the field of communication and the highest overall reliability value was obtained in this category.

The change in the reliability values of the negative and positive subscales of AI according to the predictor variable, mean age, was analyzed by meta-regression. Mean age was positively correlated with Cronbach's α in the subscales and significantly predicted it. It can be stated that as the average age of the participants increases, their answers are more consistent. A similar relationship exists between the standard deviation of age and Cronbach's α. This is expected because the change in the standard deviation of age indicates that the sample group is heterogeneous in terms of age. As a result of this heterogeneity, it is expected that the overall reliability values will be high. In the literature, it has been observed that there are studies with similar results (Caruso & Edward, 2001; Youngstrom & Green, 2003).

An interesting result obtained from the research is that there is a negative relationship between the rate of females and the scale's Cronbach's α reliability value. The reliability value obtained increases as the rate of females participating in the study decreases. It can also be stated that as the rate of men participating in the study increases, the consistency of the answers given regarding the scale increases. The fact that attitudes towards AI may differ according to gender may also cause this result. Similar to the results of this study, some studies in the literature have found that reliability decreases as the proportion of females increases (Beretvas et al., 2008; Eser & Dogan, 2023). In contrast to the results of this study, Beretvas et al. (2002) determined that reliability decreases as the proportion of men increases in their reliability generalization study.

In this study conducted within the scope of AI, one of the popular topics today, the most cited GAAIS scale was selected. Due to the increase in the number of studies in this field and the fact that the effect sizes are affected by the reliability of the measurement tools, it is vital to examine the reliability of the measurement tools and to determine the change according to the variables specified. With the increase in the number of related studies, moderator analyses can be performed by considering different variables than the variables addressed in this study.

Declaration of Conflicting Interests and Ethics

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

Contribution of Authors

Melek Gülşah Şahin: Investigation, Methodology, Screening Primary Studies, Coding Primary Studies, Software, Data Analysis, Resources, and Writing-original draft. **Yıldız Yıldırım:** Investigation, Methodology, Screening Primary Studies, Coding Primary Studies, Visualization, Software, Data Analysis, Resources, Writing-original draft.

Orcid

Melek Gülşah Şahin https://orcid.org/0000-0001-5139-9777

Yıldız Yıldırım https://orcid.org/0000-0001-8434-5062

REFERENCES

*: Included in the meta-analysis

Alcocer-Bruno, C., Ferrer-Cascales, R., Rubio-Aparicio, M., & Ruiz-Robledillo, N. (2020). The medical outcome study-HIV health survey: A systematic review and reliability generalization meta-analysis. *Research in Nursing & Health, 43*(6), 610-620. https://doi.org/10.1002/nur.22070

Arslan, K. (2020). Eğitimde yapay zekâ ve uygulamaları [Artificial intelligence and applications in education]. *The Western Anatolia Journal of Educational Sciences, 11*(1), 71-88. https://dergipark.org.tr/tr/pub/baebd/issue/55426/690058

Aslan, Ö.S., Gocen, S., & Sen, S. (2022). Reliability generalization meta-analysis of mathematics anxiety scale for primary school students. *Journal of Measurement and Evaluation in Education and Psychology, 13*(2), 117-133. https://doi.org/10.21031/epod.1119308

Begg, C.B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50*(4), 1088. https://doi.org/10.2307/2533446

*Bellaiche, L., Shahi, R., Turpin, M.H., Ragnhildstveit, A., Sprockett, S., Barr, N., ... & Seli, P. (2023). Humans versus AI: Whether and why we prefer human-created compared to AI-created artwork. *Cognitive Research: Principles and Implications, 8*(1), 1-22. https://doi.org/10.1186/s41235-023-00499-6

Beretvas, S.N., Meyers, J.L., & Leite, W.L. (2002). A reliability generalization study of the Marlowe-Crowne Social Desirability Scale. *Educational and Psychological Measurement, 62*(4), 570-589. https://doi.org/10.1177/0013164402062004003

Beretvas, S.N., Suizzo, M.A., Durham, J.A., & Yarnell, L.M. (2008). A reliability generalization study of scores on Rotter's and Nowicki-Strickland's locus of control scales. *Educational and Psychological Measurement*, *68*(1), 97-119. https://doi.org/10.1177/0013164407301529

*Bergdahl, J., Latikka, R., Celuch, M., Savolainen, I., Mantere, E.S., Savela, N., & Oksanen, A. (2023). Self-determination and attitudes toward artificial intelligence: Cross-national and longitudinal perspectives. *Telematics and Informatics*, 82, 102013. https://doi.org/10.1016/j.tele.2023.102013

Borenstein, M., Hedges, L.V., Higgins, J.P., & Rothstein, H.R. (2009). *Introduction to meta-analysis*. John Wiley & Sons.

Breazeal, C. (2004). *Designing sociable robots*. MIT.

Card, N. (2012). *Applied meta-analysis for social science research.* Guilford.

*Carolus, A., Koch, M., Straka, S., Latoschik, M.E., & Wienrich, C. (2023). MAILS-Meta AI Literacy Scale: Development and testing of an AI Literacy Questionnaire based on well-founded competency models and psychological change-and meta-competencies. arXiv preprint. https://doi.org/10.48550/arXiv.2302.09319

Caruso, J.C., & Edwards, S. (2001). Reliability generalization of the Junior Eysenck Personality Questionnaire. *Personality and Individual Differences*, 31, 173-184. https://doi.org/10.1016/S0191-8869(00)00126-4

Cochran, W.G. (1954). The combination of estimates from different experiments. *Biometrics*, 10, 101–129. https://doi.org/10.2307/3001666

*Cruz, J.P., Sembekova, A., Omirzakova, D., Bolla, S.R., & Balay-odao, E.M. (2023). General attitudes towards and readiness for medical artificial intelligence among medical and health sciences students in Kazakhstan. https://doi.org/10.2196/preprints.49536.

*Darda, K., Carre, M., & Cross, E. (2023). Value attributed to text-based archives generated by artificial intelligence. *Royal Society Open Science*, 10: 220915. https://doi.org/10.1098/rsos.220915

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled clinical trials*, *7*(3), 177-188. https://www.biostat.jhsph.edu/~fdominic/teaching/bio656/references/sdarticle.pdf

Eser, M.T., & Dogan, N. (2023). Life Satisfaction Scale: A meta-analytic reliability generalization study in Turkey sample. *Turkish Psychological Counseling and Guidance Journal*, *13*(69), 224-239. https://doi.org/10.17066/tpdrd.1223320mn

*Gabbiadini, A., Dimitri, O., Cristina, B., & Anna, M. (2023). Does ChatGPT pose a threat to human identity. *SSRN*, 4377900. https://doi.org/10.2139/ssrn.4377900

*Gozzo, M., Woldendorp, M.K., & De Rooij, A. (2021, December). Creative collaboration with the "brain" of a search engine: Effects on cognitive stimulation and evaluation apprehension. In *International Conference on ArtsIT, Interactivity and Game Creation* (pp. 209-223). Springer International Publishing.

Grassini, S. (2023). Development and validation of the AI attitude scale (AIAS-4): A brief measure of general attitude toward artificial intelligence. *Frontiers in Psychology*, 14: 1191628. https://doi.org/10.3389/fpsyg.2023.1191628

*Hadlington, L., Binder, J., Gardner, S., Karanika-Murray, M., & Knight, S. (2023). The use of artificial intelligence in a military context: Development of the Attitudes Toward AI in Defense (AAID) Scale. *Frontiers in Psychology*, 14, 1164810. https://doi.org/10.3389/fpsyg.2023.1164810

Hedges, L.V., & Pigott, T.D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods*, *9*(4), 426-445. https://doi.org/10.1037/1082-989x.9.4.426

*Heim, S., & Chan-Olmsted, S. (2023). Consumer trust in AI–human news collaborative continuum: preferences and influencing factors by news production phases. *Journalism and Media*, *4*(3), 946-965. https://doi.org/10.3390/journalmedia4030061

Henson, R.K., & Thompson, B. (2002). Characterizing measurement error in scores across studies: Some recommendations for conducting "reliability generalization" studies.

Measurement and Evaluation in Counseling and Development, *35*(2), 113-127. https://doi.org/10.1080/07481756.2002.12069054

Hess, T.J., McNab, A.L., & Basoglu, K.S. (2014). Reliability generalization of perceived ease of use, perceived usefulness, and behavioral intentions. *MIS Quarterly*, 38, 1-28. https://doi.org/10.25300/MISQ/2014/38.1.01

Higgins, J.P.T., & Thompson, S.G. (2002), Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21, 1539-1558. https://doi.org/10.1002/sim.1186

Hopcan, S., Turkmen, G., & Polat, E. (2023). Exploring the artificial intelligence anxiety and machine learning attitudes of teacher candidates. *Education and Information Technologies*, 1-21. https://doi.org/10.1007/s10639-023-12086-9

Huang, S.P. (2018). Effects of using artificial intelligence teaching system for environmental education on environmental knowledge and attitude. *Eurasia Journal of Mathematics, Science and Technology Education*, *14*(7), 3277-3284. https://doi.org/10.29333/ejmste/91248

Kandlhofer, M., Steinbauer, G., Hirschmugl-Gaisch, S., & Huber, P. (2016, October). Artificial intelligence and computer science in education: From kindergarten to university. In *2016 Institute of electrical and electronics engineers - Frontiers in education conference (IEEE-FIE)* (pp. 1-9). IEEE. https://doi.org/10.1109/FIE.2016.7757570

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who’s the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62*(1), 15-25. https://doi.org/10.1016/j.bushor.2018.08.004

*Kaya, F., Aydin, F., Schepman, A., Rodway, P., Yetisensoy, O., & Demir Kaya, M. (2022). The roles of personality traits, AI anxiety, and demographic factors in attitudes toward artificial intelligence. *International Journal of Human–Computer Interaction*, 1-18. https://doi.org/10.1080/10447318.2022.2151730

Kieslich, K., Lünich, M., & Marcinkowski, F. (2021). The threats of artificial intelligence scale (TAI) development, measurement and test over three application domains. *International Journal of Social Robotics*, 13, 1563-1577. https://doi.org/10.1007/s12369-020-00734

Kristof, W. (1974). Estimation of reliability and true score variance from a split of a test into three arbitrary parts. *Psychometrika*, 39, 491-499. https://doi.org/10.1007/BF02291670

*Kwak, Y., Ahn, J.W., & Seo, Y.H. (2022). Influence of AI ethics awareness, attitude, anxiety, and self-efficacy on nursing students’ behavioral intentions. *BMC Nursing*, *21*(1), 1-8. https://doi.org/10.1186/s12912-022-01048-0

*Kwak, Y., Seo, Y.H., & Ahn, J.W. (2022). Nursing students' intent to use AI-based healthcare technology: Path analysis using the unified theory of acceptance and use of technology. *Nurse Education Today*, 119: 105541. https://doi.org/10.1016/j.nedt.2022.105541

McCarthy, J. (2004). *What is artificial intelligence?*. http://www.formal.stanford.edu/jmc/whatisai/

*Mohamed, H.A., Awad, S.G., Eldiasty, N.E.M.M, & ELsabahy, H.E. (2023). Effect of the artificial intelligence enhancement program on head nurses' managerial competencies and flourishing at work. *Egyptian Journal of Health Care*, *14*(1), 624-645. https://doi.org/10.21608/EJHC.2023.287188

Nica, E., Sabie, O.M., Mascu, S., & Luţan, A.G. (2022). Artificial intelligence decision-making in shopping patterns: Consumer values, cognition, and attitudes. *Economics, Management and Financial Markets*, *17*(1), 31-43. https://doi.org/10.22381/emfm17120222.

*Nguyen, E. (2023). Trust and algorithmic decision making. *UC Santa Barbara*, *3*(2022), 1-15. https://escholarship.org/content/qt5z86t0dx/qt5z86t0dx.pdf

Novick, M.R., & Lewis, C.L. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, 32, 1-13. https://doi.org/10.1007/BF02289400

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

Osburn, H.G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*, *5*(3), 343–355. https://doi.org/10.1037/1082-989X.5.3.343

Ozdemir, V., Yildirim, Y., & Tan, S. (2020). A meta-analytic reliability generalization study of the Oxford Happiness Scale in Turkish sample. *Journal of Measurement and Evaluation in Education and Psychology*, *11*(4), 374-404. https://doi.org/10.21031/epod.766266

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron I., Hoffmann, T.C., Mulrow, C.D., …, & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372: 71. https://doi.org/10.1136/bmj.n71

Persson, A., Laaksoharju, M., & Koga, H. (2021). We mostly think alike: Individual differences in attitude towards AI in Sweden and Japan. *The Review of Socionetwork Strategies*, *15*(1), 123-142. https://doi.org/10.1007/s12626-021-00071-y

Pinto dos Santos, D., Giese, D., Brodehl, S., Chon, S.H., Staab, W., Kleinert, R., ... & Baeßler, B. (2019). Medical students' attitude towards artificial intelligence: A multicentre survey. *European radiology*, 29, 1640-1646. https://doi.org/10.1007/s00330-018-5601-1

Polesie, S., Gillstedt, M., Kittler, H., Lallas, A., Tschandl, P., Zalaudek, I., & Paoli, J. (2020). Attitudes towards artificial intelligence within dermatology: An international online survey. *British Journal of Dermatology*, *183*(1), 159-161. https://doi.org/ 10.1111/bjd.18875

Rothstein, H.R., Sutton, A.J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. John Wiley & Sons.

Rosenthal, R. (1979). The ‘‘file drawer problem’’ and tolerance for null results. *Psychological Bulletin*, 86, 638–641. https://doi.org/10.1037/0033-2909.86.3.638

*Saddique, F., Usman, M., Nawaz, M., & Mushtaq, N. (2020). Entrepreneurial orientation and human resource management: The mediating role of Artificial Intelligence. *Elementary Education Online*, *19*(4), 4969-4978. https://doi.org/10.17051/ilkonline.2021.05.777

Sánchez-Meca J, Marín-Martínez F, López-López JA, … & López-Nicolás, P. (2021). Improving the reporting quality of reliability generalization meta-analyses: The REGEMA checklist. *Research Synthesis Methods*, 12, 516-536. https://doi.org/10.1002/jrsm.1487

*Schepman, A., & Rodway, P. (2020). Initial validation of the general attitudes towards Artificial Intelligence Scale. *Computers in Human Behavior Reports*, 1, 100014. https://doi.org/10.1016/j.chbr.2020.100014

*Schepman, A., & Rodway, P. (2022). The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory validation and associations with personality, corporate distrust, and general trust. *International Journal of Human–Computer Interaction*, *39*(13), 2724-2741. https://doi.org/10.1080/10447318.2022.2085400

*Seo, Y.H., & Ahn, J.W. (2022). The validity and reliability of the Korean version of the General Attitudes towards Artificial Intelligence Scale for nursing students. *The Journal of Korean Academic Society of Nursing Education*, *28*(4), 357-367. https://doi.org/10.5977/jkasne.2022.28.4.357

Sindermann, C., Sha, P., Zhou, M., Wernicke, J., Schmitt, H.S., Li, M., ... & Montag, C. (2021). Assessing the attitude towards artificial intelligence: Introduction of a short measure in German, Chinese, and English language. *KI-Künstliche Intelligenz*, 35, 109-118. https://doi.org/10.1007/s13218-020-00689-0

Thompson, B., & Cook, C. (2002). Stability of the reliability of libqual+™ scores a reliability generalization meta-analysis study. *Educational and Psychological Measurement*, *62*(4), 735-743. https://doi.org/10.1177/0013164402062004013

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement*, 60, 174-195. https://doi.org/10.1177/00131640021970448

Turkle, S., Breazeal, C., Dasté, O., & Scassellati, B. (2006). Encounters with kismet and cog: Children respond to relational artifacts. *Digital media: Transformations in human communication*, 120. http://web.mit.edu/people/sturkle/encounterswithkismet.pdf

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58*(1), 6–20. https://doi.org/10.1177/0013164498058001002

Vacha-Haase, T., Kogan, L.R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement*, *60*(4), 509-522. https://doi.org/10.1177/00131640021970682

Vassar, M.A. (2008). Note on the score reliability for the Satisfaction with Life Scale: An RG study. *Soc Indic Res*, 86, 47–57. https://doi.org/10.1007/s11205-007-9113-7

Waliszewski, K., & Warchlewska, A. (2020). Attitudes towards artificial intelligence in the area of personal financial planning: A case study of selected countries. *Entrepreneurship and Sustainability Issues*, *8*(2), 399-420. https://doi.org/10.9770/jesi.2020.8.2(24)

Wallace, K.A., & Wheeler, A.J. (2002). Reliability generalization of the life satisfaction index. *Educational and Psychological Measurement*, *62*(4), 674-684. https://doi.org/10.1177/0013164402062004009

*Wang, H., Sun, Q., Gu, L., Lai, K., & He, L. (2022). Diversity in people's reluctance to use medical artificial intelligence: Identifying subgroups through latent profile analysis. *Frontiers in Artificial Intelligence*, 5: 1006173. https://doi.org/10.3389/frai.2022.1006173

Warrens, M.J. (2014). On Cronbach's alpha as the mean of all possible-split alphas. *Advances in Statistics*. 742863. https://doi.org/10.1155/2014/742863

Yin, P., & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement, 60*(2), 201-223. https://doi.org/10.1177/00131640021970466

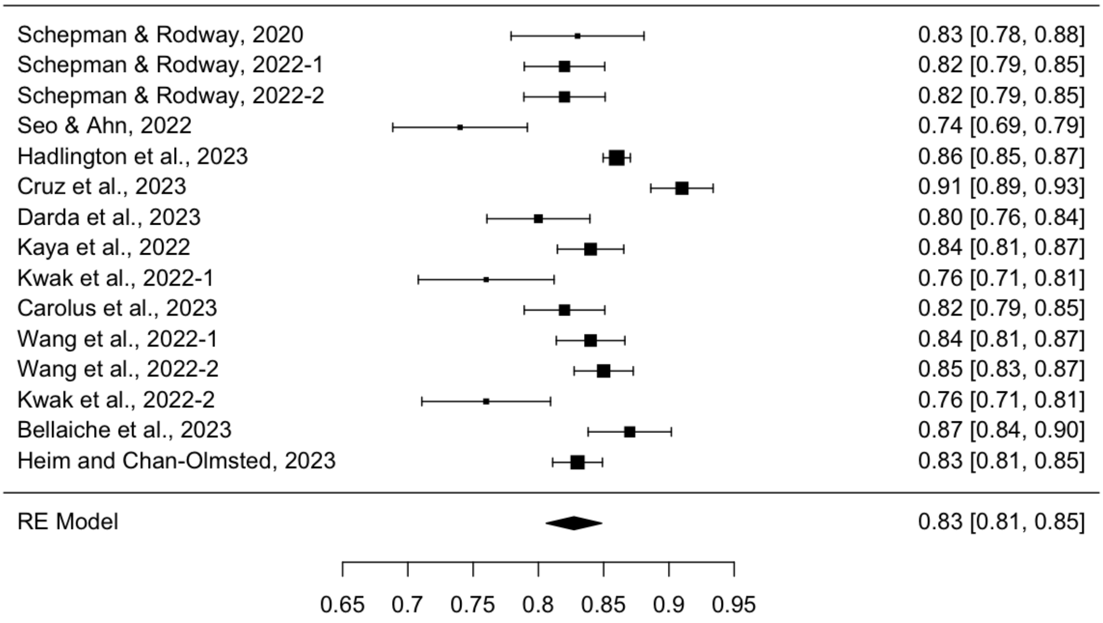
Youngstrom, E.A., & Green, K.W. (2003). Reliability generalization of self-report of emotions when using the Differential Emotions Scale. *Educational and Psychological Measurement*, *63*(2), 279-295. https://doi.org/10.1177/00131644032532

Yoruk, S., & Sen, S. (2023). A reliability generalization meta-analysis of the creative achievement questionnaire. *Creativity Research Journal*, *35*(4), 714-729. https://doi.org/10.1080/10400419.2022.2148073

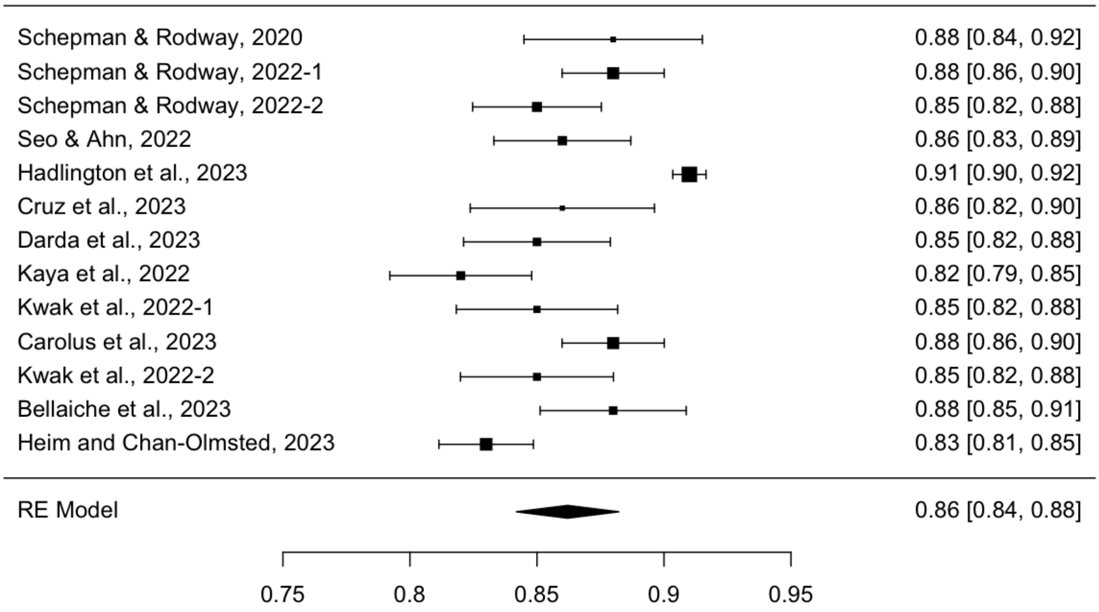
Yuzbasioglu, E. (2021). Attitudes and perceptions of dental students towards artificial intelligence. *Journal of Dental Education, 85*(1), 60-68. https://doi.org/10.1002/jdd.12385

APPENDIX

Appendix 1. *Forest plot for negative subscale (Cronbach's α).*



Appendix 2. *Forest plot for positive subscale (Cronbach's α).*



International Journal of Assessment Tools in Education
2024, Vol. 11, No. 2, 320–344

https://doi.org/10.21449/ijate.1342439

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

Experiences with emergency distance education: A dilemma between face-to-face education and distance education in tour guiding

Meltem Altınay Özdemir [1a,b*], **Zeynep Tombaş** [2]

[1a]University of Algarve, The Research Centre for Tourism, Sustainability and Well-being (CinTurs), Campus of Gambelas, Faro, Portugal
[b]Mugla Sitki Kocman University, Faculty of Tourism, Campus of Kötekli, Muğla, Türkiye
[2]Istanbul Arel University, Vocational School, Türkiye

ARTICLE HISTORY

Received: Aug. 13, 2023

Accepted: Mar. 22, 2024

Keywords:
Online learning,
Distance learning,
Covid-19 pandemic,
Tourism education,
Crisis management.

Abstract: Universities switched from face-to-face to emergency distance education as a solution to the crisis during the Covid-19 pandemic. This enabled face-to-face students to experience distance education. This study examined these experiences for tour guiding education. Distance education was available in tour guiding departments at a few Türkiye universities before the pandemic, and this was a discussion topic in academic. However, emergency distance education presented a dilemma for students: distance education or face-to-face education. Thus, the research includes students in the face-to-face tour guiding departments. A qualitative, phenomenological approach was employed to collect data using semi-structured interviews and an online questionnaire. Data were analyzed using thematic content analysis. The findings revealed that students preferred face-to-face education while being uncertain about the pros and cons of distance education. However, this decision should be underlined as not definitive. The study emphasizes that distance education is ineffective for tour guiding education due to the absence of practical courses, which are crucial for tour guiding, as well as effective communication. The study provides theoretical insights into the educational strategies used in tourism during crises and offers practical implications for enhancing distance education in higher education institutions.

1. INTRODUCTION

The Covid-19 pandemic forced a move to online learning at all education levels in most countries due to the risk of continuing face-to-face education (Masalimova et al., 2022). Emergency distance education (EDE) refers to online education activities due to the Covid-19 pandemic. In Türkiye, for example, the Chairman of the Council of Higher Education announced in a press statement dated March 18, 2020, that all university programs in Türkiye would be conducted via distance education (Saraç, 2020). Like institutions worldwide, universities in Türkiye began distance education in the spring semester of 2019-2020, which continued until the end of that academic year (Durak & Çankaya, 2020a). Then, except for departments requiring applied education, universities under the Council of Higher Education continued with distance education throughout the autumn semester of 2020-2021. Many

*CONTACT: Meltem Altınay Özdemir ✉ meltemozdemir@mu.edu.tr, maozdemir@ualg.pt University of Algarve, The Research Centre for Tourism, Sustainability and Well-being (CinTurs), Portugal

e-ISSN: 2148-7456 /© IJATE 2024

universities also prioritized infrastructure improvements, such as software licenses and personnel recruitment, to be better prepared for distance education activities in the pandemic period (Durak & Çankaya, 2020b). Nevertheless, it cannot be said that universities adjusted swiftly to distance education procedures (Durak et al., 2020).

The EDE reflected the lack of time and opportunity to train teachers or arrange distance education methodically during the pandemic (Toquero, 2020). Hence, the Turkish Council of Higher Education defines EDE as the "temporary transfer of face-to-face education to the technological environment in a crisis" (Turkish Higher Education Quality Council, 2020). In this case, the main goal is not to rebuild a sustainable education ecosystem but to provide temporary access to learning and teaching support that can be easily set up and made available during an emergency or crisis (Bakhov et al., 2021). Despite using similar components, EDE differs from normal online education in terms of terminology and functionality. Turkish universities implemented EDE effectively during the pandemic, which indicates that this innovative concept may grow and spread in the future (Karataş & Tuncer, 2020).

The Tourist Guiding Professional Law (Law 6326) establishes the acceptance requirements for the tourist guide profession in Türkiye. According to the law, there are two ways to meet the requirements. The first is through tour guiding education (TGE) provided by institutions (i.e. vocational schools, and universities). The second is through TUREB's (Turkish Tourist Guides Association) regional and national certification programs for tour guides under the direction of the Ministry of Culture and Tourism. Vocational schools offer both face-to-face and distance-learning TGE programs. As debates on distance tour guiding curricula continue **(**Arıcı & Karaçay, 2023; Köksalanlar & Çözeli, 2021; Yağcı et al., 2019)**,** the EDE has generated dilemmas in TGE. This is because students who receive face-to-face education in tour guiding departments do not perceive distance education favorably (Yağcı et al., 2019). Due to Covid-19, these students had to engage in distance education, allowing them to experience distance TGE's efficiency.

The main purpose of this study is to examine the perceptions and experiences of students studying in face-to-face tour guiding departments regarding EDE, which is compulsory due to the Covid-19 pandemic. The study helps to solve students' dilemma about distance education given that previous studies of EDE in tourism education have identified both advantages and disadvantages of distance education (Choi et al., 2021, Qiu et al., 2021; Munoz et al., 2021; Ritonga, 2022; Ye & Law, 2021) and discussions regarding the adequacy of distance education to provide practical gains in tourism education (Bilsland et al., 2020; Choi et al., 2021; Munoz et al., 2021). The study thus aims to evaluate the EDE for TGE within this framework, considering both theoretical and practical learning. The theoretical justification of data collection tool comes from various previous studies (Agyeiwaah et al., 2022; Arıcı & Karaçay, 2023; Bilsland et al., 2020; Chandra et al., 2022; Choi et al., 2021; Goh, 2020; Köksalanlar & Çözeli, 2021; Munoz et al., 2021; Qiu et al., 2021; Ritonga, 2022; Shyju et al., 2021; Şanlıöz-Özgen & Küçükaltan, 2023; Tavitiyaman et al., 2021; Xu et al., 2022; Ye & Law, 2021; Zhong et al., 2021). The present study contributes to tourism education by revealing students' awareness and perceptions regarding the efficacy of distance education for acquiring qualifications for the tour guiding profession.

This study answers a previous research dilemma: students may learn practical skills offline but should still be aware of technology advances. In this line, tour guiding students who are undecided between face-to-face education and distance education are investigated. Addressed the following research questions:

RQ$_1$: Is EDE sufficient for TGE?
RQ$_2$: What are the perceptions of distance tour guiding education?
RQ$_3$: What are the perceptions of face-to-face TGE?
RQ$_4$: What are the students' perceptions of the advantages of EDE for TGE?

RQ$_5$: What are the students' perceptions of the disadvantages of EDE for TGE?
RQ$_6$: Is distance education practically sufficient for tour guiding?
RQ$_7$: What are the perceived differences between distance education and face-to-face education in TGE?
RQ$_8$: What is the level of students' comprehension of distance education courses?

1.1. Background

1.1.1. *Distance education*

Distance education is an education method based on the internet and interactive technology that enables instructors and students to connect in a real-time setting from different locations (Simonson & Seepersaud, 2019). Newby et al. (2000: 210) define it as "*the teaching program in which teachers and students are physically independent of each other*", while Gunawardena and McIsaac (2013) define it as "*education provided using electronic communication tools at a different time or place than the instructors*". Advances in both business and science are now essential due to the rapid development of information and communication technology. Within this trend, earlier major communication tools of distance education, such as the telephone, television, and audio/video recordings, have become irrelevant (Kim & Jeong, 2018) to be replaced by online learning technologies, such as active learning tools (Kim & Jeong, 2018) and online courses (Qiu et al., 2021). Meanwhile, there is increasing familiarity with platforms like Zoom, Google Meet, and Google Courses, which were widely used during the Covid-19 pandemic (Kapasiaa et al., 2020). The use of these technologies, which have a function at each stage of the education process, is effective in promoting teaching techniques like distance education and online learning. Finally, earlier forms of distance education have been modified by new conceptualizations, such as hybrid combinations of distance learning, flexible learning, distributed learning, and web-enhanced instruction (Gunawardena & McIsaac, 2013: 355).

1.1.2. *Advantages and disadvantages of distance education*

There are both advantages and disadvantages to distance education, which students and instructors can access from different locations (Kim & Jeong, 2018: 120). According to Fojtík (2018:16), the advantages include the opportunity to attend courses at work, study at a time determined by the student, individually plan the studying mode at the workplace and at school, the absence of school every day, and the completion of tasks over the internet. Similarly, Klisowska et al. (2021), list the advantages of time management, the ability to study at the student's own speed, and access to a vast variety of educational materials. However, one of the most important advantages of online education is overcoming physical location barriers (Chandra et al., 2022).

Fojtík (2018:16) lists the following disadvantages of distance: limited communication with instructors and classmates, missing classes and seminars, self-study, occasional escapism from the information that the student records while attending, difficulty in organizing time effectively, and motivation problems. Klisowska et al. (2021) also underline the absence of social connection as well as the need to spend a lot of time in front of a computer, and the lack of direct contact with the instructor. Köksalanlar and Çözeli (2021) emphasize the serious challenge of motivation in distance education. Due to motivational issues, sometimes referred to as reluctance towards the lesson, students frequently put off tasks and struggle with time management because they cannot adapt to distance education, thereby losing interest in the lesson.

From their investigation of tour guiding students' perceptions of distance education, Köksalanlar and Çözeli (2021) found that students have negative perceptions due to the lack of a physical classroom environment, education based solely on study notes that may also be incomprehensible, inability to communicate, technological issues, and failure to understand the course. According to Arıcı and Karaçay (2023:304), the disadvantages of distance education include a lack of motivation, the loss of instructional and socializing roles, and a lack of control over the education process. In addition, there are communication problems, a lack of face-to-face connection, and the requirement for technical support (Korkmaz & Toraman, 2020) while Pesha and Kamarova (2021) state that the primary disadvantages of distance education include restricted communication, the need for additional help for students with difficulties understanding their courses, lack of self-discipline, lack of technological support, and unclear working hours.

1.2. EDE

Distance education is a very important tool during emergencies (Jiang et al., 2021), and the Covid-19 pandemic demonstrated how important it was for higher education particularly (Li & Agyeiwaah, 2023; Qiu et al., 2021). Ideally, distance education and online learning, require planning studies and instructional designs based on theory and models. However, owing to the quick transition to EDE, which Adedoyin and Soykan (2020) identified as one of the migration techniques in the struggle against the crisis, several planning, design, and development shortcomings emerged during the pandemic. Given that EDE implemented during a pandemic differs from traditional distance education (Wang et al., 2020), Adedoyin and Soykan (2020) assert that EDE should not be regarded as effective online learning or the digital transformation of universities. Instead, they suggest examining it through the framework of "*emergency distance education platforms*".

EDE is the temporary transfer of face-to-face education to an online environment during a crisis (Turkish Higher Education Quality Council, 2020). That is, it describes online learning activities implemented in response to the pandemic crisis environment to minimize disruption to the educational process (Sezgin, 2021). EDE initiatives globalized education, with problems like climate change, terrorism, refugee crises, natural catastrophes, and the battle against diseases becoming global issues (Qiu et al., 2021). Furthermore, similar crises will likely arise in the future, so educational institutions are now required to be prepared to respond to emergencies at any moment. For example, since the Covid-19 pandemic, Türkiye has experienced two earthquake disasters centered in Kahramanmaraş on February 6, 2023 (Kandilli Observatory and Earthquake Research Institute, 2023). Thus, EDE has been required in Turkish higher institutions due to both the pandemic and seismic disasters. Although EDE was implemented in all education institutions during the pandemic, it was only done in higher institutions after the earthquakes.

During the Covid-19 pandemic, higher education institutions accelerated the implementation of online EDE courses. These began in March 2020, in the middle of the Spring semester of the 2019-2020 academic year and continued in both semesters of the 2020-2021 academic year. During the 2021-2022 academic year, hybrid education initiatives were increasingly integrated into face-to-face education. While face-to-face education returned in the Fall semester of the 2022-2023 academic year, EDE reemerged as one of the government's disaster management policies, after student dormitories were allocated to earthquake victims following the Kahramanmaraş earthquakes mentioned above. Therefore, universities completed the spring semester of the 2022-2023 academic year with EDE until April, and hybrid and distance education thereafter. In short, since the pandemic, EDE has become a crucial crisis intervention in Türkiye.

1.3. EDE in Tourism Education

As in other sectors, the Covid-19 pandemic damaged tourism education (Ye & Law, 2021; Zhong et al., 2021). The severe restrictions imposed by Covid-19 have made the transition to online hospitality and tourism education an obligation rather than an option (Agyeiwaah et al., 2022: 9). Although the Covid-19 pandemic significantly hindered tourism education (Ye & Law, 2021), many institutions are likely to continue with online courses as part of hybrid education programs if the shortcomings due to the rapid shift to distance education platforms during the pandemic can be resolved (Adedoyin & Soykan, 2020). This would confirm Goh's (2020) prediction that as technology use grows, so will its application to tourism and hospitality education (Ritonga, 2022).

Studies conducted during the pandemic indicate that distance education will become a popular trend in tourism education (Choi et al., 2020; Qiu et al., 2021; Ritonga, 2022). In addition to Tavitiyaman et al. (2021) who reported a sudden migration to distance education in tourism, other studies focus on the advantages of EDE for tourism programs in this migration (Goh & Sigala, 2020; Lei & So, 2021). However, EDE activities implemented outside the norm impacted the method of teaching practice-based courses for tourism (Hsu, 2021). Therefore, various challenges have emerged. One of these is an inability to gain practical skills (Agyeiwaah et al., 2022). Academic institutions play a crucial role in transforming students into qualified professionals with essential skills for the tourism industry (Prifti et al., 2020). However, the pandemic resulted in the virtualization of classroom practical training (Kaushal & Srivastava, 2020; Sharma, 2020), compromising the benefits that students derive from classroom training (Shyju et al., 2021). Even though advanced technologies like virtual tour platforms, provide innovative ways to give application-based information and enhance the learning experience (Patiar et al., 2021), gaps remain in internship training and sector-specific practice courses (Qiu et al., 2021). Consequently, practical training outcomes, which are key components of tourism education, were significantly impacted by the pandemic. Although distance education during the pandemic process assisted tourism students in managing their daily lives, tourism and accommodation education requires a certain level of applied learning, as Kaushal and Srivastava (2020) emphasized in their study of tourism students in India. Similarly, Choi et al. (2020) believe that offline education is vital for students to obtain practical experience in the tourism industry. With the transition from traditional to creative evaluation, however, application training criteria may change in response to the pandemic (Qiu et al., 2021). Another advantage that EDE revealed is that tourism students can work part-time or full-time in the tourism industry. That is, online learning allows students to continue their education while meeting family and professional obligations (O'Connor, 2021).

Previous evaluations of EDE show that tourism students found their online courses to be clear, organized, practical, and fluent (Agyeiwaah et al., 2022). Although the virtual format presents some technological challenges, both students' and instructors' computer proficiency is growing (Hodges et al., 2020). Additionally, tourism students claimed to be ready for online learning and using the internet and technological devices (Poláková & Klímová, 2021). Given that students also need the knowledge and skills regarding widely used technology in the tourism industry (Xu et al. 2022), distance education has demonstrated, the need for tourism students to have essential technology-related equipment (Bucak & Yigit, 2021).

1.4. Overview of Studies on EDE in Tourism Education

Numerous studies have been conducted on use of EDE in tourism education due to Covid-19, focused on students' online experiences (Agyeiwaah et al., 2022; Munoz et al., 2021), perceptions (Arıcı & Karaçay, 2023; Korkmaz et al., 2023; Köksalanlar and Çözeli, 2021; Tavitiyaman et al., 2021), satisfaction (Chandra et al., 2022; Choi et al., 2020; Choi et al., 2021; Li, & Agyeiwaah, 2023; Shyju et al., 2021), and psychological situations (Tavitiyaman et al., 2021; Zapata-Cuervo et al. al., 2023; Zhong et al., 2021). Other studies have focused on EDE's effectiveness (Qiu et al., 2021; Patiar et al., 2021; Ritonga, 2022; Ye, & Law, 2021), the future of tourism education (Xu et al., 2022), and instructors' experiences with EDE (Şanlıöz-Özgen, & Küçükaltan, 2023). Agyeiwaah et al. (2022) claim that Covid-19 seriously disrupted pedagogical practices. They also emphasise that educational institutions that instruct students in the field of hospitality and tourism should design online course presentations in a visually appealing and encouraging environment. Arıcı and Karaçay (2023) found that despite problems with technical support and communication at their universities, tour guiding students are satisfied with the advantages of online education, such as convenience and low cost. Chandra et al. (2022) point out the importance of practical lessons and on-site training. To meet industry expectations for student employability, they emphasize the need for efficient tools and curricular adjustments. According to Choi et al. (2021), blended education should be considered to support learning if online learning is to be successful. They also emphasize that communication between faculty and students continues to be a key factor for success. Choi et al. (2020) also state that improvements in online learning are achieved when stronger relationships are established between instructors and students. Additionally, Kaushal and Srivastava (2020) noted sectoral concerns about the practical benefits of accommodation and tourism education. According to Korkmaz et al. (2022), although tourism students have favorable perceptions of distance education, they prefer to attend classes face-to-face. In addition, they discussed the disadvantages of distance education, including isolation from the social environment, technical issues, and the difficulty of communicating with the instructor. Köksalanlar and Çözeli (2021), in one of the few studies on tour guiding education during the Covid-19 period, reported that students perceive distance education negatively due to a lack of one-on-one education and classroom environment, inability to communicate, internet problems, and lack of technical tools like computers. They also found that most students were unwilling to study, unable to concentrate, and disengaged from their courses and school. On the other hand, some students evaluated distance education positively due to factors like convenience, accessibility, and efficient use of time. O'Connor (2021) investigated the active learning methodologies used in higher education travel and tourism programs in Ireland. They highlighted the significance of applied learning in bridging the gap between academia and industry, where students learn to perform properly.

Patiar et al. (2021) evaluated the function of the Virtual Field Trip (VFT) platform for meeting practical skills in online education. They concluded that VFT provides a technology-enhanced option for acquiring employability skills. Qiu et al. (2021) recommends the internationalization of online tourism education given that any country may face the problem of how to address crises like climate change, terrorism, refugee flows, and natural disasters. They suggest internationalizing by diversifying platforms, internationalizing the curriculum, internationalizing professors, and internationalizing students. According to Amin et al. (2022), motivational factors are important in e-learning. The quality of e-learning impacts both student competency and satisfaction. Kallou and Kikilia (2021) call EDE as "transformative" and state that "The latest Covid-19 pandemic developments have led to a new perspective of education through digital technologies, changing how universities perceive the teaching and the learning process" (p.37). Finally, Justin et al. (2022) examined students' online learning experiences and found that, although they agree that online learning makes their work life easier, they prefer to attend in-person classes.

2. METHOD

A qualitative, phenomenology research design was adopted for this study. Phenomenology refers to the conscious experience of a person's own life environments (Schram, 2003:71). That is, it studies experience or consciousness structures and examines the structure of perception, cognition, memory, imagination, emotion, desire, volition, physical awareness, embodied action, and social interaction. It examines conscious experience from the first-person point of view as well as the conditions of experience that are important to those structures (Smith, 2018). Phenomenology is a popular approach in the social sciences because it allows individual experiences to be studied (Merriam, 2018). In this line, this study examines the EDE experiences of the students to evaluate the efficacy of distance education for tour guiding.

2.1. Sampling Design

Purposive sampling was preferred for "obtaining in-depth information about specific attributes of the person, event, or situation most appropriate to answering the research questions" (Maxwell, 2012: 97). The sample comprised students registered in face-to-face tour guiding departments at universities in Istanbul. The sample selection criterion was to have experience of at least one semester in EDE (hybrid education or distance education) applied during the Covid-19 pandemic, as distinct from normal learning processes. There are also distance education programs in Türkiye, mainly in Istanbul. Thus, they were not included in the research. The study was conducted with 81 students registered in face-to-face tour guiding departments in Istanbul universities during the academic year 2022-2023 (Table 1).

2.2. Data Collection

The data were collected during the Fall and Spring semesters of the 2022-2023 academic year, with approval from the university ethics committee, (Istanbul Arel University Ethics Committee's decision dated 06 June 2022, numbered Istanbul 2022/10). Data were collected using semi-structured interviews and an online questionnaire. The questionnaire consisted of thirteen questions (four demographic questions and nine TGE questions). Three of TGE questions were close-ended, while the remaining six were open-ended. The interview questions were adapted for tour guiding education from previous studies of EDE in tourism education during the Covid-19 pandemic (Agyeiwaah et al., 2022; Arıcı & Karaçay, 2023; Bilsland et al., 2020; Chandra et al., 2022; Choi et al., 2021; Goh, 2020; Köksalanlar & Çözeli, 2021; Munoz et al., 2021; Qiu et al., 2021; Ritonga, 2022; Shyju et al., 2021; Şanlıöz-Özgen & Küçükaltan, 2023; Tavitiyaman et al., 2021; Xu et al., 2022; Ye & Law, 2021; Zhong et al., 2021). Merriam (2018) suggests using triangulation and participant confirmation to assure the internal validity, reliability, and generalizability of qualitative research, particularly when based on an interpretive paradigm. Hence, a "confirmation email" was forwarded to all participants, whose e-mail addresses were acquired with their permission, to ensure participant confirmation and scope validity in the study, after receiving their responses to confirm their responses. To achieve triangulation, the data were validated by two researchers.

2.3. Data Analysis

The data were analyzed by thematic content analysis. Maxqda software was used to compile and code the data, create the main and sub-themes, define the code frequencies, and determine the relationships between the codes. A descriptive research design was adopted to determine the key themes underlying the students' experiences of EDE in TGE. The relationships between the main and sub-themes were examined through code relationship analysis, a code map, and complex code configuration analysis. Seven key themes were identified*: (1) perception of face-to-face education in TGE, (2) perception of distance education in TGE*, (3) *Difference between face-to-face and distance education in TGE*, (4) *Sufficiency of EDE in TGE*, (5) *Practical sufficiency of distance education in TGE*, (6) *Disadvantages of EDE*, (7) *Advantages of EDE*. The themes were determined based on studies of pre-pandemic tour guiding education via distance learning (Yağcı et al., 2019) and tourism education during the pandemic period (Arıcı & Karaçay, 2023; Köksalanlar & Çözeli, 2021; Agyeiwaah et al., 2022; Chandra et al., 2022; Choi et al., 2021; Goh, 2020; Qiu et al., 2021; Shyju et al., 2021; Şanlıöz-Özgen & Küçükaltan, 2023; Tavitiyaman et al., 2021; Xu et al., 2022; Ye & Law, 2021; Zhong et al., 2021). In the following sections, while interpreting the findings, representative statements are quoted in accordance with the qualitative research writing principle of "identifying expressions that symbolically represent a subject and frequently indicate the opinions of other participants with similar perceptions".

3. FINDINGS

A total of 6089 words were evaluated using the software program. Word frequency analysis revealed there are 306-word groups. The most frequently repeated words were “more” (115), “formal” (66), “sufficient” (52), and “distance” (47).

3.1. Sample Profile

The research participants were students registered in tour guiding departments at three universities in Istanbul. Over half were female (56%), single (77%), and between the ages of 18 and 33 (64%). Most participants were associate students (80%) (Table 1).

Table 1. *Sample profile**.

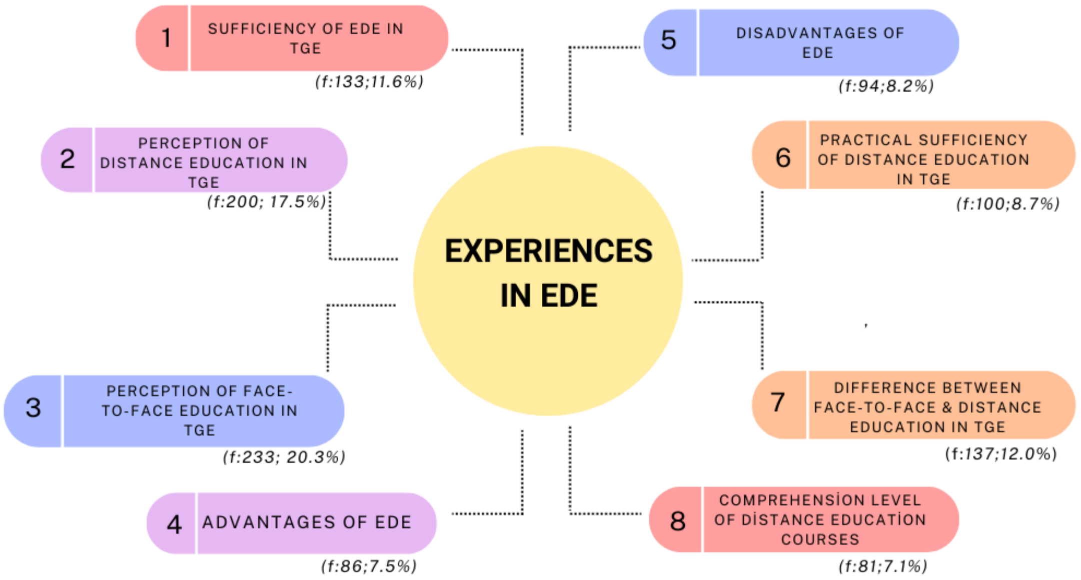
Gender	*f* (81)	%	*Marital status*	*f*	%
Female	46	56.79	Single	60	74.07
Male	35	43.21	Married	21	25.93
Age			*Education level*		
18-25	30	37.04	Associate student	65	80.25
26-33	22	27.16	Undergraduate student	16	19.75
34-41	14	17.28			
42-49	10	12.35			

* Information for all participants is given in Appendix1.

3.2. EDE Experiences in TGE

As the code system in Figure 1 shows, EDE experiences in TGE were divided into eight main themes. We coded educational level as an additional main code. Therefore, all main codes and 31 sub-codes total 1,145 codes. According to the super-code results, the codes with the most frequency were the perception of face-to-face education in TGE (20%), the perception of distance education in TGE (17.5%), the difference between face-to-face and distance education in TGE (12%) and the sufficiency of EDE in TGE (11%). The students’ experiences mostly centered on these four themes.

Figure 1. *Experiences in EDE.*


EXPERIENCES IN EDE
1 SUFFICIENCY OF EDE IN TGE
(f:133;11.6%)
2 PERCEPTION OF DISTANCE EDUCATION IN TGE
(f:200; 17.5%)
3 PERCEPTION OF FACE-TO-FACE EDUCATION IN TGE
(f:233; 20.3%)
4 ADVANTAGES OF EDE
(f:86;7.5%)
5 DISADVANTAGES OF EDE
(f:94;8.2%)
6 PRACTICAL SUFFICIENCY OF DISTANCE EDUCATION IN TGE
(f:100;8.7%)
7 DIFFERENCE BETWEEN FACE-TO-FACE & DISTANCE EDUCATION IN TGE
(f:137;12.0%)
8 COMPREHENSİON LEVEL OF DİSTANCE EDUCATİON COURSES
(f:81;7.1%)

Notes: *Total codes are f:1145; 100% including education level (f:81; 7.1%); TGE: Tour guiding education; EDE: Emergency distance education*

Appendix 2 presents the students' perceptions of distance education in TGE. The sub-codes are ranked from the highest to the least frequent. The most mentioned sub-codes were "Perceptions of strengths of face-to-face education", "Neutral perceptions of face-to-face education" and "Perceptions of weaknesses of DE)". The students predominantly concerned perceptions of the strengths and general characteristics of face-to-face TGE and the weaknesses of distance TGE. They stated that distance education and face-to-face education contribute differently to successful learning.

3.2.1. *Sufficiency of EDE in TGE*

The students stated that EDE was not sufficient for TGE because of insufficient vocational courses (46.1%), motivation problems (21.1%), and lack of effective communication (32.2%) (RQ1).

3.2.1.1. Insufficient for Vocational Courses. The students claimed EDE did not provide the qualifications for the tour guiding profession. Due to the theoretical importance of the courses, they retained knowledge more effectively in face-to-face than in distance courses. They also underlined that this knowledge should be supported by field trips: "*I think that courses should be put into practice and that verbal education is better when it is done face-to-face*" (P1); "*A program that needs to be supported by field studies/trips*" (P28).

While the course's conceptual framework and the instructor's skills are important for the students, tour guiding departments must include field trips to provide practical training, as mentioned by various participants: "*Some courses require practice and a field trip*" (P34); "*I don't believe that distance education will allow us to learn this profession effectively. We must see it with our own eyes, touch it, and experience it since this is not a virtual profession*" (P66).

Two key skills required in the tour guiding profession are the ability to communicate with others and the ability to use at least one foreign language. Neither of these skills can be obtained solely through distance education: "*I believe that the best way to improve at learning foreign languages is to have a face-to-face education that emphasizes practice*" (P66). P16 offered the following explanation:

> *In certain courses, regardless of the quality of the instructor, the course content demands physical presence in the classroom or on the trip. In the case of tourist guiding, distance education will not provide successful practice tours or classroom presentations. Presentations in the classroom can help students express themselves in front of a group.*

3.2.1.2. Lack of Effective Communication. Given that tour guides are extroverts with effective communication skills. The students highlighted the limitations of EDE in providing this:

> *Distance education may be beneficial for some courses, but it is preferable to have practical courses. To practice speaking, storytelling, and conversation in crowded environments, face-to-face education is essential in tour guiding.* (P8).

Additionally, tour guides need to be able to express themselves well, make a good impression, and communicate both verbally and nonverbally (P3):

> *The profession of tour guiding is narrative-based. Face-to-face schooling allows us to study mimicry, posture, expression style, and how teachers control their body language when teaching. One-on-one classes with our instructors and questions, ideas, opinions, and discussions are more productive. Distance education cannot do this*.

In comparing the advantages of face-to-face education to EDE, the students claimed that the latter was insufficient. They stressed how crucial instructor-student connection and communication are to the course's efficacy:

> *I consider that EDE needs additional instructor-student engagement. Although synchronous-asynchronous courses are possible in EDE, I believe that face-to-face education is more effective. (P26).*

Emphasizing the theoretical lessons of tour guiding departments, the students emphasized that it would be more productive to have face-to-face courses because of the EDE's interaction problem:

> *Tourist guiding communication should be high quality; however, EDE communication is virtually nonexistent. Distance education is insufficient to better comprehend courses such as Anatolian Civilizations and Art History, to share information, and to ask questions (P36).*

3.2.1.3. Motivation Problems. EDE activities in tour guiding departments tend to have low student concentration and motivation. Students said they were not successful because they were not motivated to attend class due to hardware challenges (P75, P70, P56, P52): "*Technical issues and abstractness prevent me from focusing on the course*. (P52)"; "*I can't study because I'm sleepy*. (P70)"; "*I'm unable to be productive, ... I can't pay as much attention as I can face-to-face* (P71)". An unexpected finding was that working students claimed that the classroom environment is preferable to that in distance education, because of the students' difficulty adapting to the courses owing to a lack of motivation (P2, P3, P32, P56, P65, P68): *"I'm not sure whether distance education is sufficient and worthwhile after a hard day of work. I believe that face-to-face education is more beneficial"* (P65).

3.2.2. *Perception of distance education in TGE*

The students were asked to list the first five words that came to mind when considering distance TGE. This word association test showed how the students think about distance TGE. The words most frequently given primarily related to perceptions of weaknesses, related to as RQ2 (35.4 %), although strengths (32.0%) were also highlighted.

Regarding the weaknesses of distance TGE, the students mainly mentioned attention problems, equipment deficiencies and socialization problems. For example, they used words like "boring, incomplete, insufficient, lack of communication" (P36), "connection problem, voice delay" (P51), "boring, carelessness, indifference" (P3), "lack of communication, harmony problem, solidarity, lack of understanding" (P14), "lack of focus, bad voice" (P56), "abstract, inattention, inadequacy" (P52), "inefficiency", "inability to perceive" (P78), "inefficiency", "inability to socialize" (P68), "connection problem" (P66), and "antisociality" (P34).

Regarding the strengths of distance TGE, the students most frequently addressed being economical and offering some conveniences: "savings" (P61), "low cost" (P22), "cheap" (P10), "time-saving, fast access, planned" (P20), "location independence" (P12), "comfort, fast communication, time-saving" (P73), "flexible" (P23), "risk-free, easy, re-watchable, accessibility to resources" (P33), "practical, placeless" (P39) and "practical, useful" (P44).

Most of the word association responses related to the technological abilities of distance education. Because perceptions and attitudes are not determinative, these words were evaluated as neutral perceptions. Examples included "culture, history, tourism, travel, art" (P48), "computer" (P69), "internet, computer (P11)", "icons, Greek and Roman sculpture art, ancient city, neolithic" (P19), "art (P81)", and "online education, presentation, zoom, connection" (P62).

To examine the co-occurrence of perceptions regarding the strengths and weaknesses of distance TGE, the code relations browser was examined. This indicated a strong relationship between the two sub-themes, with 135 concurrences. That is, the students mentioned both strengths and weaknesses while expressing their cognitive perceptions of distance TGE.

3.2.3. *Perception of face-to-face education in TGE*

When the students were asked to write the first five words that came to mind about face-to-face education in TGE, about half of their responses focused on the strengths of face-to-face education (51.9%) as well as their neutral perceptions (37%), and weaknesses (10.3%). These findings help answer RQ3.

Regarding face-to-face TGE's weaknesses, "way" (P63), "expensive" (P33, P79), and "waste of time" (P20, P43) were used, while "ease of communication" (P12), "socialization" (P34), "sincerity" (P55), "motivation" (P39), "healthy education" (P39), "efficiency" (P42) and "interaction" (P26) were used in association with strengths. Finally, neutral perceptions were expressed through words like "education" (P48), "knowledge" (P51), "school" (P60), "class" (P48), and "book" (P8).

3.2.4. *Advantages of EDE for TGE*

In relation to RQ4, the students identified five main advantages of EDE for TGE: Effective time management (46.5%), compensation (18.6%), independence from location (11.6%), savings (5.8%), and ease of access to materials (4.6%).

3.2.4.1. Effective Time Management. Several students found distance education advantageous particularly those caring for families: "*I can work and take care of my family, and I can also attend classes; I can do both*" (P80). Other students noted how they save time by avoiding transportation problems: "We can manage our time more efficiently by avoiding Istanbul's traffic" (P73). Finally, 46.5% of the participants gave effective time management as EDE's greatest advantage.

3.2.4.2. Compensation. According to 18.6% of the responses, distance education gives more chances to repeat courses and compensate. One significant advantage, for example, is the ability to watch recordings of missed courses and revise poorly understood material: "*We can watch the course's record anytime we like*" (P9, P26); "*Because the courses are recorded, if a course is missed due to force majeure, the missing parts are readily finished, and a more productive working environment is attained through repetition…"* (P30).

3.2.4.3. Independence from Location. Studying regardless of their location via the Internet was another significant advantage for 11.6 % of the participants: "*You don't need to go, you can receive a diploma from home, anywhere*" (P18); "*Education from anywhere*" (P39).

3.2.4.4. Savings. Another advantage of EDE due to its independence from location is savings, particularly transportation costs: "m*inimizing unneeded travel costs*" (P7); "*eliminating travel costs*" (P20).

3.2.4.5. *Ease of Access to Materials.* The final advantage of distance education mentioned was easier access to course materials and course records: "*Everyone has access to course materials*" (P33); "*Students have faster access to more resources for self-training*" (P41); "*Courses are videotaped weekly*" (P8).

3.2.5. *Disadvantages of EDE for TGE*

In relation to RQ5, the students identified five main disadvantages of EDE for TGE: Inappropriacy for the TGE (29.7%), lack of motivation (23.4%), lack of communication (18%), technical problems (9.5%), and poor course attendance (6.3%).

3.2.5.1. EDE's Inappropriacy for TGE. Because tour guiding is an interactive profession based on practice, the students wanted to learn not only theoretical information but also how it is used in the field. As P30 put it:

> *Due to inadequate practice, the education at the associate, undergraduate, and graduate levels in our country is insufficient. I think it would be useful to share information and teach students to utilize it in the field. Distance education isn't enough; field education is needed.*

The students believe that distance education is inappropriate because tour guiding is based on conversation and engagement. Distance education is thus deficient in terms of learning how to talk in front of a group and acquiring expressive abilities: "*It produces a lack of experience for some courses, and students avoid the communication skills required for the guiding profession from the start*." (P34); "*Not being able to make trips, having trouble speaking in front of the group" (P16); "Since tour guiding is all about communicating with people, it's not a good idea to teach lessons without ever seeing anyone or talking to them"* (P37).

3.2.5.2. Lack of Motivation. A primary disadvantage of distance education identified by the students was motivation. They claimed that they did not attend courses, particularly because they were unable to pay attention and that, even when they did listen, they had trouble understanding the subject. The statements, respectively, are as follows: "*There are situations in which we do not comprehend what we are listening to as a result of our negligence and haphazard attendance at the course*" (P52); "*Loss of attention, low motivation and lack of interest*" (P40); "*The most serious disadvantage is the difficulty in comprehending courses and obtaining information*" (P24).

3.2.5.3. Lack of Communication. A few participants (P65, P56, P48, P46, P42, P9, P2) listed, a lack of communication as an additional disadvantage. More importance should be given to the communication process between instructor-student and student-student in distance education. P42 suggested the lack of feedback, which is the most essential aspect of effective communication, as another problem. Furthermore, synchronous courses are challenging even though communication is simultaneous (P56, P46, P16, P9): "*Sociability and productivity become less. There is a problem in one-to-one communication with the instructor*" (P65); "*The rate of feedback about whether the student has received the information is poor*" (P42); "*I may claim that face-to-face education is more conducive to the expression of ideas, whereas distance education is predominantly unidirectional and restricts student participation*" (P16).

3.2.5.4. Technical Problems. The internet and information communication technologies are key components for effective distance education. They are the most essential elements for ensuring effective communication, engaging coursework, and course motivation: "*I cannot take classes because the internet is bad*" (P79); "*Courses are not effective due to internet problems*" (P57); "*Technological and hardware problems can negatively affect communication*" (P39).

3.2.5.5. Low Attendance to Courses. In addition to motivation problems, poor course attendance has detrimental effects on distance education students. This may be exacerbated by the lack of attendance requirements at some universities and the flexibility of the distance education process: "*Lack of classroom environment, no obligation to attend classes*" (P33); "*Insufficient attendance in the course. In contrast to face-to-face education, the instructor and students become unmotivated when there are few participants*" (P30); "*During the course, there's not enough involvement*" (P61).

3.2.6. *Practical sufficiency of distance education in TGE*

Regarding RQ6, nearly two thirds of the students (59.2%) considered that distance education provides inadequate practical training for the tour guiding profession. The students who claimed that distance education is inappropriate for TGE also stated that fieldtrips are essential for tour guiding courses (P9, P13, P25, P29, P40, and P73). They stated that face-to-face education activities should be prioritized over distance education activities in developing the expressive abilities of tour guides, utilizing the information in the field, and ensuring its sustainability:

> *Because a tour guide must go to a site that is discussed in class, experiencing it in the context of that lesson always makes it more memorable. As a way of preparation for the profession, we may test it out for ourselves by telling our other friends the information we gained in the class*. (P14)

> *As field-specific education is necessary, I do not think distance education is useful, but insufficient in and of itself*. (P30)
>
> *This is not a profession that can be attained through distance learning, but it is quite challenging anyway. This profession must be learned by sight, sound, and touch*. (P66)
>
> *Along with comprehensive education, it's important to teach students how to behave and how a tourist guide should behave, and practices should be prepared for them to conduct tour guides*. (K27)

3.2.7. *Differences between face-to-face & distance education in TGE*

In comparing face-to-face education and distance education for TGE to address RQ7, most. Most of the students (81.4%) claimed that they differ from one another. These differences were attributed to efficiency (57.1%), socialization (14.2%), concentration (14.2%), unidirectionality (8.9%), and self-expression (5.3%).

3.2.7.1. Efficiency. The most frequently mentioned difference is efficiency, with face-to-face education being considered more efficient than distance education. The students identified various advantages of face-to-face education, including encouraging participation in the course (P38), focusing on the course better (P77), providing opportunities for socialization (P5), making effective use of body language in communication (P71), increasing the permanence of information (P66), and making communication easier (P44). They also recognized that experience sharing (P5, P24) is possible in face-to-face education and that students pay attention to this: *"It is easier to share knowledge with faculty members and other students in face-to-face education*" (P12); "*In terms of comprehension and involvement, face-to-face education is more effective*" (P57); "*Sharing experience, socializing*" (P5). Regarding efficiency, the students criticized distance education in various ways: "*Not benefiting from the experience of other students. The distance education student makes an extra effort in terms of acquiring information*" (P24); "*Lack of communication, lack of socialization*" (P48).

3.2.7.2. Socialization. Socializing is important for TGE students, especially during distance education, because it is a social profession. Hence, the students noted this: "*You cannot socialize; this is the most important problem*" (P68); "*Class communication can be established more healthily in face-to-face education*." (P44).

3.2.7.3. Concentration. Face-to-face education makes it easier for students to pay attention to the lessons, for example through the instructors' use of body language. Several students stated that they paid more attention in face-to-face lessons: "*Face-to-face education allows easier idea sharing and concentration*" (P18); "*We can pay more attention in face-to-face lessons*" (P72).

3.2.7.4. Unidirectionality. Another difference is that face-to-face education provides a two-way communication process, whereas distance education usually presents a one-way one. Hence, in face-to-face education (P8), it is easy for students to ask questions and engage in discussions whereas in distance education, students only concentrate on listening to the lesson which forces them to participate in a tedious process (P36): "*Online education is very simple and one-way*" (P29); "*We watch it [the lesson] during distance education as though we were watching a documentary by ourselves. It eventually becomes boring*" (P36).

3.2.7.5. Self-expression. Students claimed that the two modes differ in providing opportunities to express themselves. In face-to-face education, they express themselves in front of the group whereas in distance education, they do so on a computer (P37). In addition, the restricted duration of distance education lessons means that students cannot effectively express themselves successfully during the course: "*Students talk in front of the public in face-to-face education and in front of the computer without seeing anyone in the other*." (P37); "*Insufficient involvement in the lesson due to the lesson's limited duration*" (P33).

3.2.8. *Complex code configuration*

The relationships between the codes were determined by complex code configuration analysis, which shows the strengths of the relationships and correlations between the two codes and their subcodes (Maxqda, 2021). The intersection code-subcode frequencies define the level of the relationship between two independent codes. The complex code configuration analysis revealed 81 relationships in 10 combinations between students' comprehension level of distance courses and their educational levels (Table 2). Regarding RQ8, most (88.8%) of students reported that they could understand the distance courses and associate students rated higher than undergraduate students.

Table 2. *Educational degree & perceived understanding of distance courses.*

	f	%
Associate student + Extremely high understanding (5)	18	22.2
Associate student + Very high understanding (4)	18	22.2
Associate student + Moderate understanding (3)	17	20.9
Associate student + Slight understanding (2)	9	11.1
Undergraduate student + Extremely high understanding (5)	7	8.6
Undergraduate student + Very high understanding (4)	4	4.9
Undergraduate student + Understanding not at all (1)	3	3.7
Associate student + Understanding not at all (1)	3	3.7
Undergraduate student + Sligh understanding (2)	1	1.2
Undergraduate student + Moderate understanding (3)	1	1.2
Total	81	100

Note: 1: Understanding not at all, …, 5: Extremely high understanding

4. DISCUSSION and CONCLUSION

4.1. A Dilemma in TGE

Based on the perceptions of Turkish TGE students studying EDE courses, distance education is not sufficient for TGE, particularly due to insufficient vocational courses, lack of effective communication, and motivation problems. The displacement effect caused by the Covid-19 pandemic apparently reduced student motivation and impaired the learning process (Prifti et al., 2020). Other studies emphasize that students were not able to adapt due to motivation problems (Koksalanlar & Çözeli, 2021; Arıcı & Karacay, 2023; Fojtík, 2018; Klisowska et al., 2021; Davis et al., 2019). Meanwhile, ineffective communication leads students to think that distance education is ineffective (Ye & Law, 2021). As Goh and Wen (2020) point out, while distance education permits instructor-student communication, it generates some communication challenges, including the psychological distance that online communication techniques produce between people (Darke et al., 2016). Hence, the students in the present study frequently highlighted the advantages of face-to-face education in TGE, particularly as being more appropriate for the tour guiding profession. This confirms previous findings (Arıcı & Karaçay, 2023) that students prefer face-to-face education over distance education.

Yet, despite preferring face-to-face learning to online learning in EDE, the students in our study also acknowledged some advantages of distance education, particularly effective time management, compensation opportunities, independence from location, savings, and ease of access to materials. Nevertheless, in line with previous studies (Arıcı & Karaçay, 2023), it is notable that these advantages have no significant impact on learning satisfaction. It should be highlighted at this point that distance education is especially advantageous for tourism students, who generally take part-time jobs to gain experience in the industry. Distance education allows them to schedule their personal and professional lives alongside their academic studies (Choi et al., 2020). Hence, tourism students tend to prefer asynchronous courses to synchronous courses (Arıcı & Karaçay, 2023; Sitosanova, 2021).

The participants in our study identified a number of disadvantages of EDE inappropriacy for TGE: lack of motivation, lack of communication, technical problems, and low attendance. Except for low attendance, these findings mirror the disadvantages reported in previous studies (Arıcı & Karaçay, 2023; Korkmaz & Toraman, 2020; Köksalanlar & Çözeli, 2021; Pesha & Kamarova, 2021). Regarding attendance, the students in our study stated that they were unwilling to attend synchronous courses if participation was low, which may reflect the importance that students attach to information sharing and correspondence in the online classroom (Munoz et al., 2021). Tour guides must be receptive to communication, social skills, presentation, speaking skills like body language, voice, language and diction, and creative skills like creating and telling stories. In addition, they should be passionate about the region and subject they are describing (Çolakoğlu et al., 2014: 147-154). Student preference in face-to-face programs for tour guiding to be in the classroom social environment supports this finding (Arıcı & Karaçay, 2023; Köksalanlar & Çözeli, 2021).

A number of differences between distance education and face-to-face education in TGE were identified through student experiences in EDE, namely efficiency, socialization, concentration, unidirectionality, and self-expression. These indicate that face-to-face education is more effective than distance education for TGE. Similarly, Arıcı and Karacay (2023) found that students considering EDE preferred face-to-face learning. Socialization and effective communication are very important for tourist guiding, which is a social profession. Therefore, using body language in face-to-face education helps to support communication and maintain students' attention during lessons (Nambiar, 2020).

While students identify attention issues (Köksalanlar & Çözeli, 2021), device deficiencies (Cao et al., 2020), and socialization challenges (Klisowska et al., 2021) associated with distance tour guiding education as weaknesses, they also note that it is cost-effective and has certain strengths. Computer opportunities and motivation are crucial for success in distance education, according to İbicioğlu and Antalyalı (2005). Similarly, Köksalanlar and Çözeli (2021) assert that students' negative perceptions of distance education are influenced by factors like technical problems, lack of motivation, and separation from peers. Yılmaz and Güven (2015) found that students believe distance education is an ineffective, monotonous, and expressionless form of education. On the other hand, distance learning can provide flexibility and convenience (Dumford & Miller, 2018; Zaveri et al., 2020), and be more affordable in terms of accommodation and travel expenses (Bączek et al., 2021).

Research into the Covid-19 pandemic period showed that distance learning can impair student concentration (Bakhov et al., 2021; Lamanauskas et al., 2021; Vlassopoulos et al., 2021). No matter how simultaneous teacher-student communication is in online learning (Poláková & Klímová, 2021), communication is predominantly unidirectional, especially in asynchronous courses. However, if two-way communication between teachers and students can be achieved, then video-based online learning appears appropriate (Shim & Lee, 2020). Students in distance education, contrary to what Duman and Gencel (2023) argue, are unable to express themselves sufficiently due to limited course time. Akti Aslan et al. (2021) revealed that limited course duration is a communication problem for instructors. Although distance education has been shown to help students express themselves (Lamanauskas et al., 2021), the students in the present study reported problems in doing so. Regarding understanding of the material, the sample in our study primarily comprised associate students. Nevertheless, the code relationship analysis showed that most students were able to understand their distance education courses. In line with previous research (Mulyanti et al., 2020), the students in the present study experienced a dilemma regarding distance education despite its disadvantages. Studies on tourism during the pandemic predict that distance learning will expand and that its beneficial aspects will predominate in the near future (Lei & So, 2021; Korkmaz et al., 2022; Şanlıöz-Özgen, & Küçükaltan, 2023).

4.2. Conclusion

This study makes important theoretical contributions and has practical implications for EDE in tourism education. Due to the Covid-19 pandemic, students enrolled in face-to-face programs were forced to experience fully online learning through EDE, thereby gaining experience in both modes. This created a dilemma for them between face-to-face education and distance education. During the pandemic, online education activities were described as EDE because they were implemented without following all the required distance education procedures. The difference from normal is made clear by the term "emergency". Although it provided a rapid solution in a crisis, distance education may have negative effects on outcomes in some programs. Accordingly, this study examined the experiences of students in tour guiding departments—which are based on both theoretical and applied courses.

The findings indicate that tour guiding students prefer face-to-face education, but their indecision about the advantages and disadvantages of distance education and their high level of understanding of distance education courses indicate a dilemma. In fact, if EDE is extended, it may be possible for them to have more beneficial experiences. However, while these students reported positive cognitive perceptions of face-to-face TGE, they had negative perceptions of distance education. While distance education enables effective time management, it may not be appropriate for TGE. Hence, the students tend to prefer face-to-face TGE for its efficiency. The study also found that despite being aware of the benefits of distance education, the students still prefer face-to-face education because it gives them more opportunities to practice speaking, interact with others, and express themselves verbally. Karadağ and Yücel (2020) also found that social science students are less satisfied with distance education than science and health science students. We can therefore conclude that students in the tour guiding department, which falls into social science, need more communication and interaction opportunities in their courses.

4.3. Theoretical Contribution

This study examined the attitudes of university students in Türkiye's face-to-face tour guiding departments toward EDE, which they experienced due to the Covid-19 pandemic. Hence, the study contributes to the EDE literature. The study provides evidence for an assessment of the distance education process from the perspective of students through their face-to-face education experience. Although the profession of tour guiding is primarily based on theoretical knowledge, there is also a need for practical activities. According to Lei and So (2021), and Goh and Sigala (2020), students had an advantage during EDE. However, Agyeiwaah et al. (2022) claim that they experienced difficulties such as the inability to learn practical skills to improve classroom learning. Although there are inequalities in the tourist guiding profession at both associate and undergraduate degree levels in Türkiye (Eser & Şahin, 2020), the main aim is to train qualified guides (Eker & Zengin, 2016). Aside from EDE, previous studies have discussed the need to support TGE with short practice trips (Eker & Zengin, 2016) and tour guides have similar perceptions (Eker, 2015). These studies have identified deficiencies in supporting theoretical courses with practice to bring well-qualified guides into the field.

The present study's other key conceptual contribution concerns students' dilemma regarding the advantages and disadvantages of distance education. This dilemma is evidenced by their uncertain perceptions regarding distance tourist guiding education having experienced EDE after previously only receiving face-to-face education. Apart from EDE, there are several distance education departments for tour guiding education in Türkiye. While this mode has been discussed by students and academics excluding from the pandemic in Türkiye, it is necessary to investigate the ambivalent attitudes of face-to-face tour guiding students toward distance education during EDE specifically. As Zapata-Cuervo et al. (2023) point out, "*Students' perceptions toward online learning would be a bit different from the pre-pandemic [period] when students had options to choose different methods of instruction*."

4.4. Practical Implications

The study has several important practical implications. Firstly, although the participating students work part or full-time, they do not find distance education sufficient for TGE. Hence, the outputs of pre-pandemic online tour guiding departments in Türkiye should be compared with the outputs of face-to-face education, separately from EDE. Secondly, EDE provided an opportunity for tour guiding students studying face-to-face to experience distance education. Based on their experiences, they prefer face-to-face education, especially since it offers practical courses. Thirdly, the findings indicate that tour guiding students prioritize socializing, in-class interaction, active engagement, and self-expression. Hence, they may not prefer distance education because it hampers communication. However, their attitudes could become more positive by using hybrid education in tour guiding. Although students prefer face-to-face education overall, their EDE experiences seem to have confused them somewhat, which can be attributed to the advantages of distance education. Given that, as in other disciplines, distance education is expected to become increasingly common in TGE, universities should offer courses based on practical experience and provide an effective communication system to meet students' expectations. In addition, institute principals encourage technology-based professional development and digital transformation, which lead to the design of an efficient learning environment (Karakose et al., 2021). This may be migrated to the EDE system as well.

4.5. Originality of the Research

Various studies have been conducted on online learning in tourism education before, during, and after the Covid-19. It is essential, nevertheless, that specific research on EDE continues because, as *a new mode of learning*, efficacy in achieving learning outcomes cannot yet be determined. As the present study has shown, research into EDE can answer the question of *how students' perspectives alter when they move to online learning, whether they had positive or negative perceptions different from the pre-pandemic period*. In Türkiye, TGE is provided in both distance and face-to-face systems, independently of EDE. Some students in face-to-face programs considered this as inequitable, and there was already tension between students in the two educational systems before the pandemic. EDE created a potential to either increase or decrease this tension. Our findings showed that while students registered in face-to-face tour guiding departments have benefited from EDE's advantages, they still prefer face-to-face education over online learning. At the same time, facing a dilemma between *the advantages of distance education* and *the outcomes of face-to-face education*, the students appear to have softened their negative opinions regarding distance tour guiding departments. The present study thus provides insight into both the debates surrounding distance education in tour guiding education and the consequences of the current EDE initiatives in tourism education.

4.6. Research Limitations and Future Directions

The study has several limitations. First, because this study primarily focused on EDE in TGE, it excluded students at universities that received full distance education in their normal curriculum. The findings are limited to EDE, specifically the transition from face-to-face to distance education because of the Covid-19 pandemic. Hence, the findings of previous studies investigating normal pre-pandemic distance education cannot be compared to those in this study. Secondly, this study focused only on TGE in Türkiye, and perceptions of EDE are likely to differ for students in other disciplines and other countries. Finally, the study was exploratory qualitative research that is limited in its generalizability. Future research can therefore investigate the EDE experiences of tourism students in other countries during different crises as well as the tendencies and attitudes of tourism academics regarding EDE.

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**: Istanbul Arel University, 06/06/2022-2022/10.

Contribution of Authors

Meltem Altınay Özdemir: Skeletal draft, Investigation, Methodology, Analysis, Software, Validation, Supervision, and Writing-original draft. **Zeynep Tombaş:** Data Collection, Supervision, Validation and Writing-original draft.

Orcid

Meltem Altınay Özdemir https://orcid.org/0000-0002-3002-6127

Zeynep Tombaş https://orcid.org/0000-0002-9370-8504

REFERENCES

Adedoyin, O.B., & Soykan, E. (2020). Covid-19 pandemic and online learning: The challenges and opportunities. *Interactive Learning Environments, 31*(2), 863-875. https://doi.org/10.1080/10494820.2020.1813180

Agyeiwaah, E., Baiden, F.B., Gamor, E., & Hsu, F.C. (2022). Determining the attributes that influence students' online learning satisfaction during COVID-19 pandemic. *Journal of Hospitality, Leisure, Sport & Tourism Education*, *30*, 100364. https://doi.org/10.1016/j.jhlste.2021.100364

Akti Aslan, S., Turgut, Y.E., & Aslan, A. (2021). Teachers' views related the middle school curriculum for distance education during the COVID-19 pandemic. *Education and Information Technologies*, *26*(6), 7381-7405. https://doi.org/10.1007/s10639-021-10587-z

Amin, I., Yousaf, A., Walia, S., & Bashir, M. (2022). What shapes E-Learning effectiveness among tourism education students? An empirical assessment during COVID19. *Journal of Hospitality, Leisure, Sport & Tourism Education*, *30*, 100337. https://doi.org/10.1016/j.jhlste.2021.100337

Arıcı, S., & Karaçay, T. (2023). Students' views on distance learning during the pandemic period. *Journal of Contemporary Tourism Research*, *7*(1), 301-323. https://doi.org/10.32572/guntad.1243985

Bączek, M., Zagańczyk-Bączek, M., Szpringer, M., Jaroszyński, A., & Wożakowska-Kapłon, B. (2021). Students' perception of online learning during the COVID-19 pandemic: A survey study of Polish medical students. *Medicine*, *100*(7). https://doi.org/10.1097/md.0000000000024821

Bakhov, I., Opolska, N., Bogus, M., Anishchenko, V., & Biryukova, Y. (2021). Emergency distance education in the conditions of COVID-19 pandemic: Experience of Ukrainian universities. *Education Sciences*, *11*(7), 364. https://doi.org/10.3390/educsci11070364

Bilsland, C., Nagy, H., & Smith, P. (2020). Virtual Internships and Work-Integrated Learning in Hospitality and Tourism in a Post-COVID-19 World. *International Journal of Work-Integrated Learning*, *21*(4), 425-437.

Bucak, T., & Yigit, S. (2021). The future of the chef occupation and the food and beverage sector after the COVID-19 outbreak: Opinions of Turkish chefs. *International Journal of Hospitality Management*, *92*, 102682. https://doi.org/10.1016/j.ijhm.2020.102682

Cao, W., Fang, Z., Hou, G., Han, M., Xu, X., Dong, J., & Zheng, J. (2020). The psychological impact of the COVID-19 epidemic on college students in China. *Psychiatry Research*, *287*, 112934. https://doi.org/10.1016/j.psychres.2020.112934

Chandra, S., Ranjan, A., & Chowdhary, N. (2022). Online hospitality and tourism education-issues and challenges. *Tourism: An International Interdisciplinary Journal*, *70*(2), 298-316. https://doi.org/10.37741/t.70.2.10

Choi, J.J., Robb, C.A., Mifli, M., & Zainuddin, Z. (2021). University students' perception to online class delivery methods during the COVID-19 pandemic: A focus on hospitality education in Korea and Malaysia. *Journal of Hospitality, Leisure, Sport & Tourism Education*, *29*, 100336. https://doi.org/10.1016/j.jhlste.2021.100336

Choi, J., Kim, N., & Robb, C.A. (2020). COVID-19 and tourism and hospitality education in South Korea: A focus on online learning improvements in higher education. *관광연구저널*, *34*(10), 17-27.

Çolakoğlu O., Epik F., & Efendi, E. (2014). *Tour management and tourist guiding*, (3th ed.). Detay Publication.

Darke, P.R., Brady, M.K., Benedicktus, R.L., & Wilson, A.E. (2016). Feeling close from afar: The role of psychological distance in offsetting distrust in unfamiliar online retailers. *Journal of Retailing*, *92*(3), 287–299. https://doi.org/10.1016/j.jretai.2016.02.001

Davis, N.L., Gough, M., & Taylor, L.L. (2019). Online teaching: Advantages, obstacles and tools for getting it right. *Journal of Teaching in Travel & Tourism*, *19*(3), 256–263. https://doi.org/10.1080/15313220.2019.1612313

Duman, B., & Gençel, N. (2023). Comparison of face-to-face and distance education: An example of a vocational high school. *International Journal of Progressive Education*, *19*(1), 131-153.

Dumford, A.D., & Miller, A.L. (2018). Online learning in higher education: Exploring advantages and disadvantages for engagement. *Journal of Computing in Higher Education*, *30*(3), 452–465. https://doi.org/10.1007/s12528-018-9179-z

Durak, G., & Çankaya, S. (2020a). Undergraduate students' views about emergency distance education during the COVID-19 pandemic. *Online Submission*, *5*(1), 122-147. https://eric.ed.gov/?id=ED609069

Durak, G., & Çankaya, S. (2020b). Emergency distance education process from the perspectives of academicians. *Asian Journal of Distance Education*, *15*(2), 159-174. http://www.asianjde.com/ojs/index.php/AsianJDE/article/view/507

Durak, G., Çankaya, S., & İzmirli, S. (2020). Examining the Turkish Universities' distance education systems during the COVID-19 pandemic. *Necatibey Faculty of Education Electronic Journal of Science and Mathematics Education*, *14*(1), 787-810. https://doi.org/10.17522/balikesirnef.743080

Eker, N. (2015). An evaluation for professional tourist guidance education [Master's dissertation]. Sakarya University, Türkiye.

Eker, N., & Zengin, B. (2016). Evaluation of tourist guide education: An application to professional tourist guides. *Journal of Research in Education and Teaching*, *5*(4), 65-74. http://www.jret.org/FileUpload/ks281142/File/08a.nuray_eker.pdf

Eser, S., & Şahin, S. (2021). A review on current problems in tourist guiding profession. *Journal of Turkish Tourism Research*, *4*(2), 1344–1355.

Fojtík, R. (2018). Problems of distance education. *ICTE Journal*, *7*(1), 14-23. https://periodicals.osu.eu/ictejournal/dokumenty/2018-01/ictejournal-2018-1.pdf#page=15

Goh, E. (2020). Educating the future hospitality and tourism workforce: Trends, issues, and directions in Australia and New Zealand. *Journal of Hospitality & Tourism Education, 32(4), 193.* https://doi.org/10.1080/10963758.2019.1688162

Goh, E., & Sigala, M. (2020). Integrating Information & Communication Technologies (ICT) into classroom instruction: Teaching tips for hospitality educators from a diffusion of innovation approach. *Journal of Teaching in Travel & Tourism, 20(2), 156-165.* https://doi.org/10.1080/15313220.2020.1740636

Goh, E., & Wen, J. (2020). Applying the technology acceptance model to understand hospitality management students' intentions to use electronic discussion boards as a learning tool. *Journal of Teaching in Travel & Tourism*, 1-13. https://doi.org/10.1080/15313220.2020.1768621

Gunawardena, C.N., & McIsaac, M.S. (2013). Distance education. In *Handbook of research on educational communications and technology* (pp. 361-401). Routledge.

Hodges, C., Moore, S., Lockee, B., Trust, T., & Bond, A. (2020). The difference between emergency remote teaching and online learning. *Educause Review*. https://er.educause.edu/articles/2020/3/the-difference-between-emergency-remote-teaching-and-online-learn

Hsu, L. (2021). Learning tourism and hospitality subjects with massive open online courses (MOOCs): A cross-sectional and longitudinal study. *Journal of Hospitality, Leisure, Sport & Tourism Education*, *29*, 100276. https://doi.org/10.1016/j.jhlste.2020.100276

İbicioğlu, H., & Antalyalı, U.Ö.L. (2005). The effects of opportunity, perception, motivation and interaction factors on the success of distance education: A comparative application. *Journal of Çukurova University Social Sciences Institute*, *14*(2), 325-338. https://dergipark.org.tr/en/pub/cusosbil/issue/4372/59838

Jiang, H., Islam, A.A., Gu, X., & Spector, J.M. (2021). Online learning satisfaction in higher education during the COVID-19 pandemic: A regional comparison between Eastern and Western Chinese universities. *Education and Information Technologies*, 1-23.

Justin, M.P., Wil, M., & Bui, N.M.C. (2022). Online learning for Vietnamese Hospitality and Tourism University students during a time of Covid-19. *VNU Journal of Science: Education Research*, *38*(4). https://doi.org/10.25073/2588-1159/vnuer.4692

Kallou, S., & Kikilia, A. (2021). A transformative educational framework in tourism higher education through digital technologies during the COVID-19 pandemic. *Advances in Mobile Learning Educational Research*, *1*(1), 37-47. https://www.syncsci.com/journal/index.php/AMLER/article/view/AMLER.2021.01.005

Kandilli Observatory and Earthquake Research Institute (2023). http://www.koeri.boun.edu.tr/scripts/lst2.asp, 06 February 2023.

Kapasiaa, N., Paul, P., Avijit, R., Saha, J., Zaveri, A., & Rahul, M. (2020). Impact of lockdow n on learning status of undergraduate and postgraduate students during COVID-19 pandemic in West Bengal, India. *Children and Youth Services Review*, *116*, 105194. https://doi.org/10.1016/j.childyouth.2020.105194

Karadağ, E., & Yucel, C. (2020). Distance education at universities during the novel coronavirus pandemic: An analysis of undergraduate students' perceptions. *Journal of Higher Education (Turkey)*, *10*(2), 181-192. https://doi.org/10.2399/yod.20.730688

Karakose, T., Polat, H., & Papadakis, S. (2021). Examining teachers' perspectives on school principals' digital leadership roles and technology capabilities during the COVID-19 pandemic. *Sustainability*, *13*(23), 13448. https://doi.org/10.3390/su132313448

Karataş, T.Ö., & Tuncer, H. (2020). Sustaining language skills development of pre-service EFL teachers despite the COVID-19 interruption: A case of emergency distance education. *Sustainability*, *12*(19), 8188. https://doi.org/10.3390/su12198188

Kaushal, V., & Srivastava, S. (2020). Hospitality and tourism industry amid COVID-19 pandemic: Perspectives on challenges and learnings from India. *International Journal of Hospitality Management*, 102707. https://doi.org/10.1016/j.ijhm.2020.102707

Kim, H. J., & Jeong, M. (2018). Research on hospitality and tourism education: Now and future. *Tourism Management Perspectives*, *25*, 119-122. https://doi.org/10.1016/j.tmp.2017.11.025

Klisowska, I., Seń, M., & Grabowska, B. (2021). Advantages and disadvantages of distance learning. *E-methodology*, *7*(7), 27-32. https://doi.org/10.15503/emet2020.27.32

Korkmaz, E.K., Şahin, G.G., & Işıkhan, S.Y. (2022). Tourism education in higher education during the COVID 19 process: A research in Ankara. *Journal of Turkish Tourism Research*, *6*(3), 859-878. https://doi.org/10.26677/TR1010.2022.1093

Korkmaz, G., & Toraman, Ç. (2020). Are we ready for the post-COVID-19 educational practice? An investigation into what educators think as to online learning. *International Journal of Technology in Education and Science*, *4*(4), 293-309. https://eric.ed.gov/?id=EJ1271308

Köksalanlar, A.A., & Çözeli, F.E. (2021). Attitudes of tourist guidance students towards the distance education due to the COVID-19. *Journal of Academic Researches and Studies, 13*(25), 539-550. https://doi.org/10.20990/kilisiibfakademik.869875

Lamanauskas, V., & Makarskaite-Petkeviciene, R. (2021). Distance lectures in university studies: Advantages, disadvantages, improvement. *Contemporary Educational Technology*, *13*(3). https://eric.ed.gov/?id=EJ1306583

Lei, S., & So, A.S. (2021). Online teaching and learning experiences during the COVID-19 pandemic: A comparison of teacher and student perceptions. *Journal of Hospitality & Tourism Education*, *33*(3), 148–162. https://doi.org/10.1080/10963758.2021. 1907196

Li, C., & Agyeiwaah, E. (2023). Online learning attributes on overall tourism and hospitality education learning satisfaction: Tourism Agenda 2030. *Tourism Review*, *78*(2), 395-410. http://dx.doi.org/10.1108/TR-05-2022-0221

Masalimova, A.R., Khvatova, M.A., Chikileva, L.S., Zvyagintseva, E.P., Stepanova, V.V., & Melnik, M.V. (2022). Distance learning in higher education during COVID-19. *Frontiers in Education*, *7*, 120. https://doi.org/10.3389/feduc.2022.822958

Maxwell, J.A. (2012). *Qualitative research design: An interactive approach*. Sage publications.

Merriam, S.B. (2018). *Nitel Araştırma Desen ve Uygulama için Bir Rehber* (S. Turan, Çev.). Nobel Akademik Yayıncılık.

Mulyanti, B., Purnama, W. & Pawinanto, R.E. (2020). Distance learning in vocational high schools during the COVID-19 pandemic in West Java Province. *Indonesia Indonesian Journal of Science & Technology*, *5*(2), 271-282. https://doi.org/10.17509/ijost.v5i2.24640

Munoz, K.E., Wang, M.J., & Tham, A. (2021). Enhancing online learning environments using social presence: Evidence from hospitality online courses during COVID-19. *Journal of Teaching in Travel & Tourism*, *21*(4), 339-357. https://doi.org/10.1080/15313220.2021.1908871

Nambiar, D. (2020). The impact of online learning during COVID-19: Students' and teachers' perspective. *The International Journal of Indian Psychology*, *8*(2), 783-793.

Newby, T.J., Stepich, D.A., Lehman, J.D., & Russell, J.D. (2000). *Instruction technology for teaching and learning*. Merrill.

O'Connor, N. (2021). Using active learning strategies on travel and tourism higher education programmes in Ireland. *Journal of Hospitality, Leisure, Sports and Tourism Education*, *29*, 100326. https://doi.org/10.1016/j.jhlste.2021.100326

Patiar, A., Kensbock, S., Benckendorff, P., Robinson, R., Richardson, S., Wang, Y., & Lee, A. (2021). Hospitality students' acquisition of knowledge and skills through a virtual field trip experience. *Journal of Hospitality & Tourism Education*, *33*(1), 14-28. https://doi.org/10.1080/10963758.2020.1726768

Pesha, A., & Kamarova, T. (2021). Socio-psychological problems of the transition of university teachers to distance employment during the COVID-19 pandemic. *SHS Web of Conferences*, *99*, 01040. https://doi.org/10.1051/shsconf/20219901040

Poláková, P., & Klímová, B. (2021). The perception of Slovak students on distance online learning in the time of coronavirus—A preliminary study. *Education Sciences*, *11*(2), 81. https://doi.org/10.3390/educsci11020081

Prifti, P., Kuo, Y.-C., Walker, A.E., & Belland, B.R. (2020). Self–efficacy and student satisfaction in the context of blended learning courses. *Open Learning: The Journal of Open, Distance and e-Learning*, 1–15. https://doi.org/10.1080/02680513.2020.1755642

Qiu, H., Li, Q., & Li, C. (2021). How technology facilitates tourism education in COVID-19: Case study of Nankai University. *Journal of Hospitality, Leisure, Sport & Tourism Education*, *29*, 100288. https://doi.org/10.1016/j.jhlste.2020.100288

Ritonga, A.K. (2022). Effectiveness of English for tourism e-learning during the Covid-19 pandemic. *Journal of Education Technology*, *6*(1), 102-109. https://doi.org/10.23887/jet.v6i1.42312

Saraç, Y. (2020, May 25). Basın Açıklaması [Press Statement]. https://www.yok.gov.tr/Sayfalar/Haberler/2020/YKS%20Ertelenmesi%20Bas%C4%B1n%20A%C3%A7%C4%B1klamas%C4%B1.aspx

Schram, T.H. (2003). *Conceptualizing qualitative inquiry*. Merrill Prentice Hall. Schwandt, TA.

Sezgin, S. (2021). Analysis of the emergency remote education process: Featured terms, problems and lessons learned. *Anadolu University Journal of Social Sciences (AUJSS), 21*(1), 273-296. https://doi.org/10.18037/ausbd.902616

Sharma, S. (2020). Post COVID-19: Change required in hospitality education. https://hospitality.economictimes.indiatimes.com/blog/post-covid-19-change-required-in- hospitality-education/422

Shim, T.E., & Lee, S.Y. (2020). College students' experience of emergency remote teaching due to COVID-19. *Children and Youth Services Review*, *119*, 105578. https://doi.org/10.1016/j.childyouth.2020.105578

Shyju, P.J., Vinodan, A., Sadekar, P., Sethu, M., & Lama, R. (2021). Determinants of online learning efficacy and satisfaction of tourism and hospitality management students during the COVID-19 pandemic. *Journal of Teaching in Travel & Tourism*, *21*(4), 403-427. https://doi.org/10.1080/15313220.2021.1998941

Simonson, M., & Seepersaud, D.J. (2019). *Distance education: Definition and glossary of terms* (4th ed.). Information Age Publishing.

Sitosanova, O. (2021). *Advantages and disadvantages of distance education at the university*. Scientific Papers Collection of the Angarsk State Technical University.

Smith, D.W. (2023, June 06). "Phenomenology", *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), accessed in https://plato.stanford.edu/archives/sum2018/entries/phenomenology/

Şanlıöz-Özgen, H.K., & Küçükaltan, E.G. (2023). Distance education at tourism higher education programs in developing countries: Case of Türkiye with a strategic perspective and recommendations. *Journal of Hospitality, Leisure, Sport and Tourism Education*, *32*. https://doi.org/10.1016/j.jhlste.2023.100419

Tavitiyaman, P., Ren, L.P., & Fung, C. (2021). Hospitality students at the online classes during COVID-19: How personality affects experience. *Journal of Hospitality, Leisure, Sport & Tourism Education*, *28*, 100304. https://doi.org/10. 1016/j.jhlste.2021.100304

Toquero, C.M. (2020). Emergency remote teaching amid COVID-19: The turning point. *Asian Journal of Distance Education*, *15*(1),185-188. http://www.asianjde.org/ojs/index.php/AsianJDE/article/view/450

Tourist Guiding Professional Law (2012). https://www.mevzuat.gov.tr/mevzuatmetin/1.5.6326.pdf.

Turkish Higher Education Quality Council (2020). Distance education during the pandemic period. https://portal.yokak.gov.tr/makale/pandemi-doneminde-uzaktan-egitim/

Vlassopoulos, G., Karikas, G.A., Papageorgiou, E., Psaromiligos, G., Giannouli, N., & Karkalousos, P. (2021). Assessment of Greek High School students towards distance learning, during the first wave of COVID-19 pandemic. *Creative Education*, 12, 934-949. https://doi.org/10.4236/ce.2021.124067

Wang, G., Zhang, Y., Zhao, J., Zhang, J., & Jiang, F. (2020). Mitigate the effects of home confinement on children during the COVID-19 outbreak. *The Lancet*, *395*(10228), 945-947. https://doi.org/10.1016/S0140-6736(20)30547-X

Xu, J., Tavitiyaman, P., Kim, H.J., & Lo, S.K. (2022). Hospitality and tourism higher education in the post-COVID era: Is it time to change?. *Journal of Hospitality & Tourism Education*, *34*(4), 278-290. https://doi.org/10.1080/10963758.2022.2056044

Yağcı, K., Efendi, M., & Akçay, S. (2019). Concept of distance education: From perspective of tourism guidance students. *Journal of Travel and Tourism Research*, 14, 118-136. https://dergipark.org.tr/en/pub/ttr/issue/59443/650216

Ye, H., & Law, R. (2021). Impact of COVID-19 on hospitality and tourism education: A case study of Hong Kong. *Journal of Teaching in Travel & Tourism*, *21*(4), 428-436. https://doi.org/10.1080/15313220.2021.1875967

Yılmaz, G.K., & Güven, B. (2015). Determining teacher candidates' perceptions towards distance education through metaphors. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *6*(2), 299-322. https://doi.org/10.16949/turcomat.75936

Zapata-Cuervo, N., Montes-Guerra, M.I., Shin, H.H., Jeong, M., & Cho, M.H. (2023). Students' psychological perceptions toward online learning engagement and outcomes during the COVID-19 pandemic: A comparative analysis of students in three different countries. *Journal of Hospitality & Tourism Education*, *35*(2), 108-122. https://doi.org/10.1080/10963758.2021.1907195

Zaveri, B., Amin, P., & Nasabinmana, V. (2020). Efficiency of online learning in the modern technology era-students' perception towards online learning. *Journal of Marketing Vistas*, *10*(1), 35-47. http://lib.jnu.ac.in/sites/default/files/pdf/JMV_Compressed.pdf#page=39

Zhong, Y., Busser, J., Shapoval, V., & Murphy, K. (2021). Hospitality and tourism student engagement and hope during the COVID-19 pandemic. *Journal of Hospitality & Tourism Education*, *33*(3), 194-206. https://doi.org/10.1080/10963758.2021.1907197

APPENDIX-1. PARTICIPANT PROFILE

No	Education level	Gender	Age	Marital status	No	Education level	Gender	Age	Marital status
P1	Associate student	Female	25	Married	P42	Associate student	Male	52	Married
P2	Associate student	Female	32	Single	P43	Associate student	Female	19	Single
P3	Associate student	Female	31	Single	P44	Associate student	Male	27	Single
P4	Associate student	Male	32	Single	P45	Associate student	Female	19	Single
P5	Undergraduate student	Male	40	Married	P46	Associate student	Male	34	Single
P6	Associate student	Male	27	Single	P47	Associate student	Male	29	Single
P7	Undergraduate student	Male	40	Married	P48	Associate student	Female	40	Single
P8	Undergraduate student	Male	33	Married	P49	Associate student	Male	22	Single
P9	Undergraduate student	Female	35	Married	P50	Associate student	Female	25	Single
P10	Undergraduate student	Female	30	Married	P51	Associate student	Male	20	Single
P11	Associate student	Male	35	Married	P52	Associate student	Female	19	Single
P12	Associate student	Female	21	Single	P53	Undergraduate student	Male	33	Single
P13	Associate student	Male	42	Married	P54	Undergraduate student	Female	19	Single
P14	Associate student	Male	21	Single	P55	Undergraduate student	Male	21	Single
P15	Associate student	Female	24	Single	P56	Associate student	Female	35	Married
P16	Undergraduate student	Male	45	Married	P57	Associate student	Female	18	Single
P17	Undergraduate student	Male	27	Single	P58	Associate student	Male	27	Single
P18	Associate student	Female	38	Married	P59	Associate student	Male	24	Single
P19	Undergraduate student	Male	42	Single	P60	Associate student	Female	25	Single
P20	Associate student	Male	28	Single	P61	Associate student	Male	52	Single
P21	Associate student	Male	23	Single	P62	Associate student	Female	32	Single
P22	Associate student	Male	28	Single	P63	Associate student	Female	31	Single
P23	Undergraduate student	Male	26	Single	P64	Associate student	Male	35	Married
P24	Undergraduate student	Female	25	Married	P65	Associate student	Female	26	Single
P25	Associate student	Female	38	Single	P66	Associate student	Female	18	Single
P26	Undergraduate student	Female	46	Married	P67	Associate student	Male	21	Single
P27	Associate student	Female	20	Single	P68	Associate student	Female	20	Single
P28	Associate student	Male	37	Married	P69	Associate student	Female	19	Single
P29	Associate student	Male	51	Married	P70	Associate student	Female	19	Single
P30	Associate student	Male	34	Single	P71	Associate student	Male	27	Single
P31	Associate student	Female	40	Single	P72	Associate student	Male	18	Single
P32	Associate student	Male	38	Married	P73	Associate student	Male	20	Single
P33	Associate student	Male	42	Married	P74	Associate student	Male	26	Single
P34	Associate student	Female	30	Single	P75	Associate student	Male	19	Single
P35	Associate student	Male	42	Married	P76	Undergraduate student	Female	20	Single
P36	Associate student	Female	51	Single	P77	Associate student	Female	21	Single
P37	Associate student	Male	30	Single	P78	Associate student	Female	22	Single
P38	Associate student	Female	58	Single	P79	Associate student	Male	33	Single
P39	Associate student	Male	35	Single	P80	Associate student	Male	44	Married
P40	Undergraduate student	Male	42	Single	P81	Associate student	Male	23	Single
P41	Associate student	Male	46	Single					

APPENDIX 2. SUB-THEMES OF EDE EXPERIENCES

Sub-themes	*f*	%
Perceptions of strengths (Perception of face-to-face education)	121	10.6
Neutral perceptions (Perception of face-to-face education)	88	7.7
Perceptions for weaknesses (Perception of DE)	71	6.2
Yes (Difference between face-to-face and distance education)	66	5.8
Associate student	65	5.7
Neutral perceptions (Perception of DE)	65	5.7
Perceptions for strengths (Perception of DE)	64	5.6
No (Sufficiency of EDE)	51	4.5
No (Practical sufficiency of distance education)	48	4.2
Effective time management	40	3.5
Yes (Practical sufficiency of distance education)	33	2.9
Efficiency	32	2.8
Yes (Sufficiency of EDE)	30	2.6
Insufficiency of TGE	28	2.4
Extremely high understanding (5)	25	2.2
Perceptions of weaknesses (Perception of face-to-face education)	24	2.1
Insufficient for vocational courses	24	2.1
Very (4)	22	1.9
Lack of motivation	22	1.9
Insufficient practice	19	1.7
Moderate (3)	18	1.6
Lack of communication	17	1.5
Motivation problems	17	1.5
Undergraduate student	16	1.4
Compensation	16	1.4
No (Difference between face-to-face & distance education)	15	1.3
None (Disadvantages of EDE)	12	1.0
None (Advantages of EDE)	11	1.0
Lack of effective communication	11	1.0
Slightly (2)	10	0.9
Independence from place	10	0.9
Technical problems	9	0.8
Socialization	8	0.7
Concentration	8	0.7
Understanding not at all (1)	6	0.5
Low attendance to courses	6	0.5
Unidirectionality	5	0.4
Savings	5	0.4
Ease of access to materials	4	0.3
Self-expression	3	0.3
Total	1145	100

*TGE: Tour Guiding Education; EDE: Emergency distance education; DE: Distance Education

International Journal of Assessment Tools in Education
2024, Vol. 11, No. 2, 345–367
https://doi.org/10.21449/ijate.1360899
journal homepage: https://dergipark.org.tr/en/pub/ijate *Research Article*

A comparison of Turkish and European English language teachers' language assessment knowledge levels and perceptions

Samet Fındıklı [1*], **Kağan Büyükkarcı** [2]

[1]Republic of Türkiye Ministry of National Education, Isparta, Türkiye
[2]Süleyman Demirel University, Faculty of Education, Department of Foreign Language Education, Isparta, Türkiye

ARTICLE HISTORY

Received: Sep. 15, 2023

Accepted: Mar. 12, 2024

Keywords:

Language assessment knowledge,
Language assessment literacy,
Assessing language skills,
EFL teachers.

Abstract: Language assessment knowledge, the capacity of language instructors to skillfully design, construct, and assess language evaluations, is pivotal for effective language education. This study investigates the language assessment knowledge, encompassing both general and skill-specific aspects, of in-service language educators from Europe and Türkiye. The primary objective is to contrast the language assessment knowledge of these two groups, highlighting potential differences in their assessment knowledge in terms of general and four language skills. Employing a mixed-methods approach, data were gathered sequentially via quantitative scale and qualitative online interviews. A total of 94 language teachers, 48 from Turkey and 46 from diverse European countries took part in this research. They completed the Language Assessment Knowledge Scale, and eight instructors engaged in semi-structured online interviews. The participants were selected using convenience sampling. The results indicated that while both groups scored above the average and were considered assessment literate, European language teachers had a significantly higher level of LAK compared to Turkish language teachers. This suggests that European teachers possess greater proficiency and competence in language assessment, potentially influencing the quality of the assessments they create and assess. Considering the importance of assessment knowledge mentioned in numerous studies, despite the limited sample size of this study, its results are important for the professional development of language educators. These outcomes can inform the development of teacher training programs, particularly for Turkish educators. The Ministry of National Education may consider prioritizing assessment-related subjects, such as assessing the four language skills, in future in-service teacher training initiatives.

1. INTRODUCTION

Assessment is accepted as the engine that drives learning (Cowan, 1998). Indeed, education and assessment are intertwined and indispensable units for each other. Although assessment and testing are mostly seen as scoring tools about how much learning has taken place in the classroom (Giraldo, 2018), they are also invaluable feedback that will guide the course of education (Mertler & Campbell, 2005). Especially in language education, which includes four

*CONTACT: Samet FINDIKLI ✉ sametfindikli@gmail.com Republic of Türkiye Ministry of National Education, Isparta, Türkiye

e-ISSN: 2148-7456

skills (reading, writing, speaking, and listening), the scope expands considerably, and the assessment knowledge level of the teachers is of great importance in terms of accurately gauging the education given and increasing the quality of the education shaped by the feedback from the assessment (Hughes, 2003; Malone, 2011; Popham, 2011; Stiggins, 1995).

While Popham (2011) agrees on the necessity of assessment literacy, he opposes what he perceives as the existing definitions in teachers' minds. According to Popham, assessment literacy is not solely "knowledge about educational tests and their roles," nor is it "the technical skills needed to construct and evaluate educational tests," or the ability “to calculate means, standard deviations, and correlation coefficients" (p. 267). Instead, he redefines assessment literacy as an individual's understanding of fundamental assessment concepts and procedures that are likely to influence educational decisions (Popham, 2011). This new definition serves as a reminder to educators that assessment not only measures but can also influence the course of education by providing feedback. Similarly, Boyles (2005) argues that "teachers and administrators need the necessary tools for analyzing and reflecting upon test data to make informed decisions about instructional practice and program design" (p. 18).

Having a better understanding of assessment procedures can have positive impacts on the quality of education, as stated by Malone (2011) who believes that “language assessment and language teaching go hand in hand. The best teaching involves high-quality assessment practices, and great assessment provides positive washback to the teaching and learning process” (p. 2). Thus, in order to apply successful assessment procedures in their classrooms and programs, educators need to have a strong foundation in assessment literacy, as emphasized by Malone (2011). Similarly, Giraldo (2018) argues that selecting, designing, and evaluating valid assessments is essential for achieving positive outcomes in learning and teaching. Furthermore, according to Büyükkarcı (2014), the systematic nature of assessment provides teachers with the opportunity to improve their teaching and provide the best learning experience for their students. This claim is supported by Cheng and Fox (2017), who noted that assessment plays a critical role in checking on learning and providing important information to teachers.

1.1. Assessment, Testing, and Evaluation

Assessment, testing, and evaluation in education play a crucial role in determining students' learning outcomes. While learning often leads to observable changes in performance, it is essential to recognize that learning is not always directly observable, as noted by Colby (2010). To bridge this gap, various methods and techniques are employed to measure unobservable behavioral changes, helping educators identify areas of mastery and improvement in learners (Douglas, 2009).

It is important to distinguish between the terms assessment, evaluation, and testing, as they are frequently used interchangeably. Assessment, defined by Coombe (2018), involves measuring an individual's performance to infer their abilities and provide feedback on their development. This process includes various methods such as tests, quizzes, and observations to gauge student learning (Brown, 2000; Rogiers, 2014). Assessment can be further categorized into formal and informal assessments. Informal assessments rely on observation and lack standardized rubrics, while formal assessments use standardized instruments and exams (Coombe, 2018). Both serve different purposes and have their advantages and disadvantages. There are also different assessment types, including diagnostic, self, peer, formative, and summative assessments, depending on their purpose and application.

Testing, as described by Nagai et al. (2020), is a specific type of assessment that involves formal tasks graded to gauge learners' language abilities. Tests are tools used to measure performance or knowledge, and they are designed with specific goals to draw desired conclusions about a student's abilities (Green, 2013; Bachman, 2004; Heaton, 1989).

Evaluation, a broader concept, involves the systematic gathering of information to make decisions (Bachman, 1990). It encompasses assessing program components, methods, or results to determine if they meet predetermined standards or objectives (Mohan, 2022). Evaluation also extends to assessing students, teachers, and curriculum effectiveness in relation to established goals.

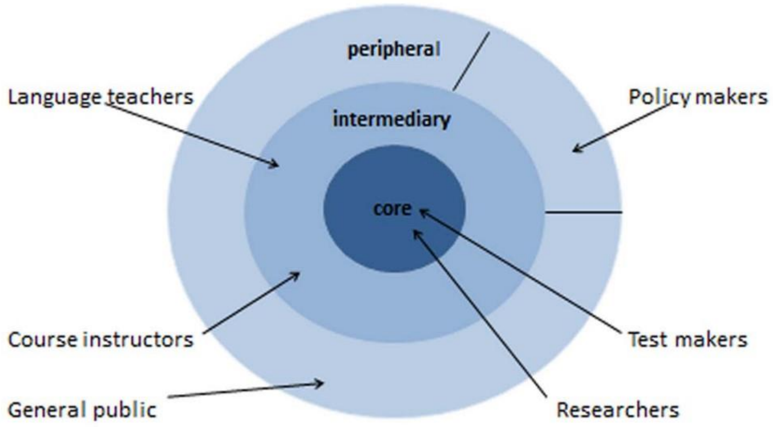
In summary, assessment, testing, and evaluation serve distinct purposes in education, with assessment focusing on measuring individual performance and providing feedback, testing concentrating on formal tasks to gauge abilities, and evaluation encompassing a broader process of gathering information to make decisions about educational programs and outcomes.

1.2. Language Assessment Literacy

Language Assessment Literacy (LAL) is an important component of assessment literacy for language teachers, as it involves their conceptual knowledge and competence in testing, assessment, and evaluation. LAL is defined by Inbar-Lourie (2017) as "the essential knowledge, skills, and principles that stakeholders involved in assessment activities must master in order to perform assessment tasks effectively." Teachers devote a significant portion of their instructional time to assessment tasks, making it critical for them to be equipped with LAL skills and knowledge. DeLuca et al. (2015), Gotch and French (2014), and Siegel and Wissehr (2011) all highlight the significance of teacher preparation in assessment, covering topics such as test item creation, administration, evaluation, analysis, statistics, and reporting.

Davies (2008) emphasizes the importance of integrating skills, knowledge, and principles into teaching, whereas Scarino (2013) proposes integrating specialized knowledge of language assessment with an understanding of the interconnectedness of language, culture, and learning. According to Popham (2011), educators' assessment literacy influences their ability to make informed educational decisions, and Wiliam (2011) emphasizes the potential of integrating assessment with instruction to improve student engagement and learning outcomes.

Figure 1. *AL/LAL stakeholders (Taylor, 2013, p. 409).*



The literature on language assessment literacy addresses the question of which stakeholders should be literate in language assessment. Taylor (2013) proposes varying levels of assessment literacy based on the roles and responsibilities of stakeholders. As can be seen in Figure 1, researchers and test developers are regarded as the core group, necessitating a thorough understanding of assessment theory, technical expertise, and moral principles. Course instructors and language teachers are at the intermediate level, as they require practical expertise for test development while putting less emphasis on theory or ethical principles. Policymakers and the general public are in the outermost circle, where a basic understanding of test instrument characteristics and score significance suffices for decision-making.

1.3. Statement of the Problem

Assessment in language education encompasses testing and evaluation methods (Clapham, 2000). Language teachers are responsible for various assessment processes, including preparation, administration, evaluation of assessment tools, feedback provision, and informal observations (Ölmezer-Öztürk & Aydın, 2019). However, there is ongoing debate regarding whether language teachers receive adequate education and training to fulfill these responsibilities.

Assessment results help identify areas of weakness in language knowledge, determine students' needs, and evaluate the effectiveness of teaching (Harding & Kremmel, 2016). Good assessment practices also enhance teaching quality and student learning outcomes (Jannati, 2015). Moreover, assessment benefits students by identifying areas needing improvement, fostering self-assessment skills, and preparing them for high-stakes standardized tests (Thomas et al., 2004).

Language assessment literacy is essential for a teacher's professional development, and it requires both theoretical knowledge and practical implementation (Inbar-Lourie, 2008). A lack of adequate training in language assessment can lead to inadequate assessment practices and hinder student progress (Giraldo, 2021). Numerous researchers emphasize the importance of language assessment literacy for language teachers (DeLuca & Klinger, 2010; Fulcher, 2012; Harding & Kremmel, 2016; Lam, 2015; Malone, 2011; Scarino, 2013; Shepard, 2000; Siegel & Wissehr, 2011; Taylor, 2009). However, there is no consensus on specific competencies for language teachers in this area.

Fulcher (2012) notes that despite significant developments since the 1990s, language assessment literacy is still in its early stages. Recent research has highlighted the need for language teachers to receive adequate training in language assessment (Lam, 2015; Sarıyıldız, 2018; Sevimel-Şahin, 2019; Sevimel-Şahin & Subaşı, 2021; Tamerer, 2019; Wardani et al., 2021; Yetkin, 2015). Studies have focused on pre-service teachers, university-level English instructors, and in-service teachers, examining their assessment literacy levels, training needs, and perceptions.

Most existing studies on language assessment knowledge have been regional, focusing on specific geographic areas. Understanding regional differences in assessment practices is crucial, as cultural orientations and learner preferences can influence language assessment effectiveness (Krajka, 2019). Furthermore, the EF English Proficiency Index 2022 report highlights lower English proficiency levels in Türkiye compared to other European countries (EF EPI, 2022). This raises questions about potential links between language teachers' assessment knowledge and variations in proficiency levels. Further investigation into assessment practices among Turkish and European language teachers is needed to understand disparities and improve language education.

Language assessment is a significant part of education in Türkiye and Europe, with teacher education programs typically including coursework and practical training in assessment. These programs cover assessment principles, types, validity, reliability, and fairness. Findings from this study will shed light on the strengths and weaknesses of language teacher education programs, contributing to improved language education practices in Türkiye and Europe.

1.4. Purpose of the Study

By examining the language assessment knowledge of in-service language teachers, with a particular focus on Türkiye, this study aims to fill a significant gap in the literature. The main goal is to gauge these teachers' levels in language assessment and then compare it with that of their counterparts in European countries. By doing this, the study hopes to identify any potential variations in assessment procedures among language teachers from various countries and investigate how they may affect assessment knowledge. The study also compares and examines the assessment abilities and knowledge of language teachers while taking into account cultural differences and standardization policies. This thorough investigation of language assessment practices will add to the body of knowledge already available on LAL by providing insightful information on the particular difficulties and variations that language teachers face in their assessment practices in various contexts. The research has a clear focus on both general and skill-based assessment knowledge, which will provide detailed information on the knowledge levels of in-service teachers. The study will also contribute to the development of effective language assessment practices in Türkiye and other countries. Overall, the research is expected to provide valuable insights into the AL of in-service language teachers and inform the development of effective language assessment policies and practices.

At the same time, this research will seek answers to the following research questions;

1. What is the Turkish and European EFL teachers' level of language assessment knowledge (LAK) in assessing students' language skills in English?
2. Is there a significant difference between the general and skill-based language assessment knowledge levels of Turkish and European language teachers?
3. How do country and demographic factors such as years of experience, educational background, school level, completion of a testing course, and attendance of testing and assessment training influence the overall LAK level and its skill-based components?
4. What are the perceptions of Turkish and European EFL teachers about their classroom-based language assessment practices?

2. METHOD

2.1. Research Design

This cross-national comparative educational study examined the assessment knowledge of English language teachers in Türkiye and European countries using mixed methods research. This approach combined qualitative and quantitative data to gain a comprehensive understanding of language teachers' assessment practices and knowledge (Cohen et al., 2018). Specifically, the study utilized an explanatory sequential design (Creswell, 2014), collecting quantitative data initially and then qualitative data to further explain the quantitative findings. This mixed methods approach enables a deeper exploration of language teachers' assessment knowledge and techniques in different educational contexts (Fox, 2016). It is claimed that mixed methods research design helps researchers address a wider range of concerns regarding the complex phenomena that are the focus of applied linguistic studies, and language assessment studies in particular, by moving beyond paradigmatic polarity (Fox, 2016). Additional sub-models are included in a mixed methods research.

Figure 2. *Explanatory design: Follow-up explanations model. Adapted from Creswell and Clark, 2007, p. 72.*

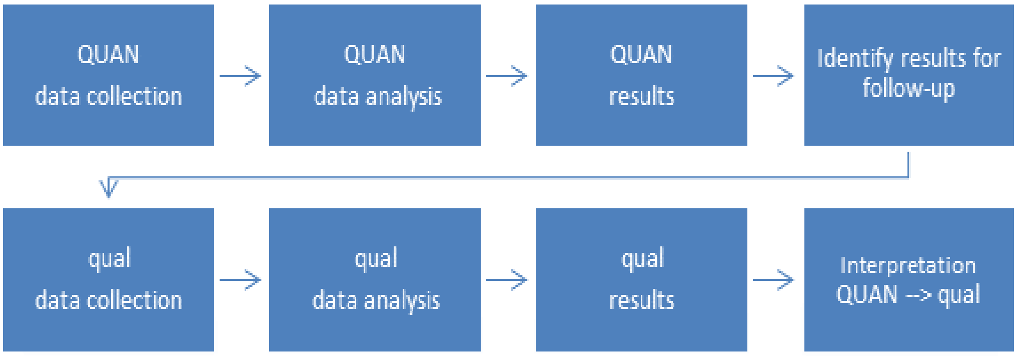


Figure 2 provides a visual representation of the quantitative and qualitative data gathered in the current study, which Creswell (2014) refers to as explanatory sequential design, aiming to gain a deeper understanding of the assessment methods and expertise employed by language teachers.

2.2. Participants

This study collected data from two groups of participants: in-service English language teachers in Türkiye and those in European countries all of whom were actively working in middle and high schools. The primary aim was to compare the LAK of these language teachers. European countries were considered as a single group for analysis due to the complexity of handling each country individually (Lor, 2019).

The selection of these two groups enabled an investigation into potential differences in language assessment knowledge between them. Given that European countries generally exhibited higher levels of English language proficiency compared to Türkiye (EF EPI, 2022), it is hypothesized that European language teachers may possess higher assessment literacy, which could contribute to more accurate assessments and tailored teaching methods.

Table 1. *Countries of participants.*

Countries	N	Percent
Türkiye	48	51.1%
Italy	12	12.7%
Spain	12	12.7%
Romania	7	7.4%
Albania	4	4.3%
Bulgaria	4	4.3%
Germany	4	4.3%
Lithuania	3	3.2%
Total	94	100%

As can be seen in Table 1, the study maintained a balanced distribution of participants, with a total of 94 in-service language teachers, including 48 from Türkiye and 46 from European countries (Italy, Spain, Romania, Albania, Bulgaria, Germany, and Lithuania). All participants' countries were selected as countries where English is a foreign language, not a first or second language, and have similar language teaching objectives.

Table 2. *Crosstabulation of gender * BA program graduated from.*

		BA program Graduated From		
		ELT	Non-ELT	Total
Gender	Female	49	10	59
	Male	27	8	35
Total		76	18	94

Table 2 shows that 59 participants were female, while 35 were male. Furthermore, the participants predominantly held degrees in English Language Teaching (ELT) programs, with 76 of them having graduated in this field. Regarding school levels, 46 participants worked at middle school level, while the remaining 48 worked at high school level. These distributions provided a comprehensive view of language assessment knowledge among participants in both Türkiye and European countries.

Quantitative data collection initially involved convenience sampling, and snowball sampling was utilized to reach more participants, especially in European countries. The goal was to include language teachers from diverse European countries rather than focusing on a single country. The selection for qualitative interviews was based on volunteers from the quantitative phase, with 8 participants representing a mix of Turkish and European teachers with varied years of experience and educational backgrounds.

2.3. Data Collection

The study was conducted in the 2022-2023 educational year and employed a mixed-methods approach to collect data, combining quantitative and qualitative research tools. For the quantitative aspect, the Language Assessment Knowledge Scale (Ölmezer-Öztürk & Aydın, 2018) was utilized to assess the language assessment knowledge of English as a Foreign Language (EFL) teachers. The Language Assessment Knowledge Scale (LAKS) underwent a rigorous development process involving expert review, teacher feedback, and validation by ELT and assessment experts. Following this, a pilot test with 50 teachers revealed issues of response consistency and participant engagement, prompting further refinement. Five experts then carefully evaluated each item and retained only those deemed fundamental for language teachers' assessment knowledge. As a result, the scale was pared down to 60 items, distributed across reading, listening, writing, and speaking constructs, representing a refined and validated version ready for wider implementation among language teachers. It included two main sections: demographic information and assessment knowledge questions. Participants were presented with 60 questions related to assessing reading, listening, writing, and speaking skills, to which they responded with "true," "false," or "don't know." In comparison to the original development process of the Language Assessment Knowledge Scale (LAKS), wherein a Cronbach's alpha coefficient of α= .91 was reported, the current study yielded a coefficient of α= .768. This discrepancy in reliability estimates may stem from differences in sample characteristics, testing conditions, or other methodological factors. It is important to note that reliability estimates can vary across different study populations and contexts. While the coefficient obtained in this study remains within an acceptable range, caution should be exercised when interpreting the scale's reliability in the specific context of this investigation.

The scale was converted into an online version using Google Forms. A combination of convenience and snowball sampling techniques was employed by the researcher to reach participants in both Türkiye and European countries through personal networks and contacts in the field of English language teaching. Participants completed the online form, and no personal information was required. However, participants were given the option to volunteer for the qualitative part of the study by providing their email addresses for further contact.

In the qualitative phase, individual semi-structured interviews were conducted with willing language teachers. These interviews followed an open-ended format, allowing participants to respond freely to a set of nine questions adapted from Jannati's study (2015). Semi-structured interviews provide a framework for exploration while allowing participants to express their perspectives in their own terms (Cohen et al., 2018). This qualitative approach complemented the quantitative data, offering deeper insights into participants' viewpoints and attitudes toward language assessment. Interviews were conducted using Zoom, with participants' consent for recording. For Turkish participants, interviews were conducted in Turkish, transcribed, and then translated into English. European participants were interviewed in English, and the interviews were transcribed into text format.

2.4. Data Analysis

The data analysis phase encompassed both quantitative and qualitative methods. In quantitative analysis, data were transferred to SPSS 26.0 for analysis. LAK levels were determined by participants' correct answers to more than half of the questions (30 out of 60 questions). As the developers of the scale applied in their own research, participants who gave 30 or more correct answers were accepted as assessment literate. Participants who gave correct answers below 30 were accepted as inadequate in terms of assessment knowledge. Inferential statistics were used to compare participants' LAK levels based on various factors such as country, gender, educational level, and years of experience.

In qualitative analysis, interview data were processed using MAXQDA Analytics Pro 2020 software. Hypothesis-related code schemes were created, and interview responses were numbered for the organization. Data were selectively included to support quantitative results and provide additional context. Anonymity was maintained by assigning code names to participants. Rigorous research techniques were employed to ensure validity and reliability, including having the interview questions analyzed by experts in the field, choosing interviewees from diverse countries, recording the interviews, and using open-ended questions to encourage participants to answer freely rather than just yes or no answers, member checking, and continually comparing data with the codes.

3. FINDINGS

3.1. Quantitative Findings

The quantitative phase of the study focused on assessing the general and skill-based Language Assessment Knowledge levels of English language teachers. Initially, an analysis was conducted to assess the LAK levels of participating teachers. Subsequently, a comparison was made between the general and skill-based LAK levels of teachers in Türkiye and Europe, organizing them into two distinct groups. Additionally, the study examined whether the demographic factors included in the research scale had a significant impact on the LAK levels of EFL teachers.

To determine the suitability of statistical tests for further analysis and comparisons between Türkiye and Europe, tests assessing the normality assumptions of the LAK level variable were performed.

Table 3. *Results of the tests of normality for general LAK level.*

		Kolmogorov-Smirnov			Shapiro-Wilk		
	Country	Statistic	*df*	Sig.	Statistic	*df*	Sig.
LAK level	Türkiye	.088	48	.200	.988	48	.911
	Europe	.123	46	.079	.957	46	.087

Table 3 displays the outcomes of these tests, including the Kolmogorov-Smirnov and Shapiro-Wilk tests, for both country groups. The *p*-values from these tests, exceeding .05 for both Türkiye and Europe, indicated that there was no compelling evidence to suggest substantial deviations from normality in the LAK level variable. Consequently, parametric tests assuming normality, such as the t-test or ANOVA, were employed to compare LAK levels between the two country groups.

3.1.1. *General and skill-based LAK levels of the participants*

Table 4 presents the general LAK levels of all EFL teachers who participated in the study. Descriptive statistics, including the mean, standard deviation, minimum, and maximum values, were used to determine the participants' LAK levels.

Table 4. *General LAK level of EFL teachers.*

N	Minimum	Maximum	Mean	Std. Deviation
94	7	51	32.11	7.33

As shown in Table 4, the 94 English language teachers who participated in the study answered an average of X= 32.11 questions correctly out of the 60-question scale. The lowest number of correct answers was 7, and the highest number of correct answers was 51. Additionally, the skill-based LAK levels of the participants were also analyzed, in addition to their LAK levels.

Table 5. *Skill-based LAK levels of the participants.*

	N	Minimum	Maximum	Mean	Std. Deviation
Assessing Reading	94	3	14	9.24	2.04
Assessing Listening	94	0	15	7.20	2.62
Assessing Writing	94	0	13	7.41	2.47
Assessing Speaking	94	0	13	8.25	2.44

The research scale comprised 15 questions for each skill, allowing participants to score between 0 and 15 for each skill. According to Table 5, out of the 94 English language teachers who participated in the study, the highest mean score in skill-based analysis was obtained in reading assessment (*X*= 9.24). Among the four skills that constitute the English language, the skill with the lowest mean value among the questions asked was listening (*X*= 7.20).

A one-sample t-test was used to determine whether the score was significantly high. The lowest possible score on the scale is 0, and the highest possible score is 60. Therefore, 30 was selected as the reference point, which represents half of the total score.

Table 6. *One-sample t-test results of participants' general LAK level scores.*

Mean Diff.	*df*	*t*	*p*
2.11	93	2.799	.003*

* *p*< .05

Table 6 indicates that the mean difference (2.11) between all of the participants' mean scores on the scale (*X*= 32.11) and half of the maximum score (30) was statistically significant, which suggests that their overall LAK level is high.

After discovering that the mean scores of the participant teachers were significantly high, the same one-sample t-test was performed for each skill individually. However, this time, since there were 15 questions for each skill, 7.5 was used as the reference value. This approach aimed to determine whether there was a significant difference between the reference value of each skill and the score that teachers received for that skill.

Table 7. *Skill-based one-sample t-test results.*

	Mean diff.	Mean	*df*	*t*	*p*
Assessing Reading	1.74	9.24	93	8.28	< .001*
Assessing Listening	-0.30	7.20	93	-1.10	.274
Assessing Writing	-0.09	7.41	93	-0.33	.739
Assessing Speaking	0.75	8.25	93	2.99	.004*

**p*< .05

Table 7 demonstrates that the mean difference (1.74) between the reference value (7.5), which was accepted as half of the total 15 points, and the mean score of the reading assessment skill (*X*= 9.24) indicated that the teachers' knowledge of measuring this skill was significantly high (*p*= < .001). A similar result was found for another skill, speaking, where the mean score for assessing speaking skills among the 94 teachers (*X*= 8.25) was slightly higher (0.75) than the reference score (7.5). The significance value (*p*= .004) suggests that the mean score of the participant teachers is also significantly high in evaluating this skill. However, the mean scores for assessing listening (*X*= 7.20) and assessing writing (*X*= 7.41) obtained by the teachers for the other two skills were slightly below the reference score. Based on the obtained data, it was found that the knowledge of the 94 participating teachers in the areas of assessing listening and writing was not significantly lower than the half scores, as indicated by the non-significant significance values (*p*= .274 and *p*= .739 respectively).

3.1.2. *A Comparison of the Turkish and European EFL teachers in terms of general and skill-based LAK*

The general LAK levels of two groups, Türkiye and Europe, were compared using an independent samples t-test, as shown in Table 8.

Table 8. *T-test results of General LAK levels by country of participation.*

				t-test		
Country	*N*	Mean	Std. Deviation	*t*	*df*	*p*
Türkiye	48	30.64	4.88	-2.02	92	.046*
Europe	46	33.65	9.02			

**p*< .05

It revealed a significant difference in general LAK levels between the two groups (*t*[92]=-2.02; *p*< .05). European participants (*X*=33.65, *SD*=9.02) demonstrated higher general LAK levels compared to Turkish participants (*X*=30.64, *SD*=4.88).

Additionally, the study assessed skill-based LAK levels in reading, listening, writing, and speaking. Each skill had a maximum score of 15, with 7.5 as the reference point for competence.

Table 9. *T-test results of Skill based LAK levels by country of participation.*

					t-test		
	Country	*N*	Mean	*SD*	*t*	*df*	*p*
Assessing Reading	Türkiye	48	9.50	1.71	1.23	92	.220
	Europe	46	8.97	2.32			
Assessing Listening	Türkiye	48	6.47	1.92	-2.82	92	.006*
	Europe	46	7.95	3.04			
Assessing Writing	Türkiye	48	6.83	2.02	-2.37	92	.020*
	Europe	46	8.02	2.75			
Assessing Speaking	Türkiye	48	7.83	2.15	-1.72	92	.089
	Europe	46	8.69	2.66			

**p*< .05

Table 9 presents the skill-based LAK levels for both groups. Participants from both groups demonstrated proficiency in reading and speaking skills, with no significant differences. In terms of assessing listening skills, European teachers displayed a mean score of *X*=7.95, while their Turkish counterparts exhibited a mean score of *X*=6.47 (*p*=.006). Additionally, concerning assessing writing skills, European language teachers attained a mean score of *X*=8.02, surpassing the mean score of *X*=6.83 achieved by Turkish teachers (*p*=.020).

In summary, European teachers generally exhibited higher LAK levels, especially in listening and writing skills, while Turkish teachers had a slight advantage in reading assessment. However, the differences in reading and speaking skills were not statistically significant.

3.1.3. BA program graduated

A comparative analysis of LAK levels between participants from Türkiye and Europe, based on their graduation from ELT or non-ELT BA programs, was conducted. Table 10 summarizes the findings.

Table 10. *T-test results according to BA program graduated.*

BA program graduated	Country	*N*	Mean	Std. Deviation	Sig.
ELT	Türkiye	45	30.64	4.97	.014*
	Europe	31	34.14	8.11	
	Total	76	32.18	6.65	
Non-ELT	Türkiye	3	30.66	4.04	.831
	Europe	15	32.06	10.80	
	Total	18	31.83	9.91	
Total		94	32.11	7.33	.856

*p< .05

For ELT graduates, Türkiye had a mean LAK level of X=30.64, Europe X=34.14, and the total X=32.18. The significant difference between Türkiye and Europe (p= .014) indicated variation.

Non-ELT graduates in Türkiye had a mean LAK level of X=30.66, Europe X=32.06, and the total X=31.83. No significant difference was found between Türkiye and Europe (p= .831). Overall, when considering both countries, the analysis revealed no significant difference in LAK levels between ELT and non-ELT graduates (p= .856).

3.1.4. *Testing course at university*

This analysis compared participants from Türkiye and Europe based on whether they took a testing course during their undergraduate studies, aiming to understand its impact on their LAK levels.

Table 11. *T-test results of LAK levels according to testing course at undergraduate education.*

Testing course at undergraduate education	Country	*N*	Mean	Std. Deviation	Sig.
Yes	Türkiye	27	30.96	4.62	.022*
	Europe	29	35.27	8.36	
	Total	56	33.19	7.10	
No	Türkiye	21	30.23	5.30	.796
	Europe	17	30.88	9.67	
	Total	38	30.52	7.46	
Total		94	32.11	7.33	.083

*p< .05

According to Table 11 for those who took the course, Turkish participants had a mean LAK level of X=30.96, European participants X=35.27, and the total X=33.19. A significant difference between the participants of Türkiye and Europe (p= .022) suggests the influence of the course.

Among those who did not take the course, Türkiye had X=30.23, Europe X=30.88, and the total X=30.52. No significant difference (p= .796) was observed between Türkiye and Europe in this group.

3.2. Qualitative Findings

The qualitative phase aimed to understand language assessment perceptions and practices among 4 Turkish and 4 European teachers. Interview questions adapted from Jannati (2005) explored their viewpoints and methods. The findings, divided into two sections, compare Turkish and European teachers' perspectives on language assessment (questions 1, 2, 4, 5, 8, 9) and their assessment methods (questions 3, 6, 7). This comparison revealed similarities and differences in their approaches to teaching and assessing students in different contexts.

3.2.1. *Findings about the EFL teachers' perceptions about language assessment*

An analysis of the perspectives shared by both Turkish and European participants revealed a range of opinions within the group. Regarding the need for assessment, several participants underscored its significance in monitoring student advancement, pinpointing areas requiring additional support, and providing constructive feedback to enhance teaching. For example, certain participants expressed:

> *"I believe we do need, and I believe we need assessment to inform our planning. As teachers, we need to know where our students are. That's what assessment for." (EU-4)*
>
> *"We need to prove that we are teaching English to children at some point. For this reason, we need to know whether children have learnt it or not. For this reason, there is a need, but under normal conditions, I think how much they can master the language is not fully measurable because it is very subjective." (TR-4)*

The fourth interview question, "Do you think students' scores represent what they have learned?", delves into the participants' perspectives on the relationship between students' scores and their actual learning outcomes. In response to this question, although they gave different answers and reasons, there was a consensus among the Turkish participants that grades do not represent what students have learned. Some responses of Turkish participants were:

> "Af*ter grading the exams, I look at them and I say that some students got higher than the score they should have gotten. Both the difficulty level of the questions we ask students and our education system unfortunately do not measure children in a multidimensional way." (TR-1)*
>
> *"It definitely does not represent. We cannot say by looking at an exam grade on a paper that this student got a hundred means that he knows everything." (TR-2)*

However, the answers of the European participants varied. Some said that the grades represent students' language knowledge, while others disagreed. Some of the views of the European participants are as follows:

> *"If it's a reliable test and a valid test if it's well prepared, yes, the scores should represent what students have learned. Also, bearing in mind it's a flexible test, so it can be adjusted to the special needs of students." (EU-4)*
>
> *"For my local students or other international students, usually the scores represent exactly what they have learned. I don't know what the mystery about the Turkish students is. I really don't understand what's happening. You know, either they are shy, they don't know how to interact. My Turkish students Rümeysa and Betül don't speak any foreign language, but in writing they are excellent." (EU-3)*

Responses to the question, "How do you increase your knowledge about assessment?" provided insights into the participants' strategies for enhancing their understanding and competence in assessment practices. Turkish and European participants presented varying perspectives on their approaches to professional development and lifelong learning in this context.

> *"We invite many foreign speakers. They're mainly from universities, Oxford University or Cambridge University, it depends. They explain to us how to use the textbooks and how to give assessments. Formal or informal. How to provide uh, well, some exercises. Also, we have some video tasks… And I think we in schools have a group of teachers of foreign language teachers where we decide what to assess, how to assess and how many points we are going to give, so there must be an agreement among language teachers." (EU-1)*
>
> *"I tend to read around the topic and I also, as I said at the beginning, have attended a number of seminars, but these were more like 2-3 day conferences or trainings where there were interesting speakers. So, teachers were given the chance to voice their concerns, to discuss the problems they have in classrooms, and I found it really beneficial." (EU-4)*
>
> *"I didn't attend any workshops or anything like that, I tried to read a little bit about the subject at the time, but it became so branched and knotted somewhere, I can honestly say that I gave up…* We discuss this with my friends all the time, let's say when the time comes, not all the time, but when the time comes, we talk about it. A small exchange of ideas, after a while, I mean, apart *from that, there is nothing else, to be honest." (TR-3)*

3.2.2. *Findings about the EFL teachers' classroom-based assessment practices*

The participants were questioned about their approach to informing students about rubrics, focusing on the transparency and communication of assessment criteria. The data analysis revealed unanimous agreement among both Turkish and European participants that students should be informed about the assessment rubric. This shared perspective underscores the importance of transparency and providing students with clear guidance on the criteria used for evaluating their work. It reflects a common commitment to promoting fairness and enabling students to understand and meet the expected assessment standards. Sample responses from participants include:

> *"It's a must, and I always tell my students, never, never sit for an exam if you don't know, for example, what that exam includes in the sense what type of rubrics does it have? Because you know all kinds of exams, for example, maybe they include, or they want to test different things." (EU-2)*
>
> *"I think it would be helpful for children or students, adults, whoever they are, to be aware of those rubrics, to know what the goal of the person receiving instruction is and to draw their path accordingly." (TR-3)*

When language teachers were queried about the specific language skills or components they assess, a notable contrast emerged. All European language teachers indicated that they assess a foreign language as a comprehensive whole. In contrast, Turkish participants predominantly mentioned that they focus on teaching and assessing reading, grammar, and vocabulary. This divergence was attributed to the examination system in Türkiye. Here are some responses from European English language teachers:

> *"We cover all skills reading, writing, listening, and speaking. The main focus is on improving communication skills, so mostly oral skills are, I don't know, practiced during the semester, but the final examination is usually written exam. So, the writing component is also very important for them." (EU-3)*
>
> *"At first, I was focusing on speaking and listening, but since I realized that I was stealing their time, I don't evaluate them, I don't measure them, and I don't spend much time on them. I work more for the exam. Since these are not in the exam, I focus more on vocabulary as a language component." (TR-2)*
>
> *"They make a whole. Yes, this is what I said before. They are just like the fingers, for example, of one hand and in a way, if one of them does not work, the whole hand does not work properly." (EU-2)*

4. DISCUSSION and CONCLUSION

This study aimed to assess the language assessment knowledge of in-service language teachers in Türkiye and compare them with European counterparts. It utilized a mixed-methods approach, combining quantitative data from a knowledge scale and qualitative data from interviews. The findings were discussed in relation to research questions, implications for teacher education, and connections with existing research. Gaps in the literature were identified, suggesting areas for future research, and expanding the current knowledge in the field.

4.1. Discussion of the First Research Question

The purpose of the first research question was to find out the general and skill-based language assessment knowledge level of EFL teachers working in secondary and high schools in Türkiye and Europe. The results indicated that these teachers generally possessed a relatively high level of general knowledge about language assessment. However, when their skill-based assessment proficiency was analyzed, it was clear that although they performed exceptionally well when it came to assessing speaking and reading, they did not meet the reference score when it came to assessing writing and listening. This suggests that while teachers may have theoretical knowledge about various types of language assessment and their purposes, they may lack the practical skills required to construct valid and reliable assessments, particularly in listening and writing areas. This underscores the importance of targeted professional development programs to enhance teachers' skill-based knowledge in language assessment, including the design, implementation, and evaluation of effective assessments.

The disparities in skill-based knowledge among language teachers can be attributed to the prioritization of reading and speaking skills in language teaching and assessment practices. These skills are often emphasized in curriculum, textbooks, and standardized tests, leading teachers to become more familiar and proficient in evaluating them. Conversely, listening and writing skills tend to receive less attention in educational settings, resulting in teachers having relatively less knowledge and experience in assessing these areas. Although EFL instructors acknowledge the significance of assessing oral skills, the findings of the study of Kim (2014) reveal a discrepancy between belief and practice. Despite their recognition of the importance of oral assessment, exams lack a dedicated speaking section, indicating that oral skills are not given as much importance in assessment practices. It is also worth noting that the findings of the current study align with a previous research conducted by Kırkgöz et al. (2018), which emphasized the underappreciation of listening and writing assessment among language teachers. The prevailing focus on reading and speaking skills is influenced by curriculum priorities and educational program objectives, with an emphasis on improving students' reading and speaking abilities (Altan, 2017).

Furthermore, the results of the study suggest that practical experience and on-the-job learning contribute significantly to teachers' assessment knowledge, reinforcing the idea that assessment literacy is primarily acquired through classroom practice rather than theoretical knowledge obtained during undergraduate education (Mertler, 2003). This distinction is evident when comparing the knowledge levels of in-service teachers in this study with those of pre-service teachers in another research (Çetin-Argün, 2020). The in-service teachers exhibited higher general LAL levels, likely due to their years of classroom experience. However, the study underscores the importance of enhancing teachers' understanding of assessment in writing and listening skills, as these areas have historically received less attention in teacher education programs. The findings of this study are consistent with various studies using the same assessment instrument, contributing to the reliability and validity of the results. Nevertheless, some contradictory findings in other studies highlight the complexity of the subject and the need for further research (Lam, 2015; Ölmezer-Öztürk & Aydın, 2018; Vogt & Tsagari, 2014; Xu & Brown, 2017).

4.2. Discussion of the Second Research Question

In this section, the second research question aimed to explore potential disparities in language assessment knowledge levels between Turkish and European language teachers, focusing on their general language assessment knowledge and skill-based language assessment knowledge. In aligning with the global trends in language education, the Turkish Ministry of Education has recently introduced a foreign language teaching program that bears striking similarities to those found in European countries (Turkish Ministry of Education, 2018). Emphasizing a communicative approach to assessment, the program underscores the importance of designing tasks that prioritize the practical application and production of language skills. This echoes the principles set forth by the Common European Framework of Reference for Languages (CEFR), a widely recognized framework that serves as a benchmark for language learning and assessment across Europe and beyond. By embracing these pedagogical principles, Turkey's language education program not only aligns with international standards but also fosters a learning environment that promotes effective communication and linguistic proficiency, mirroring the goals and objectives seen in European language education systems.

Analysis of the mean values revealed that European participants had higher scores than Turkish participants, with a statistically significant difference noted. When assessing skill-based knowledge, European respondents demonstrated higher mean values in all skills except for assessing reading. This indicates that, on average, European language teachers possess greater LAK compared to their Turkish counterparts, highlighting a potential knowledge gap between the two groups. However, it is essential to note that these results are based on aggregate mean values and do not necessarily represent individual performances.

Several factors may contribute to the observed disparities in LAK levels between European and Turkish language teachers. Past research suggests that differences in educational systems and resources dedicated to language assessment practices across regions can play a role (Bonnet, 2007; Cheng et al., 2004; Jones & Saville, 2009; Vogt & Tsagari, 2014; Vogt et al., 2020). European countries, known for their well-established assessment frameworks such as CEFR and ample resources, may have an advantage, leading to the higher mean values among European respondents. In contrast, studies have highlighted challenges faced by Turkey, including limited availability of standardized assessments and inadequate resources allocated to language assessment (Büyükkarcı, 2016; Krajka, 2019; Mede & Atay, 2017), which may affect the performance of Turkish participants in these assessments. A research carried out by Ünlücan-Tosun and Glover (2020) found that Turkish language instructors expressed a lack of confidence in integrating CEFR levels into their classroom assessments. Additionally, they noted that course materials lacked sufficient guidance for effectively implementing the CEFR.

Additionally, the findings align with the study conducted by Çakır (2020), which examined language teachers' beliefs about assessment types, content, and skills across different countries. Çakır's study revealed that while beliefs about assessment types and content did not significantly differ, variations were observed in the reasons for utilizing classroom assessment among different countries. These findings, in conjunction with the present study, emphasize the influence of country-specific factors on language assessment practices and outcomes. Similarly, Cheng et al. (2004) conducted research across different countries and identified diverse assessment methods and procedures employed in ESL/EFL teaching and learning. This diversity in assessment approaches underscores the role of context and culture in shaping assessment practices in ESL/EFL education.

Although not statistically significant, Turkish participants showed superior performance in assessing reading skills. This could be attributed to the examination-oriented approach of the Turkish education system, which places a strong emphasis on reading comprehension skills, frequently assessed in high-stakes examinations (Hatipoğlu, 2010).

4.3. Discussion of the Third Research Question

The third research question sought to investigate the impact of the country and two demographic factors, educational background, and testing course at undergraduate education, on both the overall language assessment knowledge level and its skill-based components. The purpose of the investigation was to determine how these factors affect teachers' proficiency in language assessment. We can gain insights into the relationship between demographic features and LAK levels by examining differences across countries. The findings shed light on the extent to which country and specific demographic factors influence variations in language assessment knowledge, both at the general and skill-specific levels.

The study examined assessment knowledge and its relationship with the educational background of BA program graduates. Results show that participants who completed English Language Teaching (ELT) programs in Europe had significantly higher assessment knowledge levels compared to those in Turkey. However, there was no significant difference in assessment knowledge levels between participants from Turkey and Europe who completed non-ELT BA programs. This suggests that the country factor played a role in ELT program graduates' assessment knowledge but not in non-ELT program graduates. Similar studies by Genç et al. (2020) and Kaya and Mede (2021) found no significant difference in assessment literacy scores between ELT and non-ELT program language teachers. This suggests that individual factors like motivation, effort, and language proficiency may have a stronger influence on assessment performance than the specific program type. Additionally, the study questions the effectiveness of assessment courses within ELT programs, aligning with Hatipoğlu's (2015) findings that despite extensive exposure to English language exams, ELT students had limited knowledge about testing in general and English language testing specifically.

The study found that the presence of a testing course during undergraduate education did not lead to a significant difference in assessment knowledge scores between language teachers who had taken the course and those who had not when considering the country factor. This suggests that having a testing course alone may not substantially impact language teachers' assessment knowledge levels, and several factors like course effectiveness, practical application, available support, and individual differences among teachers may be at play.

Furthermore, when focusing on participants who took the testing course, those from Europe had higher mean assessment knowledge levels compared to their Turkish counterparts. This discrepancy raises the possibility that the efficacy or nature of testing programs may differ between these areas, which could affect the assessment literacy of language instructors. According to Şahin (2009), a single LTA course is insufficient for adequately enhancing the assessment knowledge of prospective language educators to handle the demanding and crucial responsibility of consistently evaluating their students for both summative and formative assessment objectives.

These findings align with the need to reevaluate the role and effectiveness of testing courses in teacher training programs worldwide. A study conducted by Ölmezer-Öztürk and Aydın (2018) also found that having a separate testing course during BA degree education does not significantly impact the LAK levels of language teachers. Additionally, insights from Stiggins (1995) suggest that taking an educational testing and measurement course may not effectively prepare teachers for the practical realities of classroom life. This collective perspective underscores the importance of reevaluating the role of testing courses in equipping teachers with the necessary skills for assessment practices in real classroom contexts.

4.4. Discussion of the Fourth Research Question

When comparing Turkish and European EFL teachers' opinions on the value of assessment in language classes, it became clear that different contexts—cultural and educational—had an impact on their viewpoints. Assessment was emphasized by Turkish teachers as a means of understanding student learning and fulfilling exam-related requirements. Lam (2015) emphasizes how putting too much emphasis on exams can make learning less important. European educators understand the value of assessment in determining student proficiency and getting them ready for national exams. The impact of participants' prior experiences as language learners and teachers on their perceptions of assessment is noted by O'Loughlin (2006).

Turkish and European EFL teachers held opposing views on whether scores accurately represent students' learning. Turkish teachers expressed skepticism, citing environmental factors and exam-related stress as limitations in score accuracy. They pointed to cases of competent students struggling with exams. In contrast, most European teachers believed scores were accurate, highlighting potential regional differences in grading policies. This raises concerns about assessment practices aligning with the broader goals of communicative teaching approaches in language education (DeLuca et al., 2017; Gkogkou & Kofou, 2021).

The perspectives of Turkish and European EFL teachers revealed disparities in the skills assessed in their classes. Turkish educators prioritized reading and grammar assessments, considering them simpler and exam-relevant. In contrast, European teachers emphasized evaluating all language skills, particularly focusing on oral communication, aligning with the assessment for learning culture prevalent in Western countries (Xu & Brown, 2016). The variations can be attributed to the contrasting assessment cultures, with Western countries emphasizing learning-oriented assessment and East Asian educational systems prioritizing high-stakes tests and rote memorization.

The study examined how Turkish and European EFL teachers approach enhancing their assessment knowledge. Turkish teachers emphasized the significance of in-service training for their professional development, and some recognized the value of academic courses, like master's degree programs, for deepening their understanding of assessment. In contrast, European teachers outlined diverse methods for expanding their assessment knowledge, including reading assessment books, attending workshops, seminars, and engaging in international conferences and partnerships. Herrera and Macias (2015) underscore the importance of integrating assessment literacy into language teacher education programs, emphasizing the need for continuous development and commitment to assessment knowledge among both novice and experienced educators.

4.5. Conclusion

In conclusion, the goal of this study was to reveal language teachers' assessment knowledge levels and compare Turkish in-service language teachers with teachers from other European countries in terms of their language assessment knowledge levels, and their in-class assessment practices. The results showed that, despite having a generally high level of knowledge about language assessment, EFL teachers in both country groups had varying skill-based levels. Teachers performed particularly well on the reading and speaking assessments, but less well on the listening and writing assessments.

Moreover, the results of this study not only confirmed the existence of varying levels of language assessment knowledge between Turkish and European language teachers but also revealed a notable disparity favoring the European group. With the exception of reading assessment, European participants showed higher mean values in both general and skill-based LAK. However, despite the observed differences in language assessment knowledge, it is noteworthy that no significant difference was found between the Turkish and European language teachers in terms of their assessing reading and speaking abilities. This finding suggests that, in these particular language skills, both groups demonstrated comparable levels of proficiency. On the other hand, when it came to assessing listening and writing skills, significant differences emerged, with the European teachers exhibiting higher expertise.

This study also sought to explore and compare the general and skill-based assessment knowledge levels of participants across different countries, taking into account various demographic factors. The findings revealed noteworthy differences in the areas of the participants' fields of study during their BA programs and, whether they had taken a testing course at university.

Based on the qualitative findings from the interviews, a clear difference in approaches to assessing the four language skills emerged between Turkish and European language teachers. Turkish teachers expressed concerns about national exams, causing them to focus solely on exam-oriented skills while ignoring the comprehensive assessment of all four skills. European teachers, on the other hand, prioritized communication skill development and recognized the importance of assessing all four language skills. This disparity can be attributed to different priorities for lifelong learning. When we examine the answers of the participants, it is clear that Turkish teachers placed little emphasis on acquiring new knowledge about their profession, whereas Europeans valued lifelong learning as a means of expanding their knowledge. Furthermore, Turkish teachers were skeptical regarding the representation of students' knowledge through grades, whereas Europeans saw grades as more important indicators of students' understanding. Despite these differences, Turkish and European teachers agreed on the importance of assessment, the use of rubrics, and various in-class assessment methods.

4.6. Limitations of the Study

The study has several limitations. Firstly, the online data collection method aimed to collect responses from a large number of English language teachers; however, only a small sample of language teachers from Türkiye and only seven countries in Europe completed the scale, which might not accurately represent the entire population of English language teachers in the study area. Secondly, the voluntary participation in the qualitative part of the research led to a limited number of interviews, potentially reducing the applicability and inclusiveness of the findings to a broader population. Furthermore, the online format of the scales could have resulted in a non-representative sample, and participants' English proficiency and interpersonal communication skills may have influenced the accuracy of their responses during the interviews.

Additionally, it is crucial to acknowledge that the education systems in diverse countries may vary significantly. These differences could impact the experiences and perspectives of English language teachers, introducing an additional layer of complexity to the interpretation of our findings. Despite our efforts to address these variations, it is important to interpret the study's outcomes with caution, given the low effect size observed, and to consider them within the context of the specific educational landscape in the study area.

4.7. Suggestions for Further Studies

The study's findings offer guidance for future language assessment research. Subsequent studies may explore specific factors, including cultural norms, educational frameworks, and institutional contexts, influencing the observed differences in language assessment knowledge between Turkish and European teachers. Expanding the sample size in future studies is crucial for broader applicability, allowing for a more comprehensive analysis of language assessment knowledge among English language teachers in both Turkish and European settings. Additionally, investigating students' perceptions and attitudes towards language assessment can inform tailored assessment design. Longitudinal studies could provide insights into the enduring impact of language assessment practices on students' language development and real-world language skill application, offering avenues for further research.

4.8. Pedagogical Implications of the Study

The findings of the study hold significant pedagogical implications for English language teaching in Turkish and European contexts, particularly regarding assessment knowledge among teachers. To start, targeted professional development programs are essential for foreign language teachers in Türkiye to enhance their language assessment knowledge. Both Turkish and European EFL teachers displayed varying levels of skill-based language assessment knowledge. Offering comprehensive training can improve teachers' grasp and application of assessment principles, ensuring more accurate and balanced language development in classrooms.

Additionally, the notable difference in language assessment knowledge between Turkish and European teachers underscores the importance of emphasizing assessment knowledge in Turkish language teacher education programs. Integrating assessment-focused courses and workshops can bridge this gap, equipping Turkish teachers with the necessary skills to assess language proficiency across all skill areas. Promoting a culture of ongoing professional development and lifelong learning can further enhance assessment knowledge among Turkish teachers. Aligning Turkish teachers' views on grades as representations of knowledge with those of European teachers can positively impact assessment practices. Providing guidance and training on interpreting and utilizing grades as indicators of language proficiency can lead to more meaningful assessment outcomes and a more student-centered approach.

Addressing these pedagogical implications can support English language teaching, fostering more effective and equitable language assessment practices that support comprehensive language development and student success.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number:** Süleyman Demirel University, 2022-6/126.14.

Contribution of Authors

Samet Fındıklı: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing-original draft. **Kağan Büyükkarcı:** Methodology, Supervision, and Validation.

Orcid

Samet Fındıklı https://orcid.org/0000-0002-3477-7980
Kağan Büyükkarcı https://orcid.org/0000-0002-7365-0210

REFERENCES

Altan, M.Z. (2017). Globalization, English language teaching & Türkiye. *International Journal of Languages' Education*, *5*(4), 764–776. https://doi.org/10.18298/ijlet.2238

Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L.F. (2004). *Statistical analyses for language assessment book (Cambridge Language Assessment)*. Cambridge University Press.

Bonnet, G. (2007). The CEFR and education policies in Europe. *The Modern Language Journal*, *91*(4), 669–672. https://doi.org/10.1111/j.1540-4781.2007.00627_7.x

Boyles, P. (2005). Assessment literacy. In M. Rosenbusch (Ed.), *National assessment summit papers* (pp. 18–23). Iowa State University.

Brown, H.D. (2000). *Principles of language learning and teaching*. Longman.

Büyükkarcı, K. (2014). Assessment beliefs and practices of language teachers in primary education. *International Journal of Instruction, 7*(1), 107-120.

Büyükkarcı, K. (2016). Identifying the areas for English language teacher development: A study of assessment literacy. *Pegem Journal of Education and Instruction, 6*(3), 333–346. https://doi.org/10.14527/pegegog.2016.017

Cheng, L., & Fox, J. (2017). *Assessment in the language classroom: Teachers supporting student learning (Applied linguistics for the language classroom)* (1st ed. 2017). Springer.

Cheng, L., Rogers, T., & Hu, H. (2004). ESL/EFL instructors' classroom assessment practices: Purposes, methods, and procedures. *Language Testing*, *21*(3), 360-389. https://doi.org/10.1191/0265532204lt288oa

Clapham, C. (2000). Assessment and testing. *Annual Review of Applied Linguistics*, *20*, 147–161. https://doi.org/10.1017/s0267190500200093

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

Colby, D.C. (2010). *Using "Assessment for learning" practices with pre-university level ESL students: A mixed methods study of teacher and student performance and beliefs* [PhD Dissertation]. McGill University, Montreal.

Coombe, C. (2018). *An A to Z of second language assessment: How language teachers understand assessment concepts*. British Council.

Cowan, J. (1998). *On becoming an innovative university teacher*. RHE & Open University Press.

Creswell, J.W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches*. SAGE Publications.

Creswell, J.W., & Clark, V.L.P. (2007). *Designing and conducting mixed methods research*. SAGE Publications.

Çakır, N. (2020). *A comparative analysis of teachers' beliefs and practices on the assessment of 4th grade-EFL students in Türkiye, Italy and Finland* [MA thesis]. Uludağ University, Bursa.

Çetin-Argün, B. (2020). *Language assessment knowledge of preservice teachers of English as a foreign language* [MA Thesis]. Çağ University, Mersin.

Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing*, *25*(3), 327–347. https://doi.org/10.1177/0265532208090156

DeLuca, C., & Klinger, D.A. (2010). Assessment literacy development: Identifying gaps in teacher candidates' learning. *Assessment in Education: Principles, Policy & Practice, 17*(4), 419–438. https://doi.org/10.1080/0969594X.2010.516643

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2015). Teacher assessment literacy: A review of international standards and measures. *Educational Assessment*, *Evaluation and Accountability*, *28*(3), 251–272. https://doi.org/10.1007/s11092-015-9233-6

Douglas, D. (2009). *Understanding language testing* (1st ed.). Routledge.

EF EPI (2022). *EF English Proficiency Index. The world's largest ranking of countries and regions by English skills*. https://www.ef.com/wwen/epi/

Fox, J. (2016). Using portfolios for assessment/alternative assessment. In E. Shohamy, I. Or, & S. May (Eds.), *Language testing and assessment. Encyclopedia of language and education* (3rd ed., pp. 135-147). Springer. https://doi.org/10.1007/978-3-319- 02261-1_9

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly*, *9*(2), 113–132. https://doi.org/10.1080/15434303.2011.642041

Genç, E., Çalışkan, H., & Yüksel, D. (2020). Language assessment literacy level of EFL teachers: A focus on writing and speaking assessment knowledge of the teachers. *Sakarya University Journal of Education*, *10*(2), 274-291. https://doi.org/10.19126/suje.626156

Giraldo, F. (2018). Language assessment literacy: Implications for language teachers. *Profile: Issues in Teachers´ Professional Development*, *20*(1), 179-195. https://doi.org/10.15446/profile.v20n1.62089

Giraldo, F. (2021). A reflection on initiatives for teachers' professional development through language assessment literacy. *Profile: Issues in Teachers' Professional Development, 23*(1), 197–213. https://doi.org/10.15446/profile.v23n1.83094

Gkogkou, E., & Kofou, I. (2021). A toolkit for the investigation of Greek EFL teachers' assessment literacy. *Languages*, *6*(188), 1-27. https://doi.org/10.3390/languages6040188

Gotch, C.M., & French, B.F. (2014). A systematic review of assessment literacy measures. *Educational Measurement: Issues and Practice*, *33*(2), 14-18. https://doi.org/10.1111/emip.12030

Green, A. (2013). *Exploring language assessment and testing: Language in action*. Routledge.

Harding, L., & Kremmel, B. (2016). Teacher assessment literacy and professional development. *Handbook of Second Language Assessment*, 413-428. https://doi.org/10.1515/9781614513827-027

Hatipoğlu, Ç. (2010). Summative evaluation of an English language testing and evaluation course for future English language teachers in Türkiye. *English Language Teacher Education and Development (ELTED), 13*, 40-51. http://www.elted.net/uploads/7/3/1/6/7316005/v13_5hatipoglu.pdf

Hatipoğlu, Ç. (2015). English language testing and evaluation (ELTE) training in Türkiye: expectations and needs of pre-service English language teachers. *ELT Research Journal*, *4*(2), 111–128. https://dergipark.org.tr/en/download/article-file/296302

Heaton, J.B. (1989). *Writing English language tests (Longman handbooks for language teachers)*. Longman Pub Group.

Herrera, L. & Macías, D. (2015). A call for language assessment literacy in the education and development of teachers of English as a foreign language. *Colombian Applied Linguistics Journal*, *17*(2), 302-312. https://doi.org/10.14483/udistrital.jour.calj.2015.2.a09

Hughes, A. (2003). *Testing for language teachers (2nd ed.).* Cambridge University Press.

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing*, *25*(3), 385-402. https://doi.org/10.1177/0265532208090158

Inbar-Lourie, O. (2017). Language assessment literacy. In E. Shohamy, L. Or, & S. May (Eds.), *Language testing and assessment* (pp. 257-270). Springer International Publishing AG. https://doi.org/10.1007/978-3-319-02261-1_19

Jannati, S. (2015). ELT teachers' language assessment literacy: Perceptions and practices. *The International Journal of Research in Teacher Education*, *6*(2), 26-37.

Jones, N.D., & Saville, N. (2009). European language policy: Assessment, learning, and the CEFR. *Annual Review of Applied Linguistics*, 29, 51-63. https://doi.org/10.1017/s0267190509090059

Kaya, T., & Mede, E. (2015). Exploring language assessment literacy of EFL instructors in language preparatory programs. *İstanbul Aydın University Journal of Education Faculty*, *7*(1), 163–189. https://doi.org/10.17932/iau.efd.2015.013/efd_v07i008

Kırkgöz, Y., Babanoğlu, M.P., & Ağçam, R. (2018). Turkish EFL teachers' perceptions and practices of foreign language assessment in primary education. *Journal of Education and E-learning Research*, *4*(4), 163–170.

Kim, A.A. (2014). Examining how teachers' beliefs about communicative language teaching affect their instructional and assessment practices: A qualitative study of EFL University instructors in Colombia. *RELC Journal*, *45*(3), 337-354. https://doi.org/10.1177/0033688214555396

Krajka, J. (2019). L1 use in language tests: Investigating cross-cultural dimensions of language assessment. *Journal of Intercultural Management*, *11*(2), 107-133. https://doi.org/10.2478/joim-2019-0011

Lam, R. (2015). Language assessment training in Hong Kong: Implications for language assessment literacy. *Language Testing*, *32*(2), 169-197. https://doi.org/10.1177/0265532214554321

Lor, P. (2019). *International and comparative librarianship: Concepts and methods for global studies*. De Gruyter Saur. https://doi.org/10.1515/9783110267990

Malone, M. (2011). Assessment literacy for language educators. *CALDigest*, October, pp. 1-2.

Mede, E., & Atay, D. (2017). English language teachers' assessment literacy: The Turkish context. *Dil Dergisi*, *168*(1), 43–60. https://doi.org/10.1501/dilder_0000000237

Mertler, C.A. (2003). *Pre-service versus in-service teachers' assessment literacy: Does classroom experience make a difference?* [Paper Presentation]. Annual Meeting of the Mid-Western Educational Research Association, Columbus, OH, USA.

Mertler, C.A., & Campbell, C. (2005). *Measuring teachers' knowledge & application of classroom assessment concepts: Development of the "Assessment Literacy Inventory"* [Paper Presentation]. Annual Meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Mohan, R. (2022). *Measurement, evaluation and assessment in education (1st ed.).* PHI Learning PVT. LTD.

Nagai, N., Birch, G.C., Bower, J.V., & Schmidt, M.G. (2020). *CEFR-informed learning, teaching and* assessment: A practical guide (Springer Texts in Education) (1st ed. 2020 ed.). Springer.

O'Loughlin, K. (2006). Learning about second language assessment: Insights from a postgraduate student on-line subject forum. *University of Sydney Papers in TESOL*, 1, 71–85.

Ölmezer-Öztürk, E., & Aydin, B. (2018). Toward measuring language teachers' assessment knowledge: Development and validation of Language Assessment Knowledge Scale (LAKS). *Language Testing in Asia*, *8*(1). https://doi.org/10.1186/s40468-018-0075-2

Ölmezer-Öztürk, E., & Aydın, B. (2019). Voices of EFL teachers as assessors: Their opinions and needs regarding language assessment. *Journal of Qualitative Research in Education, 7*(1), 373-390. doi:10.14689/issn.21482624.1.7c1s.17m

Popham, W.J. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator*, *46*(4), 265-273.

Rogier, D. (2014). Assessment literacy: Building a base for better teaching and learning. *English Teaching Forum 3* (pp. 3-13).

Sarıyıldız, G. (2018). *A study into language assessment literacy of preservice English as a foreign language teachers in Turkish context* [MA thesis]. Hacettepe University, Ankara.

Scarino, A. (2013). Language assessment literacy as self-awareness: *Understanding* the role of interpretation in assessment and in teacher learning. *Language Testing*, *30*(3), 309–327. https://doi.org/10.1177/0265532213480128

Sevimel-Şahin, A. (2019). *Exploring foreign language assessment literacy of pre-service English language teachers* [Doctoral dissertation]. Anadolu University, Eskişehir, Türkiye.

Sevimel-Sahin, A., & Subaşı, G. (2021). Exploring foreign language assessment literacy training needs of pre-service English language teachers. *International Online Journal of Education and Teaching (IOJET)*, *8*(4), 2783-2802.

Shepard, L.A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4–14.

Siegel, M.A., & Wissehr, C. (2011). Preparing form the plunge: Preservice teachers' assessment literacy. *Journal of Science Teacher Education, 22*(4), 371-391. https://doi.org/10.1007/978-3-319-02261-1_19

Stiggins, R. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan*, *77*(3), 238–245.

Şahin, S. (2009). *An analysis of English language testing and evaluation course in English language teacher education programs in Turkey: Developing language assessment literacy of pre-service EFL teachers* [Master's Thesis]. Middle East Technical University.

Tamerer, R.B. (2019). *An investigation of Turkish pre-service EFL teachers' language assessment literacy* [MA thesis]. Kocaeli University, Kocaeli.

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing*, *30*(3), 403-412. https://doi.org/10.1177/0265532213480338

Thomas, J., Allman, C., & Beech, M. (2004). Assessment for the diverse classroom: A handbook for teachers. *Tallahassee, FL: Florida Department of Education, Bureau of Exceptional Education and Student Services.*

Turkish Ministry of Education. (2018). *High-school English language teaching program.* [Curriculum]. https://mufredat.meb.gov.tr/ProgramDetay.aspx?PID=342

Ünlücan-Tosun, F., & Glover, P. (2020). How do school teachers in Turkey perceive and use the CEFR? *International Online Journal of Education and Teaching (IOJET)*, *7*(4), 1731-1739.

Vogt, K., Guerin, E., Sahinkaraks, S., Pavlou, P., Tsagari, D., & Afiri, Q. (2008). *Assessment literacy of foreign language teachers in Europe.* Poster session presented at the 5th Annual meeting of EALTA, Athens, Greece.

Vogt, K., & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly*, *11*(4), 374-402. https://doi.org/10.1080/15434303.2014.960046

Wardani, W.O., Sukyadi, D., & Purnawarman, P. (2021). Exploring teacher assessment literacy in EFL classroom. *Proceedings of the 2nd International Conference on Progressive*

Education, ICOPE 2020, 16–17 October 2020, Universitas Lampung, Bandar Lampung, Indonesia. https://doi.org/10.4108/eai.16-10-2020.2305253

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

Xu, Y., & Brown, G.T.L. (2016). Teacher assessment literacy in practice: A reconceptualizati on. *Teaching and Teacher Education*, *58*(1), 149-162. https://doi.org/10.1016/j.tate.2016.05.010

Xu, Y., & Brown, G.T.L. (2017). University English teacher assessment literacy: A survey-test report from China. *Papers in Language Testing and Assessment*, *6*(1), 133–158.

Yetkin, C. (2015). *An investigation on ELT teacher candidates' assessment literacy* [MA thesis]. Cag University, Mersin.

International Journal of Assessment Tools in Education
2024, Vol. 11, No. 2, 368–387

https://doi.org/10.21449/ijate.1376160

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

The difference between estimated and perceived item difficulty: An empirical study

Ayfer Sayın[1*], **Okan Bulut**[2]

[1]Gazi University, Gazi Faculty of Education, Department of Educational Sciences, Ankara, Türkiye
[2]University of Alberta, Faculty of Education, Department of Educational Psychology, Alberta, Canada

ARTICLE HISTORY

Received: Oct. 15, 2024

Accepted: May 02, 2024

Keywords:

Item difficulty,
Expert prediction,
Feedback,
Language test,
LGS.

Abstract: Test development is a complicated process that demands examining various factors, one of them being writing items of varying difficulty. It is important to use items of a different range of difficulty to ensure that the test results accurately indicate the test-taker's abilities. Therefore, the factors affecting item difficulty should be defined, and item difficulties should be estimated before testing. This study aims to investigate the factors that affect estimated and perceived item difficulty in the High School Entrance Examination in Türkiye and to improve estimation accuracy by giving feedback to the experts. The study started with estimating item difficulty for 40 items belonging to reading comprehension, grammar, and reasoning based on data. Then, the experts' predictions were compared with the estimated item difficulty and feedback was provided to improve the accuracy of their predictions. The study found that some item features (e.g., length and readability) did not affect the estimated difficulty but affected the experts' item difficulty perceptions. Based on these results, the study concludes that providing feedback to experts can improve the factors affecting their item difficulty estimates. So, it can help improve the quality of future tests and provide feedback to experts to improve their ability to estimate item difficulty accurately.

1. INTRODUCTION

Item difficulty is essential not only for test development but also for creating a large item pool (Bock et al., 1988; Segall et al., 1997), providing items of varying difficulty (Huang et al., 2017), creating equivalent test forms (Förster & Kuhn, 2021; Kolen & Brennan, 2004; Van der Linden & Pashley, 2009), developing adaptive testing (Hontangas et al., 2000; Van der Linden & Pashley, 2009), and establishing Angoff standard setting (Berk, 1986; Dalum et al., 2022). The factors affecting item difficulty are first to be examined to determine item difficulty.

Understanding the factors that affect item difficulty can help test developers have better control over the statistical features of the items they create. This knowledge could also help reduce the need for pre-application, improve test statistics control, such as item difficulty distributions, and enhance test specifications (Bejar, 1983; Boldt, 1998). Therefore, there are many studies examining the factors affecting item difficulty. Some research stated that the item difficulty is affected by the item types (Freedle & Kostin, 1993), item length (Lin et al., 2021), readability

*CONTACT: Ayfer SAYIN ✉ ayfersayin@gazi.edu.tr Gazi Faculty of Education, Department of Educational Sciences, Ankara, Türkiye

e-ISSN: 2148-7456

(AlKhuzaey et al., 2021; Lumley et al., 2012), taxonomy (Hamamoto Filho et al., 2020), degree of cognitive complexity (Valencia et al., 2017), visual content (Stiller et al., 2016) and several other variables. Despite this examination, predicting item difficulty remains challenging in educational assessment and empirical attempts to explain low variance (El Masri et al., 2017; Ferrara et al., 2022). Because it is difficult to say that any variable is directly effective on item difficulty in every test. Therefore, there is a need to predict item difficulties before each test is administered.

1.1. Item Difficulty Prediction Methods

Several methods are used to predict item difficulty, including pre-testing, automatic estimation methods, and expert opinion (Attali et al., 2014). A pretest is often very costly and time-consuming and can potentially expose the items to test takers. For automatic estimation, factors affecting item difficulty must be defined, but these factors can vary based on item features, content, and the target population of test takers. The third method of estimating item difficulty is the judgement of subject matter experts (SME), which is often subjective and difficult to scale. However, teachers use their judgment in preparing classroom tests, and some testing centers seek expert opinions when developing achievement tests (e.g., licensure examinations). The Angoff standard-setting method, especially for medical education and high-stake examinations, involves consulting a group of experts in the relevant subject area to establish a standard setting that predicts the difficulty of test items and the overall exam (Benton, 2020; Berk, 1986; Dalum et al., 2022).

The information obtained from SMEs can be evaluated together with the information obtained from other sources and can be used for automatic estimation of item statistics (Attali et al., 2014; Mislevy et al., 1993; Swaminathan et al., 2003). If pilot or field testing cannot be performed, the overall test difficulty is usually adjusted based on the SMEs' judgment of item difficulty (Choi & Moon, 2020). Therefore, experts need to know the factors affecting item difficulty. Especially with the recent increase in cognitive diagnostic assessments, the importance of expert prediction on item content has come to the fore (Liu & Read, 2021). Furthermore, the information obtained from SMEs can be evaluated together with those obtained from other sources and can be used to estimate item statistics (Swaminathan et al., 2003).

Predicting item difficulty is multifaceted and influenced by various factors, including text complexity, decision-making processes, test item intricacies, and the diversity of examinee populations. Studies by Embretson and Wetzel (1987) underscore the importance of incorporating a comprehensive approach to accurately gauge item difficulty, emphasizing text-related variables, as further supported by Freedle and Kostin (1993). The utility of response time data, particularly in naming tasks, was demonstrated by Fergadiotis et al. (2018), highlighting the significance of behavioural metrics. Expertise emerges as a crucial factor, with Berenbon and McHugh (2023) showing that trained item writers excel in predicting item difficulty, contrasting with the challenges highlighted by Sydorenko (2011) and Giguère et al. (2022) regarding the limitations of hypothesized difficulty and the uncorrelated nature of difficulty in Rasch models. The work of Kibble and Johnson (2011) and Herzog et al. (2021) further illustrates the challenges in prediction due to significant individual variation and the limited predictive power of certain item characteristics. Therefore, the endeavour to predict item difficulty is difficult because of a multitude of factors, including the intricacies of text and decision processes, the diversity of test items and populations, and the evolving nature of educational standards and curricula, coupled with the essential roles of response time data, expertise, and individual variability.

Items are still commonly written by experts in high-stakes tests, as well as in-class tests, so the experts who write items must have detailed knowledge about the factors that affect the difficulty of the items. Additionally, improving experts' (teacher, item writer, professor, etc.) ability to predict how students will perform on assessments and how individual items will perform can help ensure greater consistency in the assessments over time. In other words, understanding the factors affecting item difficulty, such as linguistic and cognitive factors, and why certain items are less predictable can guide the writer and practitioners (Davies, 2021). The present research aims to identify the variables that explain SMEs' prediction of item difficulty and provide feedback to improve their predictions in the High School Entrance Examination language test in Türkiye.

1.2. Comparing the Estimated and Perceived Item Difficulty

Previous studies show that the relationships between estimated and perceived item difficulty depend on variables such as subject matter and profession of the predictors. It also shows that many test-related factors affect the difference between estimated and perceived item difficulty. For instance, Hamamoto Filho et al. (2020) investigated the psychometric properties of items used in a progress test, a longitudinal assessment of students' knowledge. The items were classified according to Bloom's taxonomy, and judges' estimates were used to assess their difficulty. The study was conducted in ten medical schools in Brazil. The study suggests that items with high-level taxonomy may better discriminate against students and that a panel of experts can provide coherent reasoning regarding the item's difficulties. Similarly, Choi and Moon (2020) investigated the factors that impact the difficulty of the reading and listening sections of the English test and found high relation difficulties. The predicted difficulties by both native and non-native speakers were significant predictors of observed difficulty. Le Hebel et al. (2019) focused on exploring the abilities of science teachers in predicting the performance of middle-low achieving students in inquiry-based tasks from the PISA science test. The study utilized a questionnaire-based approach with a sample of 125 French science and technology teachers. The study's findings suggest that the teachers could predict the difficulty levels of inquiry tasks for medium-low achieving students. Additionally, they identified potential sources of difficulty or ease in the tasks. Wauters et al. (2012) compared alternative methods to IRT-based calibration for estimating item difficulty used in adaptive item-based learning environments. The research assessed how well seven different ways of estimating something performed. To do this, the estimates produced by each method were compared to item difficulty that was obtained from a larger study conducted by Selor, which is the selection agency for the Belgian government. The larger study involved 2961 participants who took a test. According to the results, learners are more accurate than experts in predicting the item difficulties. However, this difference disappears when learners and experts are asked to rank the items based on their difficulty. Sydorenko (2011) purposed to investigate the accuracy of item difficulty prediction made by item writers and to examine whether factors affecting item writer judgments corresponded to actual item difficulty predictors. The study used online videos containing conversational dialogues centred on pragmatic functions and was completed by 35 students in their second, third, and fourth years of learning Russian. The outcomes revealed that the predicted item difficulty had a weak but significant association with the estimated item difficulty. The study also discovered that the item writer successfully anticipated linguistic focus and response format but did not consider the influence of topical knowledge.

1.3. Giving SMEs Feedback on Item Difficulty

The results of previous research show that based on understanding the underlying reasons for expert opinions, giving feedback or training to the experts for predicting item difficulty leads to improved prediction accuracy. For example, as part of a project, Davies (2021) explored the ability of examiners and item writers to predict the item difficulty in language tests, focusing on Welsh tests. The study aims to identify the factors affecting item difficulty and understand why certain items are less predictable. The method includes a panel of 13 participants who predicted the difficulty of 320 items on a 5-point scale, followed by a workshop and a second prediction round. The research also investigates whether the workshop training improves predictions and asks panellists to predict their confidence in their judgments for each item. It found that participants' predictions were correlated with estimated value, and the feedback improved the experts' perceived item. Similarly, González-Brenes et al. (2014) introduced a new method called Feature Aware Student Knowledge Tracing (FAST) that integrates general features into Knowledge Tracing, the standard for inferring student knowledge from performance data. It was determined that teachers' predictions of the difficulty of the tasks improved by 25% with the FAST method they used. Fortus et al. (2013) aimed to identify the factors that affect the difficulty level of multiple-choice items, particularly reading comprehension items, in the English test of Israel's Inter-University Psychometric Entrance Test. The researchers found that the vocabulary and grammatical complexity of the reading comprehension text had the greatest impact on item difficulty. Other variables significantly correlated with difficulty in reading comprehension items include the amount of processing, type of item, length of distractors, and level of vocabulary in stem and distractors. The study also aimed to provide feedback to experts in the context of factors affecting item difficulty, and it found that the correlation between raters' predictions of item difficulty and estimated item difficulty significantly improved from .24 to .82 after giving feedback to the experts. In a similar way, Lumley et al. (2012) discussed the importance of understanding the features that influence the difficulty of reading tasks to improve the reliability of a priori estimates of item difficulty in reading tests. This research developed a schema for describing the difficulty reading items used in PISA. This schema includes 10 variables that can be used by trained raters to predict item difficulty with reasonable success. 5 experts who participated in the study found that raters trained on the schema developed in the research showed better agreement in their predictions. Hambleton et al. (2003) aimed to create and evaluate anchor-based judgmental methods allowing LSAT test specialists to predict item difficulty statistics. The results indicated that even though it needed a long process, the specialists believed they could be trained to predict item difficulty accurately. They demonstrated some proficiency in doing so. After the training, the average error in the predictions of item difficulty ranged from about 11-13%. The panellists found the discussions helpful and were able to improve the prediction of item difficulty. Furthermore, the study discovered that test specialists benefited from the descriptions of items and information about the item statistics of many items in the training. Similarly, MacGregor et al. (2008) stated that participants' prediction of item difficulty improved after feedback; the correlation between estimated and perceived item difficulty was .48 to .65.

1.4. Present Study

Previous studies show that the factors that affect the difficulty of items in different tests differ. They also indicated that several variables affect the accuracy of experts' item difficulty perceptions, and experts can provide valuable information in estimating item difficulty. It reveals that feedback provided to experts improves their item difficulty predictions. Previous research in this field has typically concentrated on examining tests within a single content domain, such as exclusively featuring cloze tests or reading comprehension items. The current study marks a significant departure from this trend by investigating a test encompassing three content domains: grammar, reasoning, and reading comprehension. This holistic approach allows for a more comprehensive analysis of item difficulty, considering the varied cognitive skills required across different test items. This study aims to improve expert estimates of the item difficulty in a language test containing three different content domains (reading comprehension, grammar, and reasoning) in a high-stake test. In addition, this study focuses on a test in the Turkish language. Research has shown that language and cultural factors can significantly influence the difficulty of test items. Oliveri and Ercikan (2011) underscore the pronounced effects of culture and language on test performance, particularly in tasks with significant linguistic demands. Allalouf et al. (1999) highlight the role of translation and cultural congruence in item difficulty, attributing disparities in item difficulty and discrimination to translation inaccuracies and cultural relevance. Further research by Masri et al. (2016) and Noroozi and Karami (2022) illustrates how acknowledging the influence of language on test takers' perceptions can refine our understanding of item evaluation and difficulty estimation. Gao and Rogers (2010) point to the dynamic interaction between test takers and tasks as a pivotal factor in item difficulty, noting variability across language groups and proficiency levels.

This study distinguishes itself by concurrently examining item characteristics like "readability" and the attributes of both the items and the experts involved in difficulty estimation, thereby contributing a novel perspective to the taxonomy of item difficulty in language testing. In the present study, expert features and item features were also examined together using a multi-faceted Rasch analysis. Since the needs of each expert differ, the effect of feedback on the feedback of individual experts was analyzed. In this case, the present study focused on estimated item difficulty based on the data and perceived item difficulty based on the experts' prediction. It investigated the features that affect estimated and expert item difficulty perceptions and, based on the results, gave feedback to the experts. Therefore, the study aims to provide feedback to experts to improve their item difficulty predictions. This study aims to

i. identify variables that experts use to predict item difficulty,

ii. provide feedback to experts to improve their item difficulty predictions.

The study will contribute to understanding the item difficulty of a high-stakes language test that includes reading comprehension, grammar, and reasoning in the domain. Additionally, the study provided feedback to teachers, professors, and test developers- all item writers-. Accurate item difficulty estimation is crucial for developing valid and reliable assessments that align with learning objectives and provide meaningful feedback to experts and policymakers. The feedback from the data is also expected to guide the item-writing process.

2. METHOD

2.1. Research Design

The research was conducted in a semi-experimental design with the current objective of providing feedback to experts to improve their predictions of item difficulty. Experimental research entails studies to test the impact of variations the researcher creates on the dependent variable. The fundamental aim of experimental designs is to examine the cause-and-effect relationships established among variables. In experimental research, causality between variables is investigated, and changes are observed while controlling variables. Experimental studies seek to elucidate relationships between variables, interpret these relationships, and how outcomes may be influenced based on independent variables (Fraenkel & Wallen, 1990). The study received ethical clearance from the Ethics Committee of Gazi University, bearing the reference number 77082166-604.01.02-711551, dated 02.08.2023.

2.2. Participants

The first stage of the study on item difficulties was estimated based on 20,000 students who attended LGS and took the A booklet. In the second stage, 32 experts predicted the item difficulty, and in the third stage, 24 experts who had at least 3 correct predictions were selected and were given individual feedback to them. The same 24 experts predicted item difficulty again in the 4th stage of the study. Table 1 shows some information about the participants of the research.

Table 1. *Participants.*

Stage1			Stage2			Stage 3			Stage 4		
Estimated item difficulty			First prediction			Feedback			Second prediction		
Characteristic	*f*	%	Characteristic	*f*	%	Characteristic	*f*	%	Characteristic	*f*	%
Students			Experts			Experts			Experts		
Test year			*Gender*			*Gender*			*Gender*		
2018	10,000	50.0	Female	18	56.3	Female	13	54.2	Female	13	54.2
2019	10,000	50.0	Male	14	43.8	Male	11	45.8	Male	11	45.8
Gender			*Years of experience*			*Years of experience*			*Years of experience*		
Female	9,913	49.6	<1 year	3	9.4	<1 year	3	12.5	< 1 year	3	12.5
Male	10,087	50.4	1-5 years	8	25.0	1-5 year	5	20.8	1-5 year	5	20.8
School Type			5-10 years	13	40.6	5-10 year	11	45.8	5-10 year	11	45.8
Public	18,366	91.8	10+ years	8	25.0	10+years	5	20.8	10+years	5	20.8
Private	1,634	8.2	*Profession*			*Profession*			*Profession*		
			Professor	13	40.6	Professor	8	33.3	Professor	8	33.3
			Teacher	10	31.3	Teacher	8	33.3	Teacher	8	33.3
			Test developer	9	28.1	Test developer	8	33.3	Test developer	8	33.3
Total	20,000	100	Total	32	100	Total	24	100	Total	24	100

2.3. Process

This study was carried out in four stages as an experimental design. In the first stage, item difficulties were estimated for 40 items based on the data in the High School Entrance Examination (known as LGS) in Türkiye, and item features that affect item difficulty were determined. In the second stage, 6 items were determined from the 40 items with different item features. Items were selected based on different content domains (reading comprehension, reasoning and grammar), some of which include visual and some non-visual content. While some items are very long, some are short; some are easy, and some are moderate or hard. In this stage, 32 experts predicted the difficulty of the same 6 items. The factors that affect experts' item difficulty predictions were studied and the experts' predictions were compared with the actual item difficulty in the second stage. In the third stage, experts who had at least 3 correct predictions were determined and gave individual feedback to experts based on the results. In the fourth stage, 24 experts predicted 34 items' difficulty on a 5-point scale in a nested way. It means that in this stage, experts predicted the item difficulty of 6 items and did not see all items. Each item was predicted by at least 3 experts. After that, the factors that affect experts' item difficulty predictions were identified, and the experts' perceptions were compared with the estimated item difficulty again.

2.4. Predictors

In the current study, certain variables that contribute to the estimation of item difficulty within the Turkish test were analyzed. This analysis encompassed several item characteristics, including item length (word count), readability, visual content, content domain, and question prompt for the item features. Additionally, attributes of the raters themselves were considered to explore factors influencing SMEs' predictions of item difficulty in the Turkish test. Specifically, the analysis took into account the gender of the raters, their years of experience in test development, and their professional backgrounds. The findings about these features are delineated in Table 2 and Table 3. The features of items of visual content, question prompt, and content domain were scrutinized based on the assessments of two experts with backgrounds in Turkish language education and item writing. These experts independently identified the attributes of the items, and their findings were subsequently synthesized for analysis. The textual properties of the items, including item length and readability, were calculated utilizing Python software. Determining item length involved computing the word count, while readability was assessed by implementing the Ateşman (1999) formula.

Table 2. *Item features to determine affecting estimated and perceived item difficulty.*

Item Features	Min	Max	*M*	*S*	*n*	%
Visual content						
Yes					7	17.5
No					33	82.5
Question prompt						
Positive phrased					31	77.5
Negative phrased					9	22.5
Content domain						
Reading comprehension					24	60.0
Grammar					10	25.0
Reasoning					6	15.0
Textual features						
Item length (word count)	24.0	416.0	113.5	77.8		
Readability	36.2	84.9	62.0	11.7		

Table 3. *Rater features to determine affecting perceived item difficulty.*

Rater Features	1st prediction		2nd prediction	
	n	%	*n*	%
Gender				
Female	18	56.3	13	54.2
Male	14	43.8	11	45.8
Years of experience				
<1 year	3	9.4	3	12.5
1-5 years	8	25.0	5	20.8
5-10 years	13	40.6	11	45.8
10+ years	8	25.0	5	20.8
Profession				
Professor	13	40.6	8	33.3
Teacher	10	31.3	8	33.3
Test developer	9	28.1	8	33.3

2.5. Feedback Process

In the second stage, 24 experts provided feedback on the difficulty of the items. For this, an instructor group was established. It consisted of three professors, two of them working in the field of Turkish education and one of them working in measurement and evaluation at the university. While preparing the feedback, the factors affecting the difficulty of the 6 items in the first stage were determined in detail by the instructor group. Based on the first stage results, they examined the purpose of the items, the formal and content features, and the order of the options together. Then, the accuracy and inaccuracy of the experts' predictions in the first stage were deduced, and the tutorial group conducted online interviews with each expert individually. The feedback was presented personally by comparing the factors that the experts paid attention to during the prediction process with the actual item statistics.

2.6. Data Collection Tool

The data collection process comprised four sequential stages within an experimental design framework. In the initial phase, item difficulty estimates were derived for 40 items based on the High School Entrance Examination data in Türkiye, with concurrent identification of item features influencing difficulty levels. The annual exam by the Ministry of National Education serves as a pivotal placement test for approximately 1 million students seeking admission to high schools. Subsequently, six items were selected from the initial pool, each characterized by distinct features such as content domain (e.g., reading comprehension, reasoning, grammar), visual or non-visual elements, varying lengths, and differing difficulty levels. Data were collected from the experts using an item difficulty estimation form. The form included the items and the item difficulty that the expert could mark the answer next to each item. Expert predictions of item difficulty on a 5-point scale (1=very difficult to 5=very easy) were obtained for these six items in the second stage, involving 32 experts. The third stage involved providing individual feedback to experts who demonstrated at least three correct predictions. Finally, in the fourth stage, 24 experts, following a nested design, predicted the difficulty of 34 items on a 5-point scale, with each item assessed by at least three experts.

2.7. Analysis

In the first stage, based on the answers of 10,000 students who participated in LGS in 2018 and 2019 and received booklet A, item difficulty was estimated based on the CTT for 40 items. Then, hierarchical regression analysis was performed to determine the features affecting the difficulty of the items by using the item length (word count), readability, visual content, content domain, and question prompt as predictors. In the second stage, the features affecting the item difficulty predictions of 32 experts were analyzed by multi-faceted Rasch analysis. Multi-faceted Rasch analysis is a statistical method used to examine the influence of different factors, such as experts and items, on expert predictions. This analysis provides individual and group-level statistics on a single comparable scale, the logit scale. The logit scale allows for meaningful comparisons and interpretations of the estimates (Myford & Wolfe, 2003). This study performed analyses using the Minifac (Facets) Rasch software program. The analysis included 6 item facets (item difficulty, item visual, question prompt, content domain, item length and item readability), and 4 rater facets (experts, experts' gender, profession, and year of experience). In the third stage, during the feedback process, the points that the experts paid attention to while predicting the difficulty of the items were determined and compared with the estimation of the items. In the fourth and final stage, 24 experts predicted the item difficulty of the remaining 34 items in the tests. In line with the experts' prediction, the difficulty of the 34 items was analyzed. In the multifaceted Rasch analysis, a 10-facet crossed design was used as items (6) x expert (24) x gender (2) x profession (3) x years of experience (4) x item visual (2) x question prompt (2) x content domain (3) x item length (2) x readability (2). In the analysis after the second prediction, predictions were similarly made based on the 10-factor crossed design. The model in the second prediction is as follows: items (34) x expert (24) x gender (2) x profession (3) x years of experience (4) x item visual (2) x question prompt (2) x content domain (3) x item length (2) x readability (2). The Spearman correlation coefficient was estimated to examine the relationship between the estimated and perceived values.

3. FINDINGS

3.1. Estimated Item Difficulty

After the estimated item difficulties, the average difficulty of the Turkish items in 2018 was estimated as 0.63. The item difficulties varied between 0.23 and 0.91. In 2019, the item difficulties varied between 0.34 and 0.75; the average difficulty was 0.59. Hierarchical regression analysis was performed to determine the extent to which item features explained the item's difficulties, and the results are shown in Table 4. As a result of the analysis, it was determined that 27% was explained by only the content domain feature. It was found that there is positive and moderate relationship between reading comprehension items and item difficulty (β=0.519; p<0.01). It shows that reading comprehension items are easier than grammar and reasoning items. However, it was determined that the item length (word count), readability, visual content and question prompt do not have a direct effect on the item difficulty (p>0.01).

Table 4. *Results of the regression analysis.*

		Unstandardized coefficients		Standardized coefficients		
Model		B	Std. Error	Beta	*t*	*p*
1	(Constant)	.519	.031		16.679	.000
	Cognitive_domain	.150	.040	.519	3.746	.001
2	(Constant)	.513	.139		3.689	.001
	Cognitive_domain	.145	.046	.502	3.184	.003
	Length	2.458E-5	.000	.013	.073	.942
	Readability	.000	.002	.015	.095	.925
	Visual_content	-.023	.077	-.061	-.295	.770
	Question_prompt	-.002	.051	-.007	-.044	.965

a. Dependent Variable: Item difficulty

3.2. First Round of Item Difficulty Prediction

In the second stage of the study, 32 experts predicted the difficulty of 6 items. The item difficulty predictions of the experts were analyzed by multi-faceted Rasch analysis with the experts and the items' features. All facet vertical rules are shown in Appendix 1, and the measurement report is shown in Table 5.

When Appendix 1 was examined, the experts indicated that the most difficult item was the 6th, and the easiest item was the 1st. It is seen that the experts' predictions of the difficulty/ease of the items were significantly divided into two categories approximately (reliability=0.70; strata=2.35; χ^2=16.1; p<0.05). It is also seen that R27 is the most generous (predicting that the items are easier), while R18 and R26 are the most rigid (predicting that the items are more difficult) experts. However, it was determined that the item difficulty predictions did not differ significantly in terms of strictness/generosity (reliability=0.27; χ^2=41.2; p>0.05). It was also determined that the item difficulty predictions of the experts did not differ significantly according to their gender (reliability=0.00; χ^2=0.4; p>0.05), profession (reliability=0.00; χ^2=0.7; p>0.05), and years of experiment (reliability=0.34; χ^2=6.4; p>0.05). When the item difficulty predictions were analyzed according to the item features, the experts tended to predict items with visual text more difficult than those with nonvisual text (the discrimination reliability values are high (>0.70) for the discrimination ratio (separation=1.74) and the discrimination index (strata=2.65); χ^2=4.0; p<0.05). Experts' item difficulty predictions also varied according to the length (number of words) of the item, and experts predicted items with more than 150 words to be more difficult than items with less than 150 words (reliability=0.71; χ^2=3.4;

$p<0.05$). However, it was determined that the predictions did not show a significant difference according to the positive-negative question prompt (reliability=0.00; $\chi^2=0.0$; $p>0.05$), content domain (reading comprehension, grammar, reasoning) (reliability=0.00; $\chi^2=0.8$; $p>0.05$) and readability (reliability=0.00; $\chi^2=0.0$; $p>0.05$).

Table 5. *Measurement report of the first prediction.*

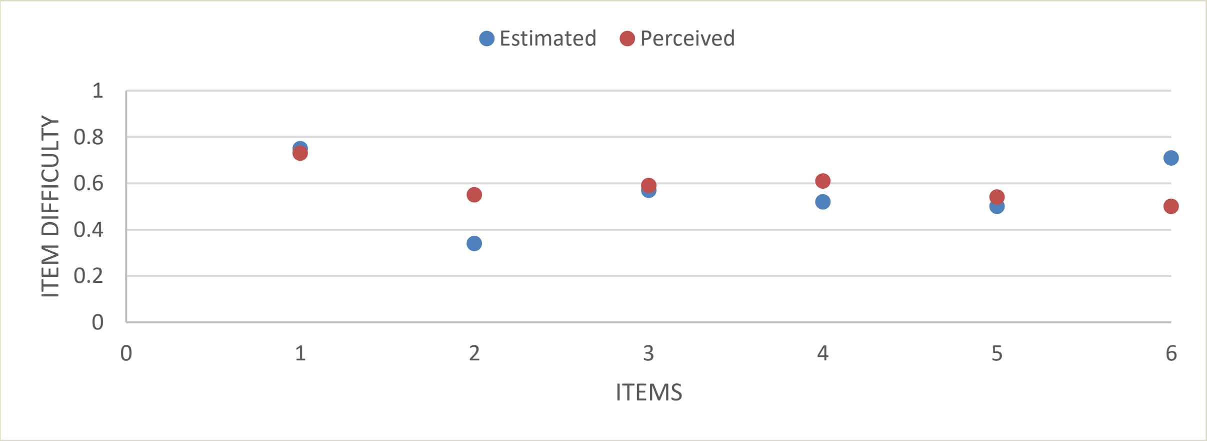
Model			Rater Features			Item Features				
Sample	Items*	Raters	Gender	Profession	Years of experience	Visual*	Question Prompt	Content Domain	Length*	Readability
RMSE	0.26	0.60	0.15	0.19	0.24	0.15	0.16	0.22	0.15	0.20
Adj (True) S.D.	0.39	0.37	0.00	0.00	0.18	0.26	0.00	0.00	0.23	0.00
Separation	1.51	0.61	0.00	0.00	0.72	1.74	0.00	0.00	1.56	0.00
Strata	2.35	1.15	0.33	0.33	1.30	2.65	0.33	0.33	2.41	0.33
Reliability	0.70	0.27	0.00	0.00	0.34	0.75	0.00	0.00	0.71	0.00
X^2	16.1	41.2	0.4	1.1	6.4	4.0	0.0	1.4	3.4	0.0
(sig.)	(0.01)	(0.10)	(0.51)	(0.58)	(0.09)	(0.04)	(0.93)	(0.50)	(0.04)	(0.93)

* Separated variables

3.3. Second Round of Item Difficulty Prediction

In the second round, 24 experts predicted the difficulty of the remaining 34 items. For this purpose, tests consisting of 6 items were prepared for the experts. For example, R4 predicted the difficulty of items 1, 5, 9, 12, 14, and 15, while R12 predicted the difficulty of items 4, 9, 23, 24, 26, and 27. In other words, a nested method was followed, not a cross method. So, each expert predicted the difficulty of 6 items, and at least 3 experts examined one item. It is shown in Figure 1.

Figure 1. *Compare the estimated and perceived item difficulty of the first prediction.*



The item difficulty predictions were analyzed using a multi-faceted Rasch analysis with the experts and item features. All facet vertical rules are shown in Appendix 2, and the measurement report is in Table 6. As a result of analyses, the raters indicated that the most difficult item was the 29th, and the easiest item was the 15th. When the item measurements are examined, the discrimination reliability values are high (>0.70) for the discrimination ratio (separation=1.88) and the discrimination index (strata=2.84). Accordingly, it is seen that the experts significantly categorized the difficulty/ease predictions of the items into approximately three categories ($\chi^2=189.3$; $p<0.05$). When the estimated values are also examined, the tests do not have very easy and very difficult items. Therefore, it can be said that the experts' item difficulty predictions are similar to the estimates. It is seen that R4 is the most generous (predicting that the items are easier), while R19 and R12 are the strictest (predicting that the items are more difficult) experts. It was determined that the experts' predictions differed significantly in terms of strictness/generosity (reliability=0.76; $\chi^2=41.2$; $p>0.05$). This is likely because the experts predicted 34 items using a nested method during the second prediction process. The second predictions did not differ significantly according to their gender (reliability=0.00; χ^2=0.8; *p*>0.05), profession (reliability=0.00; χ^2=0.6; *p*>0.05) and seniority (reliability=0.19; χ^2=4.1; *p*>0.05). When the predictions were analyzed according to the item features, it was determined that the item difficulty predictions varied according to the content domain (reliability=0.86; strata=3.57; χ^2=21.2; *p*<0.05). Accordingly, the experts stated that the most difficult items belonged to the grammar content domain, followed by the reasoning content domain. They stated that the reading comprehension items were easier than the items in the other content domain. Experts' item difficulty predictions also varied according to the length (number of words) of the item, with more than 150 words being more difficult than items with fewer than 150 words (reliability=0.84; strata=3.44; χ^2=6.4; *p*<0.05). Experts' predictions were also affected by the readability; as the readability of the items increased, experts tended to evaluate the items more difficult (reliability=0.85; strata=3.51; χ^2=12.0; *p*<0.05). However, it was determined that the item difficulty predictions did not show a significant difference according to the visual content (reliability=0.06; χ^2=1.1; *p*>0.05) and positive-negative question prompt (reliability=0.52; χ^2=2.1; *p*>0.05).

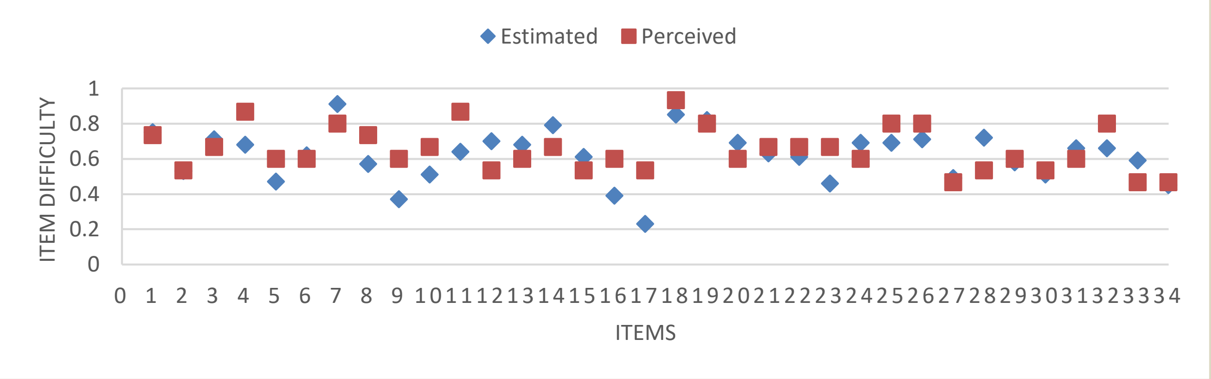
In the fourth stage of the study, after giving feedback to the experts, it was also found a positive and moderate correlation between the estimated and perceived item difficulty (*r*=0.410; *p*<0.01). It was observed that the experts tended to predict the items as easily as they were (Figure 4).

Table 6. *Measurement report of the second prediction.*

Model Sample	Items*	Raters*	Rater Features			Item Features				
			Gender	Profession	Years of experience	Visual	Question Prompt	Content Domain*	Length*	Readability*
RMSE	1.79	1.65	.41	.47	.62	.54	.48	.54	.44	.55
Adj (True) S.D.	3.37	2.97	.00	.00	.30	.13	.50	1.31	1.03	1.30
Separation	1.88	1.80	.00	.00	.48	.24	1.04	2.43	2.33	2.38
Strata	2.84	2.74	.33	.33	.98	.66	1.72	3.57	3.44	3.51
Reliability	0.78	0.76	.00	.00	.19	.06	.52	.86	.84	.85
X^2	189.3	98.7	.8	.6	4.1	1.1	2.1	21.2	6.4	12.0
(sig.)	(.00)	(.00)	(.36)	(.73)	(.25)	(.30)	(.15)	(.00)	(.01)	(.00)

* Separated variables

Figure 2. *Compare the estimated and perceived item difficulty of the second prediction.*



r=0.410; *p*<0.01

4. DISCUSSION and CONCLUSION

This study aimed to increase the accuracy of experts' item difficulty estimates by focusing on estimated item difficulty based on data and perceived item difficulty based on expert estimates. All results of the current study are summarized in Table 7.

Table 7. *Summary of the results.*

Turkish test	Estimated item difficulty	1st prediction	2nd prediction
LGS 2018	0.34 - 0.75	√	
LGS 2019	0.23 - 0.91		√
Rater features			
Rater (strictness/generosity)	---	x	√
Gender	---	x	x
Profession	---	x	x
Years of experience	---	x	x
Item features			
Visual content	x	√	x
Question prompt	x	x	x
Content domain	√	x	√
Item length	x	√	√
Readability	x	x	√

It was determined that 27% of the difficulty of the items in the Turkish test was significantly explained by only the content domain features. Although a limited number of studies have established models that explain a significant portion of the variation in item difficulty (53.5% (Sung et al., 2015), research showed that a significant variance in item difficulty is not explained by the models. For instance, despite identifying many explanatory predictors, they explained 23% of the variance in item difficulty in a science test (El Masri et al., 2017). The difficulty of 214 reading and listening comprehension items was modeled as a function of 12 predictor variables with item and text interaction. Seven of the 12 variables in the model explained approximately 31% of the variance in item difficulty (Rupp et al., 2001). In another study examining how task features affect item difficulty in EFL listening tests, regression analyses were conducted by using 20 predictors. As a result of the research, it was determined that item features explained 31.6% of the difficulty. (Ying-hui, 2006). The reason why a significant portion of the item difficulties were not explained may be that the difficulty varies according to the field, language, purpose, item types and other different structures of the test (Sydorenko, 2011).

The present study found that the reading comprehension items were easier than the grammar and reasoning items. The results also showed that the length of the items (word count), readability, visual or non-visual content, and positive or negative phrasing did not directly affect the item difficulty. Some research showed that longer items (i.e. length of distractors, item length) could be more difficult because they required more cognitive effort to process and comprehend (Fortus et al., 2013; Freedle & Kostin, 1993; Gorin & Embretson, 2006; Lin et al., 2021; Stenner, 2022; Stiller et al., 2016; Trace et al., 2017), and some studies also indicated that as the readability of items increases, their difficulty also increased (AlKhuzaey et al., 2021; Choi & Moon, 2020; Toyama, 2021). However, similar to the present research, some studies found that item length or readability might not always affect item difficulty directly (Aljehani et al., 2020). In this case, it is important to consider the specific context in which item length and readability are being considered. For example, in a language test where the primary goal is to assess reading comprehension skills, longer passages may be easier, even if they need more time to read, because they provide more information, and it might be easy to find the main idea or other indicators. The test also included grammar, reading comprehension and reasoning items in this research. Although grammar items were the shortest in the test, reading comprehension items were the easiest. For all these reasons, although experts thought length and readability are affected, the textual features (length and readability) examined in the study may not have effectively affected the item difficulty. In general, visual content can affect item difficulty by either aiding or hindering the test-taker's ability to comprehend the item. For example, if a test item includes a visual aid that effectively illustrates the content of the item, it may be easier for test-takers to understand the item and answer the item correctly. Conversely, if the visual aid is confusing or it is necessary to read the information in the visual and compare it with the information in the text and reach an inference, it may make the item more difficult for test-takers to understand and respond correctly (Santi et al., 2015; Stiller et al., 2016). In this study, it was determined that the visual content did not directly affect the difficulty of the item. The students had enough time to solve the items, the visual items were carefully designed in the item writing, the visual content was clearly expressed, the visuals were designed by the level of the students, and the students were familiar with the items in the visual content. Question prompting, another variable examined in this study, can also affect item difficulty. Research showed that negatively worded items can be more difficult than positive ones for test-takers to understand and answer correctly compared to positively worded items (Haladyna et al., 2002). However, some studies found that visual content or question prompts might not affect item difficulty (Caldwell & Pate, 2013). This study, conducted on a Turkish test, found that question prompts did not directly affect item difficulty. However, as with item length, it is important to consider other factors that may have influenced this finding. The findings that reading comprehension items were easier than grammar and reasoning items may indicate that students encounter greater challenges with grammar and reasoning items, which likely demand higher cognitive efforts. Reading comprehension items, relying on the ability to understand and interpret text, may enable students to locate answers more easily using information that is directly related to and retrievable from the text. In contrast, grammar and reasoning items might require more complex cognitive processes such as abstract thinking, knowledge of rules, and problem-solving skills. The result that the length of items (word count), readability, presence of visual or non-visual content, and the use of positive or negative phrasing did not directly impact item difficulty suggests the complexity of factors determining item difficulty, indicating that these features alone may not significantly influence the challenge level of an item. This implies that other variables, such as the cognitive abilities of the students being tested, their pre-existing knowledge, and their familiarity with the text or type of items, might be more determinative in influencing item difficulty. Although some research indicates that longer items might be more challenging due to the increased cognitive effort required to process and understand them, the findings of this study could suggest that students may have developed strategies to manage these lengths and remain unaffected in their question-solving process. Moreover, features like readability and visual content may not significantly affect item difficulty if they contain information that students are already familiar with or can easily understand.

In the first prediction, while the visual content in the items affected the experts’ prediction, it did not affect the second estimation. This is consistent with the real situation. While the content domain of the items did not affect the experts' predictions in the first prediction, it did in the second one. Experts stated that reading comprehension items were easier. This is exactly consistent with the estimated situation. The length of the items was effective in both predictions of the experts. The readability of the items was also effective in the experts' second prediction. The changes, which impact the experts' item difficulty predictions, are consistent with the estimates. In other words, there has been an improvement in the factors affecting the experts' predictions in line with the feedback given to the experts. A positive and moderate correlation was also found between the experts' perceptions and the estimated item difficulty (*r*=410; *p*<0.01). This finding is generally consistent with the results in the literature. For example, a study by Choi and Moon (2020) determined that the experts' prediction and estimated item difficulty were moderately or highly correlated in the reading comprehension items. Le Hebel et al. (2019) found that teachers could identify relevant potential sources of difficulty or easiness in the items that come from the PISA science test. Similarly, Attali et al. (2014) discovered that judges could accurately rank various items according to their difficulty level, and this trend remained consistent across multiple judges and subject areas in the SAT. Impara and Plake (1998) also stated that experts could predict item difficulty with 54% accuracy. Some research also showed that experts predict item difficulties significantly (Enright et al., 1993; Hamamoto Filho et al., 2020; Wauters et al., 2012), whereas some research showed the opposite of these results. For example, Sydorenko (2011) found a low correlation (*r* = .30) between the estimated and perceived difficulty, which could be due to the item writer not taking into account the difficulty of the topic and the similarity of intermediate and advanced items (Sydorenko, 2011). Kibble and Johnson (2011) stated that there is a significant but relatively low correlation between the perceived and estimated item difficulty in multiple-choice items (*r*=-0.19; *p*<0.01). Therefore, research suggests that experts should be aware of their potential biases and take steps to mitigate them, such as seeking feedback. In this research, it was found that there was an improvement in item difficulty prediction after giving feedback to the experts. It was consistent with research results that feedback or training on item difficulty improves experts' predictions (Davies, 2021; Fortus et al., 2013; González-Brenes et al., 2014; Hambleton & Jirka, 2011; Lumley et al., 2012; MacGregor et al., 2008).

In this study, it was also observed that the experts tended to predict the items as easily as they were. Urhahne and Wijnia (2021) reviewed 10 studies that examined the correlation between teachers' perceptions of task difficulty and the actual difficulty of those tasks with meta-analysis. The review found that in 8 out of the 10 studies, teachers tended to underestimate the level of challenge posed by the tasks or overestimate the expected performance of their students.

4.1. Limitation and Future Research

The study focuses on the High School Entrance Examination in Türkiye, which limits the generalizability of the findings to other contexts or examinations. Furthermore, 40 items can also be considered relatively small, potentially affecting the representativeness of the findings. In addition, the study primarily examines the factors that influence experts' item difficulty predictions and does not consider other potential sources of variability, such as test-taker characteristics. Based on the outcomes of this research, the practical implications for test developers, item writers, and educational practitioners are substantial and can significantly enhance the development and evaluation process of test items. The improvement in experts' predictions of item difficulty following feedback underscores the value of continuous training and development for item writers. Implementing feedback mechanisms and training programs that focus on the nuanced aspects of item design, such as the influence of visual content, content domain, item length, and readability on item difficulty, can empower item writers to make more accurate predictions. Similarly, the fluctuating impact of the content domain on expert predictions across different estimations highlights the importance of iterative review processes in accounting for various factors that may influence item difficulty. Furthermore, the findings suggest that training programs for item writers should cover the technical aspects of item construction and include modules on cognitive psychology and how test-takers interact with different item types. Such comprehensive training can enhance item writers' awareness of their potential biases and improve their ability to predict item difficulty accurately. In other words, the results may also serve as a source of guidance for item writers. It highlights the importance of validating expert judgments and using multiple sources of information when assessing item difficulty or other constructs in research.

Acknowledgments

This study was presented as an oral presentation at the NCME (National Council on Measurement in Education) conference held in Chicago, USA, from May 12-15, 2023.

The authors would like to thank the blind reviewers and SMEs for their useful comments and insightful suggestions. We also thank the Ministry of National Education (MEB) for providing the data necessary for this research.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Gazi University Ethics Committee, E77082166-604.01.02-711551, dated 02.08.2023.

Contribution of Authors

Ayfer Sayın: Design, Data Collection and/or Processing, Materials, Analysis and Interpretation, Literature Review, Writing. **Okan Bulut**: Conception, Design, Supervision, Writing, Critical Review.

Orcid

Ayfer Sayın https://orcid.org/0000-0003-1357-5674
Okan Bulut https://orcid.org/0000-0001-5853-1267

REFERENCES

Aljehani, D.K., Pullishery, F., Osman, O., & Abuzenada, B.M. (2020). Relationship of text length of multiple-choice questions on item psychometric properties–A retrospective study. *Saudi J Health Sci*, *9*, 84-87. https://doi.org/10.4103/sjhs.sjhs_76_20

AlKhuzaey, S., Grasso, F., Payne, T.R., & Tamma, V. (2021). A Systematic Review of Data-Driven Approaches to Item Difficulty Prediction. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, & V. Dimitrova, *Artificial Intelligence in Education* Cham. https://doi.org/10.1007/978-3-030-78292-4_3

Allalouf, A., Hambleton, R., & Sireci, S. (1999). Identifying the causes of dif in translated verbal items. *Journal of Educational Measurement, 36*(3), 185-198. https://doi.org/10.1111/j.1745-3984.1999.tb00553.x

Attali, Y., Saldivia, L., Jackson, C., Schuppan, F., & Wanamaker, W. (2014). Estimating item difficulty with comparative judgments. *ETS Research Report Series*, *2014*(2), 1-8. https://doi.org/10.1002/ets2.12042

Bejar, I.I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, *7*(3), 303-310. https://doi.org/10.1002/j.2333-8504.1981.tb01274.x

Benton, T. (2020). How Useful Is Comparative Judgement of Item Difficulty for Standard Maintaining? *Research Matters*, *29*, 27-35.

Berenbon, R., & McHugh, B. (2023). Do subject matter experts' judgments of multiple-choice format suitability predict item *quality*?. *Educational Measurement Issues and Practice, 42*(3), 13-21. https://doi.org/10.1111/emip.12570

Berk, R.A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, *56*(1), 137-172. https://doi.org/10.3102/00346543056001137

Bock, R.D., Murakl, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, *25*(4), 275-285. https://doi.org/www.jstor.org/stable/1434961

Boldt, R.F. (1998). GRE analytical reasoning item statistics prediction study. *ETS Research Report Series*, *1998*(2), i-23. https://doi.org/10.1002/j.2333-8504.1998.tb01786.x

Caldwell, D.J., & Pate, A.N. (2013). Effects of question formats on student and item performance. *American Journal of Pharmaceutical Education*, *77*(4). https://doi.org/10.5688/ajpe77471

Choi, I.-C., & Moon, Y. (2020). Predicting the difficulty of EFL tests based on corpus linguistic features and expert judgment. *Language Assessment Quarterly*, *17*(1), 18-42. https://doi.org/10.1080/15434303.2019.1674315

Dalum, J., Christidis, N., Myrberg, I.H., Karlgren, K., Leanderson, C., & Englund, G.S. (2022). Are we passing the acceptable? Standard setting of theoretical proficiency tests for foreign-trained dentists. *European Journal of Dental Education*. https://doi.org/10.1111/eje.12851

Davies, E. (2021). Predicting item difficulty in the assessment of Welsh. Collated Papers for the ALTE 7th International Conference, Madrid, Spain.

El Masri, Y.H., Ferrara, S., Foltz, P.W., & Baird, J.-A. (2017). Predicting item difficulty of science national curriculum tests: the case of key stage 2 assessments. *The Curriculum Journal*, *28*(1), 59-82. https://doi.org/10.1080/09585176.2016.1232201

Embretson, S., & Wetzel, C. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement, 11*(2), 175-193. https://doi.org/10.1177/014662168701100207

Enright, M.K., Allen, N., & Kim, M.I. (1993). A Complexity Analysis of Items from a Survey of Academic Achievement in the Life Sciences. *ETS Research Report Series*, *1993*(1), i-32. https://doi.org/10.1002/j.2333-8504.1993.tb01529.x

Fergadiotis, G., Swiderski, A., & Hula, W. (2018). Predicting confrontation naming item difficulty. *Aphasiology, 33*(6), 689-709. https://doi.org/10.1080/02687038.2018.1495310

Ferrara, S., Steedle, J.T., & Frantz, R.S. (2022). Response Demands of Reading Comprehension Test Items: A Review of Item Difficulty Modeling Studies. *Applied Measurement in Education*, *35*(3), 237-253. https://doi.org/10.1080/08957347.2022.2103135

Förster, N., & Kuhn, J.-T. (2021). Ice is hot and water is dry: Developing equivalent reading tests using rule-based item design. *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000691

Fortus, R., Coriat, R., & Fund, S. (2013). Prediction of item difficulty in the English Subtest of Israel's Inter-university psychometric entrance test. In *Validation in language assessment* (pp. 61-87). Routledge.

Fraenkel, J.R. & Wallen, dan Norman E. (2006). *How to Design and Evaluate Research in Education.* McGraw-Hill Education, USA.

Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading comprehension item difficulty for expository prose passages for three item types: Main idea, inference, and supporting idea items. *ETS Research Report Series*, *1993*(1), i-48. https://doi.org/10.1002/j.2333-8504.1993.tb01524.x

Gao, L., & Rogers, W. (2010). Use of tree-based regression in the analyses of l2 reading test items. *Language Testing, 28*(1), 77-104. https://doi.org/10.1177/0265532210364380

Giguère, G., Brouillette-Alarie, S., & Bourassa, C. (2022). A look at the difficulty and predictive validity of ls/cmi items with rasch modeling. *Criminal Justice and Behavior, 50*(1), 118-138. https://doi.org/10.1177/00938548221131956

González-Brenes, J., Huang, Y., & Brusilovsky, P. (2014). General features in knowledge tracing to model multiple subskills, temporal item response theory, and expert knowledge. The 7th international conference on educational data mining (pp. 84–91), London. https://doi.org/pdfs.semanticscholar.org/0002/fab1c9f0904105312031cdc18dce358863a6.pdf

Gorin, J.S., & Embretson, S. E. (2006). Item diffficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, *30*(5), 394-411. https://doi.org/10.1177/0146621606288554

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

Hamamoto Filho, P.T., Silva, E., Ribeiro, Z.M.T., Hafner, M.d.L.M.B., Cecilio-Fernandes, D., & Bicudo, A.M. (2020). Relationships between Bloom's taxonomy, judges' estimation of item difficulty and psychometric properties of items from a progress test: a prospective observational study. *Sao Paulo Medical Journal*, *138*, 33-39. https://doi.org/10.1590/1516-3180.2019.0459.R1.19112019

Hambleton, R.K., & Jirka, S.J. (2011). Anchor-based methods for judgmentally estimating item statistics. In *Handbook of test development* (pp. 413-434). Routledge.

Hambleton, R.K., Sireci, S.G., Swaminathan, H., Xing, D., & Rizavi, S. (2003). *Anchor-Based Methods for Judgmentally Estimating Item Difficulty Parameters*. LSAC Research Report Series, Newtown, PA.

Herzog, M., Sari, M., Olkun, S., & Fritz, A. (2021). Validation of a model of sustainable place value understanding in Turkey. *International Electronic Journal of Mathematics Education, 16*(3), em0659. https://doi.org/10.29333/iejme/11295

Hontangas, P., Ponsoda, V., Olea, J., & Wise, S.L. (2000). The choice of item difficulty in self-adapted testing. *European Journal of Psychological Assessment*, *16*(1), 3. https://doi.org/10.1027/1015-5759.16.1.3

Hsu, F.-Y., Lee, H.-M., Chang, T.-H., & Sung, Y.-T. (2018). Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. *Information Processing & Management*, *54*(6), 969-984. https://doi.org/10.1016/j.ipm.2018.06.007

Huang, Z., Liu, Q., Chen, E., Zhao, H., Gao, M., Wei, S., Su, Y., & Hu, G. (2017). Question Difficulty Prediction for READING Problems in Standard Tests. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1). https://doi.org/10.1609/aaai.v31i1.10740

Impara, J.C., & Plake, B.S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, *35*(1), 69-81. https://doi.org/10.1111/j.1745-3984.1998.tb00528.x

Kibble, J.D., & Johnson, T. (2011). Are faculty predictions or item taxonomies useful for estimating the outcome of multiple-choice examinations? *Advances in physiology education*, *35*(4), 396-401. https://doi.org/10.1152/advan.00062.2011

Kolen, M.J., & Brennan, R.L. (2004). *Test equating, scaling, and linking methods and practices*. Springer New York, NY. https://doi.org/10.1007/978-1-4939-0317-7

Le Hebel, F., Tiberghien, A., Montpied, P., & Fontanieu, V. (2019). Teacher prediction of student difficulties while solving a science inquiry task: example of PISA science items. *International Journal of Science Education*, *41*(11), 1517-1540. https://doi.org/10.1080/09500693.2019.1615150

Lin, C.-S., Lu, Y.-L., & Lien, C.-J. (2021). Association between Test Item's Length, Difficulty, and Students' Perceptions: Machine Learning in Schools' Term Examinations. *Universal Journal of Educational Research*, *9*(6), 1323-1332. https://doi.org/10.13189/ujer.2021.090622

Liu, X., & Read, J. (2021). Investigating the Skills Involved in Reading Test Tasks through Expert Judgement and Verbal Protocol Analysis: Convergence and Divergence between the Two Methods. *Language Assessment Quarterly*, *18*(4), 357-381. https://doi.org/10.1080/15434303.2021.1881964

Lumley, T., Routitsky, A., Mendelovits, J., & Ramalingam, D. (2012). *A framework for predicting item difficulty in reading tests* Proceedings of the annual meeting of the American educational research association (AERA), Vancouver, BC, Canada.

MacGregor, D., Kenyon, D., Christenson, J., & Louguit, M. (2008). Predicting item difficulty: A rubrics-based approach. *American Association of Applied Linguistics. March, Washington, DC*. https://doi.org/10.1109/FIE.2015.7344299

Masri, Y., Baird, J., & Graesser, A. (2016). Language effects in international testing: the case of pisa 2006 *science* items. *Assessment in Education Principles Policy and Practice, 23*(4), 427-455. https://doi.org/10.1080/0969594x.2016.1218323

Mislevy, R.J., Sheehan, K.M., & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement*, *30*(1), 55-78. https://doi.org/www.jstor.org/stable/1435164

Noroozi, S., & Karami, H. (2022). A scrutiny of the relationship between cognitive load and difficulty estimates of language test items. *Language Testing in Asia, 12*(1). https://doi.org/10.1186/s40468-022-00163-8

Oliveri, M., & Ercikan, K. (2011). Do different approaches to examining construct comparability in multilanguage assessments lead to similar conclusions?. *Applied Measurement in Education, 24*(4), 349-366. https://doi.org/10.1080/08957347.2011.607063

Rupp, A.A., Garcia, P., & Jamieson, J. (2001). Combining multiple regression and CART to understand difficulty in second language reading and listening comprehension test items. *International Journal of Testing*, *1*(3-4), 185-216. https://doi.org/10.1080/15305058.2001.9669470

Sano, M. (2015). Automated capturing of psycho-linguistic features in reading assessment text. Annual meeting of the National Council on Measurement in Education, , Chicago, IL, USA.

Santi, K.L., Kulesz, P.A., Khalaf, S., & Francis, D.J. (2015). Developmental changes in reading do not alter the development of visual processing skills: an application of explanatory item response models in grades K-2. *Frontiers in Psychology*, *6*, 116. https://doi.org/10.3389/fpsyg.2015.00116

Segall, D.O., Moreno, K.E., & Hetter, R.D. (1997). Item pool development and evaluation. In *Computerized adaptive testing: From inquiry to operation.* (pp. 117-130). American Psychological Association. https://doi.org/10.1037/10244-012

Septia, N.W., Indrawati, I., Juriana, J., & Rudini, R. (2022). An Analysis of Students' Difficulties in Reading Comprehension. *EEdJ: English Education Journal*, *2*(1), 11-22. https://doi.org/10.55047/romeo

Stenner, A.J. (2022). Measuring reading comprehension with the Lexile framework. In *Explanatory Models, Unit Standards, and Personalized Learning in Educational Measurement: Selected Papers by A. Jackson Stenner* (pp. 63-88). Springer. https://doi.org/10.1007/978-981-19-3747-7_6

Stiller, J., Hartmann, S., Mathesius, S., Straube, P., Tiemann, R., Nordmeier, V., Krüger, D., & Upmeier zu Belzen, A. (2016). Assessing scientific reasoning: A comprehensive evaluation of item features that affect item difficulty. *Assessment & Evaluation in Higher Education*, *41*(5), 721-732. https://doi.org/10.1080/02602938.2016.1164830

Sung, P.-J., Lin, S.-W., & Hung, P.-H. (2015). Factors Affecting Item Difficulty in English Listening Comprehension Tests. *Universal Journal of Educational Research*, *3*(7), 451-459. https://doi.org/10.13189/ujer.2015.030704

Swaminathan, H., Hambleton, R.K., Sireci, S.G., Xing, D., & Rizavi, S.M. (2003). Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates. *Applied Psychological Measurement*, *27*(1), 27-51. https://doi.org/10.1177/0146621602239475

Sydorenko, T. (2011). Item writer judgments of item difficulty versus actual item difficulty: A case study. *Language Assessment Quarterly*, *8*(1), 34-52. https://doi.org/10.1080/15434303.2010.536924

Toyama, Y. (2021). What Makes Reading Difficult? An Investigation of the Contributions of Passage, Task, and Reader Characteristics on Comprehension Performance. *Reading Research Quarterly*, *56*(4), 633-642. https://doi.org/10.1002/rrq.440

Trace, J., Brown, J.D., Janssen, G., & Kozhevnikova, L. (2017). Determining cloze item difficulty from item and passage characteristics across different learner backgrounds. *Language Testing*, *34*(2), 151-174. https://doi.org/10.1177/0265532215623581

Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educational Research Review*, *32*, 100374. https://doi.org/10.1016/j.edurev.2020.100374

Valencia, S.W., Wixson, K.K., Ackerman, T., & Sanders, E. (2017). Identifying text-task-reader interactions related to item and block difficulty in the national assessment for educational progress reading assessment. In: San Mateo, CA: National Center for Education Statistics.

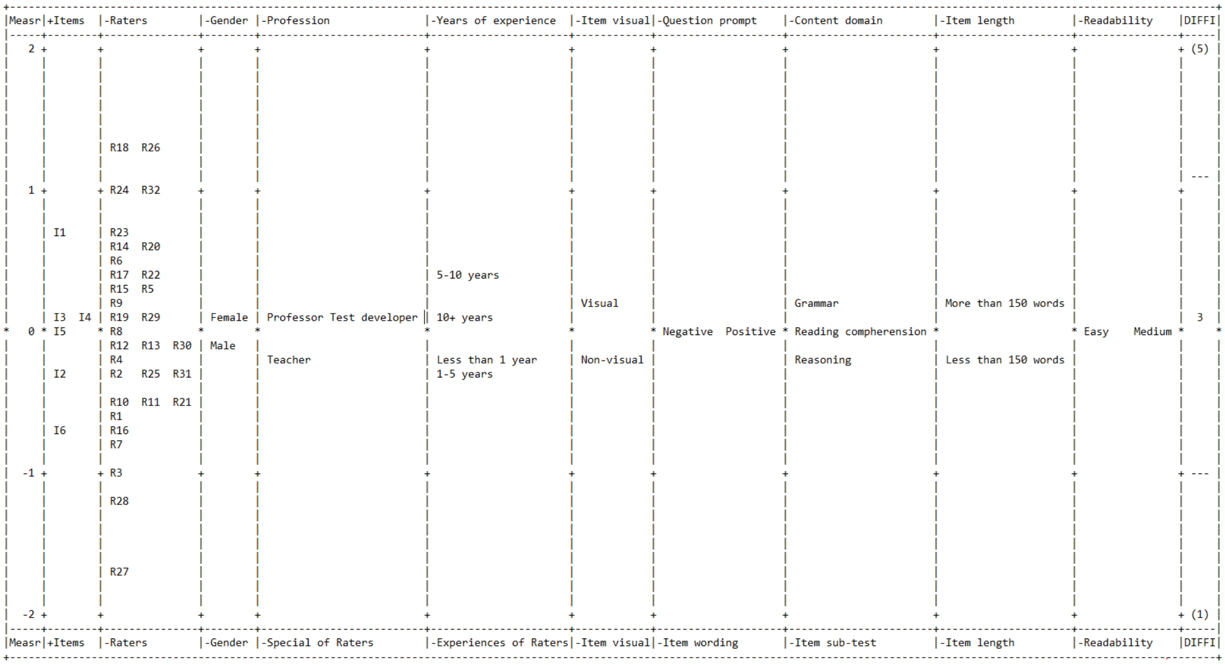
Van der Linden, W.J., & Pashley, P.J. (2009). Item selection and ability estimation in adaptive testing. In *Elements of adaptive testing* (pp. 3-30). Springer, New York, NY. https://doi.org/10.1007/978-0-387-85461-8_1

Wauters, K., Desmet, P., & Van Den Noortgate, W. (2012). Item difficulty estimation: An auspicious collaboration between data and judgment. *Computers & Education*, *58*(4), 1183-1193. https://doi.org/10.1016/j.compedu.2011.11.020

Ying-hui, H. (2006). An investigation into the task features affecting EFL listening comprehension test performance. *The Asian EFL Journal Quarterly*, *8*(2), 33-54.

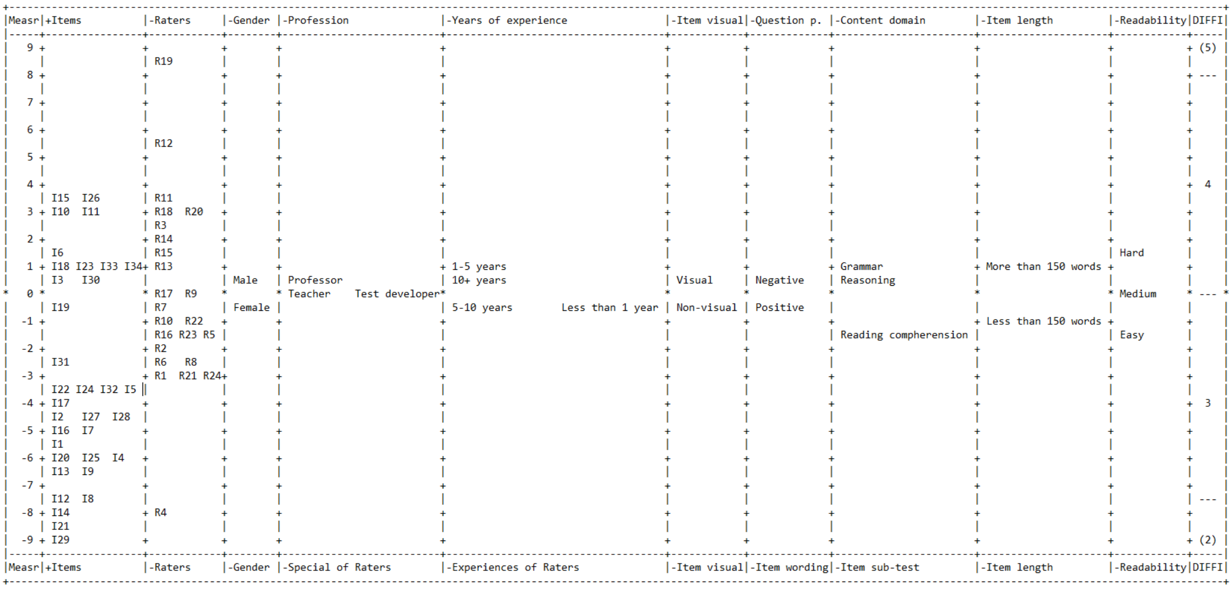
APPENDIX

Appendix 1. *All facet vertical "rulers" of the first prediction.*



```
+------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
|Measr|+Items  |-Raters        |-Gender |-Profession               |-Years of experience  |-Item visual|-Question prompt    |-Content domain        |-Item length         |-Readability    |DIFFI|
|-----+--------+---------------+--------+--------------------------+----------------------+------------+--------------------+-----------------------+---------------------+----------------+-----|
|   2 +        +               +        +                          +                      +            +                    +                       +                     +                + (5) |
|     |        |               |        |                          |                      |            |                    |                       |                     |                |     |
|     |        |               |        |                          |                      |            |                    |                       |                     |                |     |
|     |        |               |        |                          |                      |            |                    |                       |                     |                |     |
|     |        |               |        |                          |                      |            |                    |                       |                     |                |     |
|     |        |               |        |                          |                      |            |                    |                       |                     |                |     |
|     |        |               |        |                          |                      |            |                    |                       |                     |                |     |
|     |        | R18  R26      |        |                          |                      |            |                    |                       |                     |                |     |
|     |        |               |        |                          |                      |            |                    |                       |                     |                |     |
|     |        |               |        |                          |                      |            |                    |                       |                     |                | --- |
|   1 +        + R24  R32      +        +                          +                      +            +                    +                       +                     +                +     |
|     |        |               |        |                          |                      |            |                    |                       |                     |                |     |
|     |        |               |        |                          |                      |            |                    |                       |                     |                |     |
|     | I1     | R23           |        |                          |                      |            |                    |                       |                     |                |     |
|     |        | R14  R20      |        |                          |                      |            |                    |                       |                     |                |     |
|     |        | R6            |        |                          |                      |            |                    |                       |                     |                |     |
|     |        | R17  R22      |        |                          | 5-10 years           |            |                    |                       |                     |                |     |
|     |        | R15  R5       |        |                          |                      |            |                    |                       |                     |                |     |
|     |        | R9            |        |                          |                      | Visual     |                    | Grammar               | More than 150 words |                |     |
|     | I3  I4 | R19  R29      | Female | Professor Test developer|| 10+ years            |            |                    |                       |                     |                |  3  |
*   0 * I5     * R8            *        *                          *                      *            * Negative  Positive * Reading compherension *                     * Easy    Medium *     *
|     |        | R12  R13  R30 | Male   |                          |                      |            |                    |                       |                     |                |     |
|     |        | R4            |        | Teacher                  | Less than 1 year     | Non-visual |                    | Reasoning             | Less than 150 words |                |     |
|     | I2     | R2   R25  R31 |        |                          | 1-5 years            |            |                    |                       |                     |                |     |
|     |        |               |        |                          |                      |            |                    |                       |                     |                |     |
|     |        | R10  R11  R21 |        |                          |                      |            |                    |                       |                     |                |     |
|     |        | R1            |        |                          |                      |            |                    |                       |                     |                |     |
|     | I6     | R16           |        |                          |                      |            |                    |                       |                     |                |     |
|     |        | R7            |        |                          |                      |            |                    |                       |                     |                |     |
|     |        |               |        |                          |                      |            |                    |                       |                     |                |     |
|  -1 +        + R3            +        +                          +                      +            +                    +                       +                     +                + --- |
|     |        |               |        |                          |                      |            |                    |                       |                     |                |     |
|     |        | R28           |        |                          |                      |            |                    |                       |                     |                |     |
|     |        |               |        |                          |                      |            |                    |                       |                     |                |     |
|     |        |               |        |                          |                      |            |                    |                       |                     |                |     |
|     |        |               |        |                          |                      |            |                    |                       |                     |                |     |
|     |        |               |        |                          |                      |            |                    |                       |                     |                |     |
|     |        | R27           |        |                          |                      |            |                    |                       |                     |                |     |
|     |        |               |        |                          |                      |            |                    |                       |                     |                |     |
|     |        |               |        |                          |                      |            |                    |                       |                     |                |     |
|  -2 +        +               +        +                          +                      +            +                    +                       +                     +                + (1) |
|-----+--------+---------------+--------+--------------------------+----------------------+------------+--------------------+-----------------------+---------------------+----------------+-----|
|Measr|+Items  |-Raters        |-Gender |-Special of Raters        |-Experiences of Raters|-Item visual|-Item wording       |-Item sub-test         |-Item length         |-Readability    |DIFFI|
+------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
```

Appendix 2. *All facet vertical "rulers" of the second prediction.*



```
+-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
|Measr|+Items          |-Raters     |-Gender |-Profession               |-Years of experience               |-Item visual|-Question p. |-Content domain        |-Item length         |-Readability|DIFFI|
|-----+----------------+------------+--------+--------------------------+-----------------------------------+------------+-------------+-----------------------+---------------------+------------+-----|
|   9 +                +            +        +                          +                                   +            +             +                       +                     +            + (5) |
|     |                | R19        |        |                          |                                   |            |             |                       |                     |            |     |
|   8 +                +            +        +                          +                                   +            +             +                       +                     +            + --- |
|     |                |            |        |                          |                                   |            |             |                       |                     |            |     |
|   7 +                +            +        +                          +                                   +            +             +                       +                     +            +     |
|     |                |            |        |                          |                                   |            |             |                       |                     |            |     |
|   6 +                +            +        +                          +                                   +            +             +                       +                     +            +     |
|     |                | R12        |        |                          |                                   |            |             |                       |                     |            |     |
|   5 +                +            +        +                          +                                   +            +             +                       +                     +            +     |
|     |                |            |        |                          |                                   |            |             |                       |                     |            |     |
|   4 +                +            +        +                          +                                   +            +             +                       +                     +            +  4  |
|     | I15  I26       | R11        |        |                          |                                   |            |             |                       |                     |            |     |
|   3 + I10  I11       + R18  R20   +        +                          +                                   +            +             +                       +                     +            +     |
|     |                | R3         |        |                          |                                   |            |             |                       |                     |            |     |
|   2 +                + R14        +        +                          +                                   +            +             +                       +                     +            +     |
|     | I6             | R15        |        |                          |                                   |            |             |                       |                     | Hard       |     |
|   1 + I18 I23 I33 I34+ R13        +        +                          + 1-5 years                         +            +             + Grammar               + More than 150 words +            +     |
|     | I3   I30       |            | Male   | Professor                | 10+ years                         | Visual     | Negative    | Reasoning             |                     |            |     |
*   0 *                * R17  R9    *        * Teacher    Test developer*                                   *            *             *                       *                     * Medium     * --- *
|     | I19            | R7         | Female |                          | 5-10 years       Less than 1 year | Non-visual | Positive    |                       |                     |            |     |
|  -1 +                + R10  R22   +        +                          +                                   +            +             +                       + Less than 150 words +            +     |
|     |                | R16 R23 R5 |        |                          |                                   |            |             | Reading compherension |                     | Easy       |     |
|  -2 +                + R2         +        +                          +                                   +            +             +                       +                     +            +     |
|     | I31            | R6   R8    |        |                          |                                   |            |             |                       |                     |            |     |
|  -3 +                + R1  R21 R24+        +                          +                                   +            +             +                       +                     +            +     |
|     | I22 I24 I32 I5 ||           |        |                          |                                   |            |             |                       |                     |            |     |
|  -4 + I17            +            +        +                          +                                   +            +             +                       +                     +            +  3  |
|     | I2   I27  I28  |            |        |                          |                                   |            |             |                       |                     |            |     |
|  -5 + I16  I7        +            +        +                          +                                   +            +             +                       +                     +            +     |
|     | I1             |            |        |                          |                                   |            |             |                       |                     |            |     |
|  -6 + I20  I25  I4   +            +        +                          +                                   +            +             +                       +                     +            +     |
|     | I13  I9        |            |        |                          |                                   |            |             |                       |                     |            |     |
|  -7 +                +            +        +                          +                                   +            +             +                       +                     +            +     |
|     | I12  I8        |            |        |                          |                                   |            |             |                       |                     |            | --- |
|  -8 + I14            + R4         +        +                          +                                   +            +             +                       +                     +            +     |
|     | I21            |            |        |                          |                                   |            |             |                       |                     |            |     |
|  -9 + I29            +            +        +                          +                                   +            +             +                       +                     +            + (2) |
|-----+----------------+------------+--------+--------------------------+-----------------------------------+------------+-------------+-----------------------+---------------------+------------+-----|
|Measr|+Items          |-Raters     |-Gender |-Special of Raters        |-Experiences of Raters             |-Item visual|-Item wording|-Item sub-test         |-Item length         |-Readability|DIFFI|
+-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
```

International Journal of Assessment Tools in Education
2024, Vol. 11, No. 2, 388–405

https://doi.org/10.21449/ijate.1397644

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

Development and validation of the IS-C psychometric tool for evaluating children's impulsivity

Fatma Özgün Öztürk[1*], Ganime Can Gür[1]

[1]Pamukkale University, Faculty of Health Sciences, Department of Psychiatric Nursing, Denizli, Türkiye

ARTICLE HISTORY

Received: Nov. 29, 2023

Accepted: Apr. 15, 2024

Keywords:

Assessment,
Impulsivity,
Instrument Development,
Reliability,
Validity.

Abstract: This research aims to develop an instrument for the evaluation of impulsivity traits in children and to examine the psychometric features of the developed scale. The process of developing the scale involved three main phases: namely, item generation, evaluation of content validity, and analysis of psychometric properties. The study sample comprised 319 children (68 females, 201 males) aged 5-18, all diagnosed with attention deficit hyperactivity disorder (ADHD), including 50 who underwent pilot testing. Both exploratory and confirmatory factor analyses were employed to assess the factor structure of the scale, resulting in an 18-item scale encompassing motor impulsivity, non-planning impulsivity, and attention-related impulsivity factors. The Confirmatory Factor Analysis *(CFA)* indicated a satisfactory model-data fit. The overall scale demonstrated high reliability, with Cronbach's Alpha coefficients reaching 0.863. The analyses indicated that the scale is both valid and reliable.

1. INTRODUCTION

Impulsivity, which is accepted as a basic feature of childhood psychopathology, has been associated with various psychopathologies, especially attention deficit hyperactivity disorder (ADHD) (Beauchaine et al., 2017; Martel et al., 2017). ADHD, one of the most widespread disorders of childhood, is characterized by issues with hyperactivity, attention deficiency, and impulse control (Öztürk & Başgül, 2015). Patients with attention deficit and hyperactivity disorder may exhibit attention issues, hyperactivity, impulsive issues, or both symptoms simultaneously (Ercan & Aydın, 2005). The prevalence of attention deficit and hyperactivity disorder is between 2-17% in children, adolescents, and adults (Öztürk & Başgül, 2015). Beginning in childhood, ADHD symptoms can last until adolescence (60-80%) for a sizeable portion of patients, and even into adulthood (40-60%) for some patients (Ercan, 2015). In this context, ADHD, which is widespread in society, has several detrimental effects on a person's ability to be successful at school as well as their ability to interact with others and do business (Ercan & Aydın, 2005; Hallowell & Ratey, 2011; Yazgan, 2010). The impulsive/hyperactive subtype of ADHD substantially influenced these negative aspects. *Willcutt et al. (1999*) reported a relationship between impulsive/hyperactive subtype and oppositional defiant disorder or conduct disorder. Similarly, it has been noted that impulsivity and hyperactivity

*CONTACT: Fatma ÖZGÜN ÖZTÜRK ✉ ftmzgn@gmail.com Pamukkale University, Faculty of Health Science, Department of Psychiatric Nursing, Denizli, Türkiye

e-ISSN: 2148-7456

symptoms in teenagers are indicators of forensic criminal behavior, while attention deficit alone is not (Willcutt et al., 1999*)*.

In studies on impulsivity, it is emphasized that high levels of impulsivity may contribute to interpersonal and social difficulties and may also cause various mental health problems such as substance use disorders. In addition, it is also reported to be an important factor in juvenile delinquency and criminal behavior (Sharma et al., 2014).

Based on this information in the relevant literature, it can be said that impulsivity negatively affects an individual's quality of life, relationships, and functionality. Impulsivity arises from the interplay of various factors, including neurological, genetic, environmental, cognitive, social, and emotional influences. The complex interaction among these factors contributes to the manifestation of impulsivity (Gladwin et al., 2020; Han et al., 2022; Kreek et al.,2005; Nomura & Nomura, 2006; Sharma et al., 2014).

The risky act of impulsivity is characterized by the premature expression of thoughts, which frequently results in unfavorable outcomes and improper circumstances (L'Abate, 1993). Eysenck (1977) described impulsivity as the taking of risks, inability to prepare, and slow mental processing. In the literature, it is seen that impulsivity is classified in various ways by researchers (Dickman, 1990; Eysenck & Eysenck, 1977; Patton et al., 1995; Whiteside & Lynam, 2001). *Patton et al. (1995*) divided it into three categories; namely, acting without sufficient planning and thought, acting without sufficient motor activation, and attention issues (lack of a plan). Motor impulsivity is an area that represents impairments in the ability to inhibit impulsive action and inappropriate responses. Attentional impulsivity refers to a tendency to switch attention quickly and can lead to inappropriate snap judgments. Inability to plan impulsivity refers to the inability to think about a current orientation or the future (Patton et al.,1995). Impulsivity is a pattern of conduct rather than one impulsive act (Moeller et al., 2001). Impulsive persons have the potential to hurt not just themselves but also other people. As a result, impulsiveness is the fast and unplanned response to internal and external stimuli without considering any potential negative effects on oneself or others (L'Abate, 1993).

To diagnose, treat, and implement necessary interventions for any potential psychopathology, it is crucial to identify and address impulsivity. Various methods have been developed by mental health professionals worldwide to assess different dimensions of impulsivity in children. Typically, self-report surveys, parent, and teacher rating scales, as well as behavioral or computer-based tasks, are employed to identify impulsivity in children (Cyders & Coskunpınar, 2011; Olson et al., 1999). Measurement tools commonly used to assess impulsivity in children include the Barratt Impulsiveness Scale for Children, UPPS Impulsive Behavior Scale, Teacher-Rated Children's Attention and Impulse Control Questionnaire (TRCAICQ), Dickman Impulsivity Inventory for Children (IDIJ-c), ADHD-IV Rating Scale for measuring inattentive, impulsive, and hyperactive behaviors, Eysenck's Impulsiveness Questionnaire, Kansas Reflection-Impulsivity Scale for Preschool Children, and the Go/No-Go task (Barkley, 1991; Cosí et al., 2008; DuPaul et al., 1998; Eysenck et al., 1984; Halperin et al., 1991; Leyva & Nolivos, 2015; Patton et al., 1995; Watts et al., 2020; Wright, 1971).

This research contributes to the limited measurement tools available on impulsivity for children in Türkiye. This scale, developed for Turkish parents to evaluate their children's impulsivity levels, can provide a more in-depth understanding of child psychopathology and behavioral problems and thus can be used in early diagnosis and intervention processes for children's mental health. Additionally, the development of this scale in Turkish culture may enable its widespread use in clinical practices and research.

2. METHOD

2.1. Design

In this study, a methodological approach that included three basic stages was used in the development of the Children's Impulsivity Scale (CDS). In the first stage, an item pool was created for IS-C. Then, in the second stage, the content validity of the scale was meticulously evaluated. Finally, the third phase focused on improving and evaluating the psychometric properties of the IS-C. Through these systematic steps, the research aimed to ensure the comprehensiveness, appropriateness, and reliability of the scale.

2.2. Participants

Data from a private child psychiatry clinic was collected throughout the development and validation of the IS-C. Individuals who were willing to participate in the research were included in the research using the convenience sampling method. Participants in the current study had to meet the following criteria: being diagnosed with ADHD, being between the ages of 6 and 16, and not having any other psychiatric disease diagnosis. Data was collected from the parents of children who met these criteria. Different sampling groups were utilized at various stages of the scale's development. In this situation, groups for confirmatory and explanatory factor analyses (N=269) and the pilot scheme (N=50) were developed. To apply factor analysis, the sample must be five to ten times larger than the number of items in the scale (Bryman & Cramer, 2002). On the other hand, Kline (1994) states that a sample size of 200 people will usually be adequate, but this number can be reduced to 100 in cases when the factor structure is clear and sparse (Kline, 2015). When looking at the study groups in the research, the study groups can be said to be sizable enough for both validity and reliability analyses.

2.3. Instruments

2.3.1. Personal information form

It was formed by the researcher using information from the literature. The personal information form includes basic information about the children's age, education level, family type, and family income status, as well as basic information about their parents.

2.3.2. Turgay DSM-IV-based child and adolescent behavioral disorders screening and rating scale (T-DSM-IV-S)

The validity and reliability studies of the Turkish form of this scale, developed by Turgay (1995), were conducted by *Ercan et al. (2001).* The scale, which comprises 41 items, was created by translating the DSM-IV diagnostic criteria into questions without altering their original intent. The scale includes 9 questions that investigate attention deficit disorder, 6 questions that focus on hyperactivity, 3 questions that focus on impulsivity, 8 questions that focus on the oppositional defiant disorder (ODD), and 15 questions that focus on behavioral disorders. Mothers, fathers, and teachers of children who are thought to have ADHD fill out the scale. Each item is given a score between 0 and 3, where 0 is the lowest and 3 is the highest. At least 6 of the 9 items examining attention deficit must be answered with a score of 2 or 3, and at least 6 of the 9 questions examining hyperactivity and impulsivity must be answered with a score of 2 or 3.

2.4. Procedure

2.4.1. Formation of the item pool

In line with the theoretical knowledge and the relevant literature, an item pool was created by considering the definitions of basic impulsivity dimensions and clinical symptom findings (American Psychiatric Association, 2013; Ercan, 2015; Hallowell & Ratey, 2011; Mukaddes, 2015). While creating the item pool, more than one item should be written about the same symptom, the items should cover all aspects of impulsivity, a single symptom should be measured with one item, there should be positive and negative items related to impulsivity, the items should be concise, each item should have a main idea, and possible attention should be paid to features such as items being written in clear, understandable and simple language. For each of the three dimensions (motor, non-planning, and attention-related impulsivity) that were determined to be included in this newly developed scale, different questions were prepared by the behavioral aspects of these dimensions. Consequently, a 32-item item pool was created and a 4-level Likert scale was used. Participants rated their level of agreement with each statement from rarely/never (1) to always (4).

2.4.2. Content validity

Following the creation of the item pool, a group of six experts in the field and the language was formed to provide feedback on whether the items in the item pool accurately reflect the relevant conceptual framework and whether the expressions are appropriate in terms of linguistic, semantic, and spelling. To test the content validity, an expert opinion form was given to the experts and they were asked to give answers to this Likert-type scale as follows: 1. Not relevant, 2. Relevant but requires a significant change, 3. Relevant but requires little change, and 4. Very relevant. The items constituting the item pool were examined by field experts as to whether they reflected the relevant theoretical structure and their opinions and suggestions were received by language experts as to whether they were linguistically, semantically, and orthographically appropriate. Necessary adjustments were made to the items in line with the opinions and suggestions. The content validity index *(I-CVI)* was determined by considering the scores given by the experts to options 3 and 4 for each question, and the scale-level content validity index *(S-CVI)* was calculated by averaging these values. This process was used to evaluate the overall validity of the scale (Polit & Beck, 2006). According to Lynn (1986), when there are six or more experts, the *I-CVI* should equal 0.83. Thus, six items having *I-CVI* values of less than 0.83 were taken from the scale. In the end, the scale's S-CVI was found to be 0.90. An *S-CVI* value of 0.90 and higher could be used to support the claim that content validity is suitable (Polit et al., 2007). Finally, the scale's 25 items were evaluated by a Turkish field expert to confirm its language validity.

2.4.3. Pilot study

The internal validity of the scale and the compatibility of each item with the scale were determined through a pilot application. Accordingly, the pilot application was conducted with a group of 50 individuals who shared characteristics with the sample used for the measurement. For each person, the amount of time it would take to complete the form after it was handed out was determined. The test's average completion time was calculated by dividing the time between the first and last finishers by the total number of test takers. The situation of those who finished the test too early or too late was not considered. The completion time of the test was determined as 5 minutes. Cronbach alpha values and item-total correlation values were examined in the pilot application. According to the analysis, the Cronbach alpha value for the pilot application is 0.787. At this point, it was determined that 7 items (4th, 7th, 9th, 11th, 17th, 20th, and 21st items) did not fit the scale total adequately and that the item-total correlation values were below the acceptable level (below 0.20), therefore these items were to be removed from the scale. Validity and reliability analyses were carried out on the scale's 18-item final form.

2.5. Psychometric Testing and Statistical Analysis

Statistical analysis of the data in the study was conducted using *LISREL* 8.8 and *SPSS* 23.0.

2.5.1. Construct validity

Factor analysis, which combines several statistical techniques to parse complex data using a correlation or covariance matrix, is the most widely used technique for evaluating the psychometric properties of scales (Brown, 2015). Therefore, exploratory factor analysis *(EFA)* and confirmatory factor analysis *(CFA*) were used to assess the construct validity of the scale.

EFA is a technique for determining the number and type of relationships that may exist between elements of a measurement instrument. Kaiser-Meyer-Olkin *(KMO)* and Bartlett Sphericity Tests were conducted to evaluate the suitability of the data set for *EFA* analysis. The fact that Bartlett's test is significant and the *KMO* value is both greater than 0.60 and close to one indicates that the data are suitable for factor analysis (Hayran, 2012; Seçer, 2015; Terwee et al., 2007). Following this, the principal component analysis technique and direct oblimin rotation with Kaiser normalization were used to clarify the factor structure. The most appropriate structure and number of elements were determined using eigenvalues of 1 and above (DeVellis, 2016; Johnson & Christensen, 2019). According to recommendations, the factor value of each item should be 0.30 or higher (Çam & Baysan-Arabacı, 2010; Grove et al., 2012; Tavşancıl, 2019). In this study, the minimum factor loading accepted in determining which item will be placed under which factor is 0.32.

The assumed structure of the scale, derived from the *EFA* test, underwent validation through both first and second-level confirmatory factor analyses. Commonly used fit index indicators were used to evaluate *CFA* model fit. According to the criteria proposed by Marcoulides and Schumacker (2001) and Seçer (2015), *RMSEA* and *SRMR* should be less than 0.08. Other fit index values should exceed 0.9. Additionally, the ratio of Chi-square to degrees of freedom (χ^2/df) should be less than 3.0.

2.5.2. Criterion-related validity

For the criterion-related validity of the scale, a correlation analysis was performed between IS-C and T-DSM-IV-S. The correlation between the IS-C and the T-DSM-IV-S was investigated using Spearman's Correlation Coefficient.

2.5.3. Reliability of the scale

Split-half reliability, internal consistency, and composite reliability analyses were used to assess the scale's reliability. Item-total score, floor and ceiling effects, and Cronbach's alpha coefficient were used to analyze internal consistency. A Cronbach's alpha value of 0.70 or higher was considered acceptable. Item-total correlations must be positive and higher than 0.25 (Kalaycı, 2010). To determine the satisfactory internal and content validity of an outcome instrument, it is advised that the percentage of ceiling and floor effect be less than 15% (Terwee et al., 2007). The two-half test reliability method is another method for calculating the scale's internal consistency coefficient. Spearman-Brown and Guttman split-half coefficients and the correlation between halves were calculated to determine split-half reliability. The minimum acceptable Spearman-Brown and Guttman split-half coefficients should be 0.70 (DeVellis, 2016; Johnson & Christensen, 2019). Hotelling's T2 test was used to determine whether the item averages were different from each other (Kartal & Bardakçı, 2018). The results of Tukey's Test for Non-additivity *(ANOVA and Tukey's Test for Non-additivity)*, which were carried out specifically to examine the additivity feature of the scale, were evaluated (Özdamar, 2016).

2.6. Ethical Considerations

Ethics committee approval was received dated 23/09/2020 and numbered 60116787-020/57785. Verbal and written information regarding the research, the "Informed Consent" principle, the "Respect for Autonomy" principle (indicating that the subjects were free to choose whether or not to participate in the study), and the "Confidentiality and Protection of Confidentiality" principle (assuring the subjects that their data would be kept private) were all provided to the parents and children.

3. RESULTS

3.1. Sample Characteristics

The study comprised 269 children in total. The average age of the children was 9.85±2.51, and 74% of them were boys. The moms' average age was 37.47±4.95, and 44.6% of them had completed high school. The fathers' average age was 40.91±4.77, and 46.5% of them had completed high school (Table 1).

Table 1. *Distribution of the Participants' Socio-Demographic Details (n:269).*

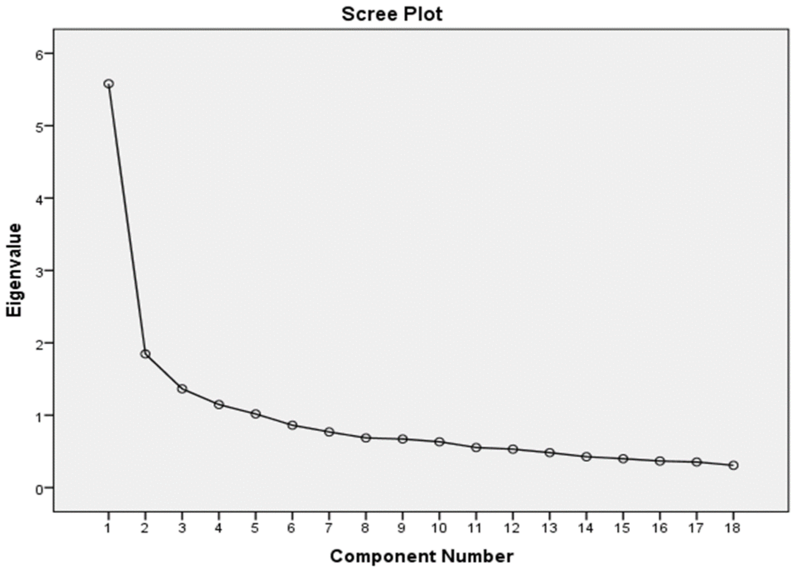
Variables	*n*	%
Child's Sex		
Female	68	25.3
Male	201	74.7
Mother's Education		
Elementary	50	18.6
High school	120	44.6
University	99	36.8
Father's Education		
Elementary	50	18.6
High school	125	46.5
University	94	35.0
	Mean±SD	Min.- Max.
Child's Age (year)	9.85±2.51	5-18
Mother's Age	37.47±4.95	26-59
Father's Age	40.91±4.77	30-58

3.2. Construct Validity

3.2.1. Exploratory factor analysis (EFA)

The *KMO* coefficient in the 18-item IS-C *EFA* was found to be 0.869, and the results of Bartlett's sphericity test (χ^2: 1511.495, *df*= 153, *p*<0.001) were significant. The Direct Oblimin method was chosen in the factor analysis to ensure that the structure of the relationship between the factors remained the same. Based on the Principal Component Analysis, it was discovered that 18 items were composed of three components (Figure 1) (scree plot).

Figure 1. *Scree plot graph.*



Following an Exploratory Factor Analysis, the first factor (seven items) was named "Motor impulsivity," the second (six items) "Non-planning impulsivity," and the third (five items) "attention-related impulsivity." This was determined by taking into consideration the conceptual structure and contents of the items. With factor loadings ranging from 0.446 to 0.792, the first factor accounted for 30.99% of the variance in total. 10.259% of the variance was explained by the factor loadings of the items in the second factor, which varied from 0.405 to 0.664. The third component's item factor loadings, which accounted for 7.582% of the variance overall, varied from 0.618 to 0.770. The total variance explained by the scale was found to be 48.840%. The eigenvalue for the first factor was determined as 5.580, 1.847 for the second, and 1.365 for the third (Table 2).

Table 2. *Explanatory factor analysis and item-total score analysis for the sub-scales.*

Sub-Scales	Explanatory Factor Analysis	Item-Subscale Total Score Analysis	
Items	Factor value of items	Item-subscale score Correlations (r)	*p*
Factor 1 Motor impulsivity)			
Q3	0.510	0.541	*p* <0.01
Q10	0.446	0.314	*p* <0.01
Q12	0.524	0.595	*p* <0.01
Q14	0.486	0.503	*p* <0.01
Q15	0.792	0.674	*p* <0.01
Q16	0.746	0.559	*p* <0.01
Q18	0.768	0.663	*p* <0.01
Eigenvalues	5.580		
Described Variance (%)	30.999		
Factor 2 (Non-planning impulsivity)			
Q5	0.664	0.471	*p* <0.01
Q6	0.660	0.483	*p* <0.01
Q8	0.622	0.333	*p* <0.01
Q13	0.596	0.539	*p* <0.01
Q24	0.450	0.451	*p* <0.01
Q25	0.405	0.325	*p* <0.01
Eigenvalues	1.847		
Described Variance (%)	10.259		
Factor 3 (Attention-related impulsivity)			
Q1	0.703	0.525	*p* <0.01
Q2	0.770	0.564	*p* <0.01
Q19	0.638	0.460	*p* <0.01
Q22	0.618	0.509	*p* <0.01
Q23	0.640	0.515	*p* <0.01
Eigenvalues	1.365		
Described Variance (%)	7.582		
Total explained variance (%)	48.840		

The correlation between the factors of the impulsivity scale was examined to determine the relationship between the factors. Table 3 shows the correlation values between the impulsivity scale's sub-dimensions. The findings indicate significant relationships between the scale's three sub-dimensions.

Table 3. *Inter-factor Correlation.*

Subscales	Factor 1	Factor 2	Factor 3
Factor 1	1		
Factor 2	0.540**	1	
Factor 3	0.503**	0.452**	1

**p<0.01 (2-tailed)

3.2.2. Confirmatory factor analysis (CFA)

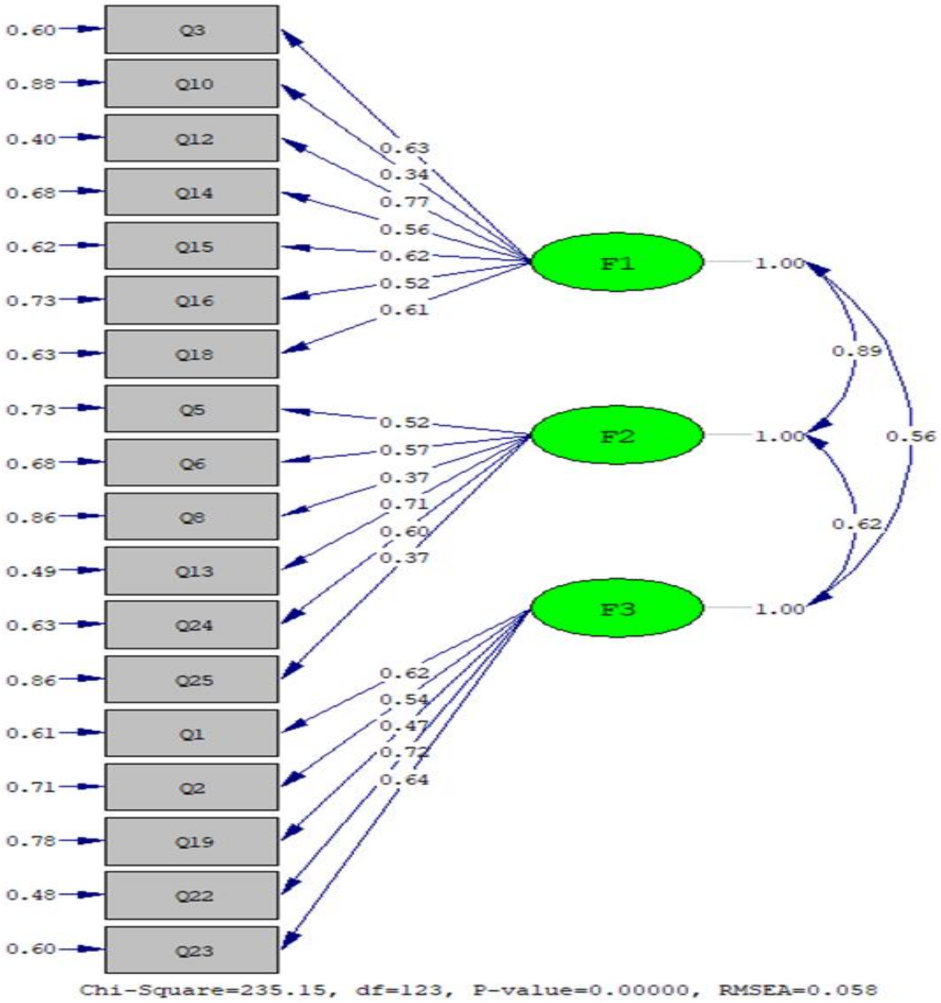
The IS-C, which has 18 items and three sub-factors, has fit indices that are significant according to the first level *CFA* results (χ^2= 235.15, *df*=123, *p*=0.000, χ^2/df=1.91) as shown in Figure 1. *RMSEA*: 0.05, *RMR*: 0.05, *SRMR*: 0.05, *CFI*: 0.96, *NNFI*: 0.95, *NFI2*: 0.93, *GFI:* 0.91, *AGFI*: 0.88, *IFI*: 0.96, and *RFI*: 0.91 are the values of the fit index (Table 4). All the fit indices for the structural model produced by the initial level *CFA* analysis were, therefore, at a good level. When the t-values between the factors and items were examined, it was seen that all the items were significant at the 0.05 level. Standardized correlation values were statistically significant (*p*<0.01); correlation values between motor impulsivity and non-planning impulsivity factors were 0.89 while the values were 0.56 between motor impulsivity and attention-related impulsivity factors and 0.62 between non-planning impulsivity and attention-related impulsivity factors (Figure 2). Standardized analysis values indicate how well each item (observable variable) represents its latent variable. When the diagram in Figure 1 is examined, one-way arrows pointing towards the observed variables from the latent variables motor impulsivity, non-planning impulsivity, and attention-related impulsivity show a linear significant relationship. This is an indicator of how well each variable represents the latent variable on which it is dependent (Şimşek, 2020). As shown in Figure 1, the standardized analysis values for each *CFA*-related item range from 0.34 to 0.77.

Table 4. *Results of the first and second level confirmatory factor analysis.*

Fit Indices Examined	Model		Result
	First-level CFA	Second-level CFA	
χ^2/df	1.91	2.07	Perfect Fit
RMSEA	0.05	0.06	Perfect Fit/ Acceptable Fit
RMR	0.05	0.05	Perfect Fit
SRMR	0.05	0.06	Perfect Fit/ Acceptable Fit
CFI	0.96	0.96	Perfect Fit
NNFI	0.95	0.95	Perfect Fit
NFI	0.93	0.92	Acceptable Fit
GFI	0.91	0.90	Perfect Fit
AGFI	0.88	0.87	Acceptable Fit
IFI	0.96	0.96	Perfect Fit
RFI	0.91	0.90	Acceptable Fit

RMSEA: Root Mean Square Error of Approximation; SRMR: Standardized Root-Mean-Square Residual; RMR: Root-Mean-Square Residual; FI: Comparative Fit index; NNFI: Non-Normed Fit Index; NFI: Normed Fit Index; GFI: Goodness of Fit Index; AGFI: Adjusted Goodness of Fit Index; IFI: Incremental Fit Index; RFI: Relative Fit Index

Figure 2. *Results of first-level confirmatory factor analysis.*



As demonstrated in Figure 2 (χ^2= 259.53, *df*=125, *p*=0.000, (χ^2/*df*=2.07), the second-level *CFA* results indicate that the fit indices of the IS-C are significant. *RMR*: 0.05, *RMSEA*: 0.06, *SRMR*: 0.06, *NNFI:* 0.95, *CFI:* 0.96, *NFI:* 0.92, *AGFI:* 0.87, *GFI:* 0.90, *RFI:* 0.90 and *IFI:* 0.96 were the values of the fit index (Table 4). Standardized correlation values were statistically significant (p<0.01); correlation values between scale and motor impulsivity factors were 0.86, while they were 0.97 between scale and non-planning impulsivity factors and 0.64 between scale and attention-related impulsivity factors. In the second level *CFA* analysis, modifications were implemented between Q2 and Q19, Q16 and Q18 items following the modification suggestions, and it was discovered that the model provided a better fit after the modifications. As shown in Figure 2, the standardized analysis values for each *CFA*-related item range from 0.36 to 0.75.

3.2.3. Item-total score analysis

EFA and *CFA* are widely acknowledged as the two most important analyses for ensuring construct validity during the scale development process. Even though item analysis is a reliability analysis, item-total correlations are calculated before *EFA* and *CFA* analyses to ensure item validity. According to the analysis of 18-item IS-C, the item correlation coefficients ranged between 0.294 and 0.643 (p<0.001) (Table 5).

Table 5. *Item-total score analysis.*

No	Items	Item-Scale Score Correlation (r)*	Cronbach's Alpha if Item Deleted
Q1	Able to regulate their behavior	0.453	0.857
Q2	When playing games and doing activities, she/he waits for her/his turn	0.420	0.858
Q3	She/he cannot wait	0.529	0.854
Q5	She/he cannot keep her/his word	0.415	0.858
Q6	She/he answers to the query without fully hearing or reading it	0.467	0.856
Q8	She/he is unaware of the risks.	0.294	0.864
Q10	She/he can tolerate situations when they arise that she does not want to	0.356	0.860
Q12	She/he wants to act in every way that comes to mind.	0.643	0.849
Q13	Does not wait for her/his turn when performing successive tasks	0.635	0.849
Q14	Is quick-paced	0.511	0.854
Q15	Till she achieves her/his goals, she/he persists even when she receives a negative answer.	0.538	0.853
Q16	She/he has angry outbursts that are excessive for the circumstance or incident that she/he is experiencing.	0.497	0.855
Q18	Promptly gets furious when any of his/her requests are rebuffed	0.599	0.851
Q19	Can maintain calm while sitting in places like theaters, movies, and classrooms	0.317	0.863
Q22	She/he is calm	0.514	0.854
Q23	She/he takes action while considering the outcome of her actions	0.453	0.857
Q24	Interrupts others as they are speaking	0.550	0.853
Q25	She/he cannot give up the tiny award at that moment, even if she/he will end up receiving a larger prize.	0.299	0.863

3.2.4. Criterion-related validity

Table 6 shows a moderate positive correlation (*r*= 0.524, 0.594, and 0.580, respectively) between the motor, non-planning, and attention-related impulsivity subscales of the IS-C and the hyperactivity and impulsivity subscale of the T-DSM-IV-S (*p*<0.01; *n*=155). According to the results, the criterion validity of the IS-C was established.

Table 6. *Criterion-related validity: Findings on the similar scale validity of the IS-C (n=155).*

Scale	IS-C		
	Motor impulsivity	Non-planning impulsivity	Attention-related impulsivity
T-DSM-IV-S (Hyperactivity and impulsivity subscale)	*r*	*r*	*r*
	0.524**	0.594**	0.580**

**p<0.01 (2-tailed); IS-C: Impulsivity scale for children; T-DSM-IV-S: DSM-IV-based child and adolescent behavior disorders screening and rating scale

3.3. Reliability of the Scale

The Cronbach Alpha reliability coefficients for "Factor 1," "Factor 2," "Factor 3," and the overall scale were determined to be 0.812, 0.702, 0.747, and 0.863, respectively (Table 7). The results of Table 3 indicate that the correlation coefficients between sub-scale item scores were statistically significant (*p*<0.001) and varied from 0.314 to 0.674 for "Factor 1," 0.325 to 0.539 for "Factor 2," and 0.460 to 0.564 for "Factor 3," respectively. The Spearman-Brown coefficients for the total scale were determined to be 0.857 by the split-half analysis, 0.827 for "Factor 1," 0.724 for "Factor 2," and 0.814 for "Factor 3." The results showed that the Guttman split-half coefficients for the overall scale, "Factor 1," "Factor 2," and "Factor 3" were 0.856, 0.820, 0.721, and 0.790, respectively. The correlation values for the two halves of the overall scale and subscale measures were found to be moderately and highly significant. The composite reliability coefficient, which was calculated using the error variance values, and the factor loadings that the *CFA* generated were 0.810 for factor 1, 0.741 for factor 2, 0.807 for factor 3, and 0.917 for the overall scale (Table 7).

The floor effect of the overall scale was 0.4, and its ceiling effect was 6.7. The floor and ceiling effects were as follows: 0.4 and 10.0 for "Factor 1," 0.7 and 13.4 for "Factor 2," and 0.7 and 12.6 for "Factor 3." According to Tukey's Test for Non-additivity, the items that make up the IS-C were found to be homogeneous and interrelated questions. Moreover, it showed that while the overall scale was not additive (Tukey Non-additivity: F= 9.532, p=0.002<0.05), the subscales of factor 1 (F=1.841, p=0.175>0.05), factor 2 (F=0.272, p=0.602>0.05), and factor 3 (F=0.056, p=0.812>0.05) were additive (Table 7). Hotelling's T-squared test was used to determine whether the test design was appropriate for ISC's reliability analysis applications, and the results showed that ISC's model had a suitable structure (F=21.390, p=0.000)

Table 7. *Reliability analysis of the total scale and sub-scales (n=269).*

Scale and Subscales	Cronbach-a	Spearman-Brown	Guttman split-half	Correlation between two halves	Composite Reliability	Floor effect %	Ceiling effect %	Tukey's Test for Non-additivity	Mean±SD	Min.-Max.
Factor 1	0.812	0.827	0.820	0.705	0.810	0.4	10.0	F=1.841 p=0.175	20.65±4.69	8-28
Factor 2	0.702	0.724	0.721	0.568	0.741	0.7	13.4	F=0.272 p=0.602	15.72±3.75	6-24
Factor 3	0.747	0.814	0.790	0.686	0.807	0.7	12.6	F=0.056 p=0.812	12.94±3.21	5-20
Scale	0.863	0.857	0.856	0.750	0.917	0.4	6.7	F=9.532 p=0.002	40.66-9.49	19-71

4. DISCUSSION and CONCLUSION

Numerous acts that are improper for the situation or that are overly dangerous, ill-thought-out, and frequently result in unfavorable outcomes are symptoms of impulsivity (Özdemir et al., 2012; Mukaddes, 2015). Thus, the purpose of this study was to develop a scale for gauging impulsivity in children. During the scale's development, a review of the literature was done, and the created item pool was presented to field experts, followed by pilot applications and item compatibility testing. The developed draft form was submitted to expert opinions on the scale's validity, and the Content Validity Index for each item on the scale was calculated. As stated in the literature, six items with values less than the determined value were removed from the test (Lynn, 1986). Furthermore, it was determined that the Content Validity Index value for the whole test is greater than the scope validity criterion, and the test's content validity is statistically significant (Polit et al., 2007).

A pilot application was given to 50 children who resembled the target demographic to reduce any issues that were likely to occur during the real application. Following the removal of seven items (4th, 7th, 9th, 11th, 17th, 20th, and 21st items) that were shown to have minimal test-related contributions, the item-total correlation analysis was conducted again. After the pilot application, a scale comprising 6 negative and 12 positive items was obtained. After that, it was decided whether the sample size was adequate and whether the variables had the appropriate degree of association by using the *KMO* and Bartlett sphericity tests. Correlation coefficients between partial and observed values were compared using the *KMO* test, an index. The ISC in the current study has a *KMO* value of 0.86, indicating that factor analysis may be performed on it. Furthermore, the p-value of the scale for Bartlett's test of sphericity was notably low (p<0.001), indicating that the correlation matrix of the scale's components is appropriate for factor analysis. In the following step, *EFA* was used to test the construct validity of the scale. None of the scale's items had overlapping features, and each item's factor loads exceeded 0.32. It was discovered that a three-dimensional structure explained 48.84% of the variation in total. Studies on scale development and adaptation should account for at least 40% of the variance according to Kline (2015). This means that the value determined by exploratory factor analysis during the research phase was adequate to determine the scale's factor structure.

The model fit of the factor structure obtained from *EFA* was examined using first- and second-level *CFA*, and the model fit indices were found to be at a good level. The *CFA* results revealed that the fit indices and factor loading values were within the ranges suggested by the literature (Marcoulides & Schumacker, 2001; Seçer, 2015). According to the relevant literature and theoretical views, the three-factor structure obtained after determining the model fit of the IS-C was named motor impulsivity, non-planning impulsivity, and attention-related impulsivity. It was determined that the standardized correlation values were statistically significant and that there were positive and significant relationships between the variables of motor, non-planning, and attention-related impulsivity. *CFA* results of the IS-C show that the scale confirms its three-factor structure and that the items adequately define and measure the concept they are intended to measure (DeVellis, 2016; Johnson & Christensen, 2019; Marcoulides & Schumacker, 2001) *EFA* and *CFA* results show that the three-dimensional factor structure of the scale is suitable for the Turkish sample and that the scale has a strong factor structure for the Turkish sample.

The criterion validity of the IS-C was examined by calculating the correlation coefficient between it and the T-DSM-IV-S hyperactivity and impulsivity subscale. In this study, a correlation coefficient between 0.70 and 0.30 was assumed to indicate a moderate correlation (Büyüköztürk, 2018). According to the findings, all subscales of IS-C were found to be moderately positively related to the hyperactivity and impulsivity subscale score of T-DSM-IV-S. It can be said that these results show that the IS-C has criterion validity. Additionally, the correlation values between the ISC subscales show that there are significant relationships between the three subscales of the scale and that there is no multicollinearity problem.

The reliability of the IS-C was assessed using split-half reliability, composite reliability, and internal consistency techniques. When the subscales and total score of the scale were examined, it was seen that it had composite reliability, split-half reliability, and internal consistency. For a scale to be considered reliable, it is typically expected to have a reliability rating of 0.70 or higher (Büyüköztürk, 2018; DeVellis, 2016; Johnson & Christensen, 2019). The internal consistency, split-half reliability, and composite reliability of the IS-C are supported by the data. In this study, the correlations between the items and the total score of the sub-dimension and the scale were both higher than 0.25 (Kalaycı, 2010). The total score correlations for item Q8 and item Q25 on the scale were 0.294 and 0.299, respectively. These items were retained in the scale because the factor loads for them ranged from 0.622 to 0.405. Because if the items in the scale have a tolerable item-total correlation (0.20-0.30 value), it is recommended not to rush to remove these items from the scale, but rather to look at the factor loading values during the factor analysis and decide accordingly (Seçer, 2015). This finding demonstrates that the items were related to both the scale and its sub-dimensions.

The results of Tukey's test for non-additive value are significant, which means that the scale's items have a structure that can account for at least three independent sub-dimensions and that the items are significantly different from one another. The probability of the total scale not being additive was determined as p<0.05, which shows that the overall scale is not additive.

When the sub-dimensions of the scale are examined, it is revealed that the probability of not being additive is p>0.05, that is, all sub-dimensions of the scale are additive (Özdamar, 2016). To determine if the item means varied from one another in this study, Hotelling's T2 test was performed (Kartal & Bardakçı, 2018). According to the results, there are differences between the means for scale items, item difficulty degrees are not all equal, participant responses to items are not all identical, and all scale items are significant. The scale's subscale is said to fall short of measuring the intended feature if the floor and ceiling percentages are higher than 15% (Terwee et al., 2007). The results of the present study demonstrated that the scale was a trustworthy measurement instrument and that the floor and ceiling effects were less than 15%.

Testing test-retest reliability in this study was not possible due to time constraints. The psychometric qualities of the scale are very strongly supported by the available data. To measure impulsivity in the context of this study, a validated and reliable instrument was developed. Furthermore, it can be applied to further research on this topic because there is no available scale like this scale in the literature.

In child and adolescent psychiatry, a scale that simply measures impulsivity and is completed by the family is not included in clinical practice in our nation. This study is the first in this field. Recognition of impulsivity, which underlies or coexists with many neurological and psychological diseases, is of great importance in terms of treatment, clinical follow-up, nursing care, and psychoeducation planning. This scale can be used to monitor pharmaceutical and cognitive-behavioral therapy in impulsivity. In the treatment strategy, the disease caused by impulsivity can be treated or impulsive behavior can be the focus of treatment. This newly created scale may help identify impulsivity and plan interventions on this issue.

4.1. Suitability for Clinical Application

We developed and validated the Children's Impulsivity Scale (IS-C) and identified the following three domains: non-planning impulsivity, motor impulsivity, and attention-related impulsivity. The impulsivity scale can be a valuable tool in understanding the effects of impulsivity on social functioning, academic performance, general attitudes, and behaviors in children. The effect of impulsivity on obesity, accident risks, behavioral problems, anger control difficulties, risky behaviors, fighting, peer bullying, screen addiction, substance addiction, etc. can be examined. In addition, the relationship of impulsivity with difficulties or problems in family processes can be investigated. The Turkish version of the scale and its evaluation are shown in the Appendix.

Acknowledgments

We are especially grateful to all patients who participated in this study. The authors of this study are grateful to all experts for translation and back-translation, for conceptual-content identification. This study was presented as a verbal presentation at Karadeniz 11th International Conference on Applied Sciences, December 17-18, 2022, in Rize, Türkiye. (Özgün Öztürk F, Can Gür G. Development of A Psychometric Tool for Assessing Impulsivity in the Children: A Validity Reliability Study, Verbal Presentation. Karadeniz 11th International Conference on Applied Sciences, December 17-18, 2022, Rize, 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). **Ethics Committee Number**: Pamukkale University, Non-Interventional Clinical Research Ethics Committee, 60116787-020/57785.

Contribution of Authors

Fatma Özgün Öztürk: Conceptualization, Investigation, Data curation, Writing – original draft, Writing – review & editing, Software. **Ganime Can Gür:** Conceptualization, Methodology, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Supervision. All authors have approved the final article.

Orcid

Fatma Özgün Öztürk https://orcid.org/0000-0001-5457-2694
Ganime Can Gür https://orcid.org/0000-0002-6013-257X

REFERENCES

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (5th ed., Vol. 10). American Psychiatric Association.

Barkley, R.A. (1991). The ecological validity of laboratory and analogue assessment methods of ADHD symptoms. *Journal of Abnormal Child Psychology*, *19*, 149-178. https://doi.org/10.1007/BF00909976

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

Bryman, A., & Cramer, D. (2002). *Quantitative data analysis with SPSS release 10 for Windows: A guide for social scientists*. Routledge.

Büyüköztürk, Ş. (2018). *Sosyal bilimler için veri analizi el kitabı* [*Handbook of data analysis for social sciences*]. Pegem Atıf İndeksi, 001–214.

Cosí, S., Morales-Vives, F., Canals, J., Lorenzo-Seva, U., & Vigil-Colet, A. (2008). Functional and dysfunctional impulsivity in childhood and adolescence. *Psychological Reports*, *103*(1), 67-76. https://doi.org/10.2466/pr0.103.1.67-76

Cyders, M.A., & Coskunpinar, A. (2011). Measurement of constructs using self-report and behavioral lab tasks: Is there overlap in nomothetic span and construct representation for impulsivity? *Clinical Psychology Review*, *31*(6), 965-982. https://doi.org/10.1016/j.cpr.2011.06.001

Çam, M., & Baysan-Arabacı, L. (2010). Qualitative and quantitative steps on attitude scale construction. *Hemar-G*, *12*(2), 59–71.

DeVellis, R.F. (2016). Scale development: Theory and applications. Sage publications.

Dickman, S.J. (1990). Functional and dysfunctional impulsivity: Personality and cognitive correlates. *Journal of Personality and Social Psychology*, *58*(1), 95. https://doi.org/10.1037/0022-3514.58.1.95

DuPaul, G.J., Anastopoulos, A.D., Power, T.J., Reid, R., Ikeda, M.J., & McGoey, K.E. (1998). Parent ratings of attention-deficit/hyperactivity disorder symptoms: Factor structure and normative data. *Journal of Psychopathology and Behavioral Assessment*, *20*, 83-102. https://doi.org/10.1023/A:1023087410712

Ercan, E.S. (2015). Dikkat eksikliği hiperaktivite bozukluğunda prognoz ve öngörücü faktörler [Prognosis and predictive factors in attention deficit hyperactivity disorder]. *Türkiye Klinikleri J Child Psychiatry-Special Topics*, *1*(1), 96–98.

Ercan, E.S., & Aydın, C. (2005). *Dikkat eksikliği hiperaktivite bozukluğu* [*Attention deficit hyperactivity disorder*]. Gendaş Kültür.

Ercan, E., Amado, S., Somer, O., & Çıkoğlu, S. (2001). Dikkat eksikliği hiperaktivite bozukluğu ve yıkıcı davranım bozuklukları için bir test bataryası geliştirme çabası [An effort to develop a test battery for attention deficit hyperactivity disorder and disruptive behavior disorders]. *Çocuk ve Gençlik Ruh Sağlığı Dergisi, 8(*3), 132–144.

Eysenck, S.B., & Eysenck, H.J. (1977). The place of impulsiveness in a dimensional system of personality description. *British Journal of Social and Clinical Psychology*, *16*(1), 57-68. https://doi.org/10.1111/j.2044-8260.1977.tb01003.x

Eysenck, S.B., Easting, G., & Pearson, P.R. (1984). Age norms for impulsiveness, venturesomeness, and empathy in children. *Personality and Individual Differences*, *5*(3), 315-321. https://doi.org/10.1016/0191-8869(84)90047-6

Gladwin, T., Jewiss, M., Banic, M., & Pereira, A. (2020). Associations between performance-based and self-reported prospective memory, impulsivity, and encoding support. *Acta Psychologica*, *206*, 103066. https://doi.org/10.1016/j.actpsy.2020.103066

Grove, S.K., Burns, N., & Gray, J. (2012). *The practice of nursing research: Appraisal, synthesis, and generation of evidence*. Elsevier Health Sciences.

Hallowell, E.M., & Ratey, J.J. (2011). *Driven to distraction* (revised): *Recognizing and coping with attention deficit disorder*. Anchor.

Halperin, J.M., Wolf, L., Greenblatt, E.R., & Young, G. (1991). Subtype analysis of commission errors on the continuous performance test in children. *Developmental Neuropsychology*, *7*(2), 207-217. https://doi.org/10.1080/87565649109540488

Han, X., Zhang, H., Xu, T., Liu, L., Cai, H., Liu, Z., …, & Yuan, T. (2022). How impulsiveness influences obesity: the mediating effect of resting-state brain activity in the dlpfc. *Frontiers in Psychiatry*, *13*. https://doi.org/10.3389/fpsyt.2022.873953

Hayran, O. (2012). *Sağlık bilimlerinde araştırma ve istatistik yöntemler* [*Research and statistical methods in health sciences*]. Nobel Tıp Kitabevi.

Johnson, R.B., & Christensen, L. (2019). *Educational research: Quantitative, qualitative, and mixed approaches*. SAGE Publications.

Kalaycı, Ş. (2010). *SPSS uygulamalı çok değişkenli istatistik teknikleri* (Vol. 5) [*Multivariate statistical techniques with SPSS*] (Vol. 5). Asil Yayın Dağıtım.

Kartal, M., & Bardakçı, S. (2018). *SPSS ve AMOS uygulamalı örneklerle güvenirlik ve geçerlik analizleri* [*Reliability and validity analysis with SPSS and AMOS applied examples*]. Akademisyen Yayınevi.

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

Kreek, M.J., Nielsen, D.A., Butelman, E.R., & LaForge, K.S. (2005). Genetic influences on impulsivity, risk taking, stress responsivity and vulnerability to drug abuse and addiction. *Nature Neuroscience*, *8*(11), 1450-1457. https://doi.org/10.1038/nn1583

L'Abate, L. (1993). *A family theory of impulsivity. In The impulsive client: Theory, research, and treatment*. (pp. 93–117). American Psychological Association.

Leyva, D., & Nolivos, V. (2015). Chilean family reminiscing about emotions and its relation to children's self-regulation skills. *Early Education and Development*, *26*(5-6), 770-791. https://doi.org/10.1080/10409289.2015.1005691

Lynn, M.R. (1986). Determination and quantification of content validity. Nursing Research.

Marcoulides, G.A., & Schumacker, R.E. (2001). New developments and techniques in structural equation modeling. Psychology Press.

Moeller, F.G., Barratt, E.S., Dougherty, D.M., Schmitz, J.M., & Swann, A.C. (2001). Psychiatric aspects of impulsivity. *American Journal of Psychiatry*, *158*(11), 1783–1793.

Mukaddes, N.M. (2015). *Yasam Boyu Dikkat Eksikligi Hiperaktivite Bozuklugu ve Eslik Eden Durumlar* [*Lifelong Attention Deficit Hyperactivity Disorder and Associated Conditions*]. Nobel Tıp Kitabevi.

Nomura, M., & Nomura, Y. (2006). Psychological, neuroimaging, and biochemical studies on functional association between impulsive behavior and the 5-ht2a receptor gene polymorphism in humans. *Annals of the New York Academy of Sciences*, *1086*(1), 134-143. https://doi.org/10.1196/annals.1377.004

Olson, S.L., Schilling, E.M., & Bates, J.E. (1999). Measurement of impulsivity: Construct coherence, longitudinal stability, and relationship with externalizing problems in middle childhood and adolescence. *Journal of Abnormal Child Psychology*, *27*, 151-165 https://doi.org/10.1023/A:1021915615677

Özdamar, K. (2016). *Ölçek ve test geliştirme yapısal eşitlik modellemesi* [*Scale and test development structural equation modeling*]. Nisan Kitabevi.

Özdemir, P.G., Selvi, Y., & Aydin, A. (2012). *Dürtüsellik ve tedavisi* [*Impulsivity and its treatment*]. *Psikiyatride Güncel Yaklaşımlar*, *4*(3), 293–314.

Öztürk, M., & Başgül, Ş. (2015). *Çocuklarda dürtüsellik* [*Impulsivity in children*]. Hayykitap.

Patton, J.H., Stanford, M.S., & Barratt, E.S. (1995). Factor structure of the Barratt impulsiveness scale. *Journal of Clinical Psychology, 51*(6), 768-774. https://doi.org/10.1002/1097-4679(199511)51:6<768::AID-JCLP2270510607>3.0.CO;2-1

Polit, D.F., & Beck, C.T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health, 29*(5), 489–497.

Polit, D.F., Beck, C.T., & Owen, S.V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health, 30*(4), 459–467.

Seçer, İ. (2015). *Psikolojik test geliştirme ve uyarlama süreci* (1. Baskı) [*Psychological test development and adaptation process*] (1. Baskı). Ankara: Anı Yayıncılık. ISBN, 978–605.

Sharma, L., Markon, K.E., & Clark, L.A. (2014). Toward a theory of distinct types of "impulsive" behaviors: a meta-analysis of self-report and behavioral measures. *Psychological Bulletin, 140*(2), 374. https://doi.org/10.1037/a0034418

Şenol, S. (2008). *Dikkat Eksikliği Hiperaktivite Bozukluğu. In Çocuk ve Ergen Psikiyatrisi Temel Kitabı* [*Attention Deficit Hyperactivity Disorder. In Basic Book of Child and Adolescent Psychiatry*] (pp. 293–311). Hekimler Yayın Birliği.

Şimşek, Ö.F. (2020). *Yapisal eşitlik modellemesine giriş: Temel ilkeler ve LISREL uygulamalari* [*Introduction to structural equation modeling: Basic principles and LISREL applications*]. Ekinoks Yayınları.

Tavşancıl, E. (2019). *Tutumların Ölçülmesi ve SPSS ile Veri Analizi* (6th ed.) [*Measurement of Attitudes and Data Analysis with SPSS*] (6th ed.). Nobel Akademik Yayıncılık.

Terwee, C.B., Bot, S.D., de Boer, M. R., van der Windt, D.A., Knol, D.L., Dekker, J., Bouter, L.M., & de Vet, H.C. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60*(1), 34–42. https://doi.org/10.1016/j.jclinepi.2006.03.012

Turgay, A. (1995). *Çocuk ve ergenlerde davranım bozuklukları için DSM-IV'e dayalı tarama ve değerlendirme ölçeği* (Yayınlanmamış ölçek) [*DSM-IV based screening and assessment scale for conduct disorders in children and adolescents* (Unpublished scale)]. Integrative Therapy Institute Toronto, Kanada.

Watts, A.L., Smith, G.T., Barch, D.M., & Sher, K.J. (2020). Factor structure, measurement and structural invariance, and external validity of an abbreviated youth version of the UPPS-P Impulsive Behavior Scale. *Psychological Assessment, 32*(4), 336. https://doi.org/10.1037/pas0000799

Whiteside, S.P., & Lynam, D.R. (2001). The five-factor model and impulsivity: Using a structural model of personality to understand impulsivity. *Personality and Individual Differences, 30*(4), 669-689. https://doi.org/10.1016/S0191-8869(00)00064-7

Willcutt, E.G., Pennington, B.F., Chhabildas, N.A., Friedman, M.C., & Alexander, J. (1999). Psychiatric comorbidity associated with DSM-IV ADHD in a nonreferred sample of twins. *Journal of the American Academy of Child & Adolescent Psychiatry, 38*(11), 1355–1362.

Wright, J.C. (1971). The Kansas Reflection-Impulsivity Scale for Preschoolers (KRISP). Merrifield, VA, USA: National Program on Early Childhood Education. https://doi.org/10.1037/t36374-000

Yazgan, Y. (2010). *Hiperaktif çocuk ve ergen okulda* [*Hyperactive children and adolescents at school*]. Evrim Yayınevi.

APPENDIX

6.1. The Evaluation of the Scores

The scale has three sub-dimensions, eighteen items, and a 4-point Likert style of design. On the scale, the answers to questions numbered Q1, Q2, Q10, Q19, Q22 and Q23 are scored reverse. In the IS-C, the scores that can be obtained from the "Motor Impulsivity" dimension can vary from 8 to 28, those that can be obtained from the "Non-planning Impulsivity" dimension from 6 to 24, and those that can be obtained from the "Attention-related Impulsivity" dimension from 5 to 20 (Table 7). The subscale scores served as the foundation for evaluating the ISC's results. The scale does not provide a total score. An elevated score on the scale denotes a heightened degree of impulsivity. The scale can be filled in by an adult (mother or father) who is familiar with the child.

6.2. Child Impulsivity Scale - Turkish Version

ÇOCUK DÜRTÜSELLİK ÖLÇEĞİ

AÇIKLAMA: Bu test bazı durumlarda çocuğunuzun nasıl düşündüğünü ve davrandığını ölçen bir testtir. Lütfen her cümleyi dikkatlice okuyunuz ve bu sayfanın sağındaki 4 seçenekten çocuğunuz için en uygun seçeneğe (X) işareti koyunuz. Her cümle için uzun süre düşünmeyiniz. Mümkün olduğu kadar çabuk ve samimi cevaplar veriniz. Kararsız kaldığınız durumlarda ilk aklınıza gelen doğrultuda hareket ediniz.

CÜMLELER:	Nadiren/ Hiçbir zaman	Bazen	Sıklıkla	Her zaman
1. Davranışlarını kontrol edebilir.	( )	( )	( )	( )
2. Oyun ve etkinliklerde sırasını bekler	( )	( )	( )	( )
3. Sabırsızdır	( )	( )	( )	( )
5. Verdiği sözleri tutamaz	( )	( )	( )	( )
6. Sorulan sorunun tamamını okumadan veya dinlemeden cevaplar	( )	( )	( )	( )
8. Tehlikeleri hesaplayamaz	( )	( )	( )	( )
10. İstemediği bir durum yaşadığında tahammül edebilir	( )	( )	( )	( )
12. Aklına ne gelirse yapmak ister	( )	( )	( )	( )
13. Sırayla yapılan işlerde sırasını bekleyemez	( )	( )	( )	( )
14. Tez canlıdır	( )	( )	( )	( )
15. İstediği bir şeyi elde edene kadar ısrar eder	( )	( )	( )	( )
16. İçinde bulunduğu durum ya da karşılaştığı olayla orantısız biçimde öfke patlaması yaşar	( )	( )	( )	( )

18. Herhangi bir isteği karşısında engellendiğinde hemen sinirlenir	( )	( )	( )	( )
19. Sınıfta veya sinema, tiyatro gibi ortamlarda sakince oturabilir.	( )	( )	( )	( )
22. Sakindir	( )	( )	( )	( )
23. Davranışlarının sonunu düşünerek hareket eder	( )	( )	( )	( )
24. Başkalarının sözünü keser	( )	( )	( )	( )
25. Daha sonra büyük bir ödül alacak olsa da o an küçük ödülden vazgeçemez	( )	( )	( )	( )

Referanslara eklemek koşulu ile ölçek izinsiz kullanılabilir.

International Journal of Assessment Tools in Education
2024, Vol. 11, No. 2, 406–423
https://doi.org/10.21449/ijate.1373629
journal homepage: https://dergipark.org.tr/en/pub/ijate
Research Article

Scoring open-ended items using the fuzzy topsis method and comparing it with traditional approaches

Aykut Çitçi[1], Fatih Kezer[2*]

[1]Republic of Türkiye Ministry of National Education, Ankara, Türkiye
[2]Kocaeli University, Faculty of Education, Department of Measurement and Evaluation, Kocaeli, Türkiye

ARTICLE HISTORY

Received: Oct. 09, 2023
Accepted: Mar. 04, 2024

Keywords:
Fuzzy logic,
Open-ended items,
Topsis,
Fuzzy topsis.

Abstract: This study investigates the application of the fuzzy logic method for scoring open-ended items, specifically comparing its effectiveness against traditional scoring methods. Utilizing the fuzzy TOPSIS method within the mathematics domain, this research established seven criteria for evaluating open-ended responses, developed in consultation with three experts. Due to constraints imposed by the pandemic, the study did not proceed with a real-world application; instead, it simulated data for 25 students to compare the rankings derived from traditional and fuzzy logic methods using the MS Excel program. The research produced three distinct rankings using the conventional method and analyzed the correlation between these rankings and those generated by the fuzzy TOPSIS method, employing the Spearman rank correlation coefficient. The findings reveal a significantly positive correlation between the rankings obtained through traditional methods and those acquired via the fuzzy logic approach, suggesting the latter's potential as an effective alternative for evaluating open-ended responses.

1. INTRODUCTION

The word “logic” in Turkish is the Arabic translation of the Greek word *logike*. It denotes both a verbal and mental concept. According to Al-Farabi, the word was derived from *nutk* (to say). Ali Sedad also indicated that *nutk* means both the utterance and the thought (Öner, 1986). As a concept, logic is a science that facilitates one to reach the knowledge of the unknown through the known or a discipline which prevents faulty thinking if one follows the rules. In other words, logic is a branch of science that examines correct and appropriate forms of thinking. The emergence of logic as a science is as old as the existence of mankind. Human beings need to think, reason, and make decisions all the time for different situations they face in their lives. There has been a need for a systematization of intellectual methods so that one can make the correct deduction and decisions (Karataş, 2018). Aristotle (384-322 BC) is the first thinker to examine and establish systematically (Öner, 1986; Paksoy et al., 2013).

According to Aristotle, logic is the science of the ideal laws of thinking (Aristotle, 1989, qtd. in Köz, 2022). Aristotle based his understanding of classical logic on the assumption that right and wrong as concepts are explicitly distinct; he argued that there could be more right or more

*CONTACT: Fatih Kezer ✉ fatih.kezer@kocaeli.edu.tr Kocaeli University, Faculty of Education, Department of Measurement and Evaluation, Kocaeli, Türkiye

e-ISSN: 2148-7456

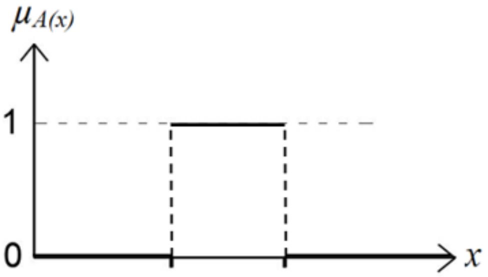
wrong situations; however, because he did not want to engage with fuzziness and thus made logic clear by defuzzying it. as such, Aristotle has established the foundation of classical logic (Erdin, 2007). Reasoning is important in classical logic, and it is its basis (Hasırcı, 2010: Öner, 1986; Taylan, 2008).

Criticism against the reasoning methods of classical logic has set the foundation of modern logic with the advent of symbols in the second half of the 19th century. Bertrand Russell's (1872-1970) contention that classical logic falls short in solving mathematical paradoxes along with his publication of *Principia Mathematica* with Whitehead in 1910 established symbolic (modern) logic (Paksoy et al., 2013). Just like in classical logic, modern logic aims to make inferences from the unknown towards to known. The use of symbolic language in modern logic studies aimed at alleviating the mistakes and shortcomings in language by turning premises and interferences into symbols. Modern logic has developed various inspection methods. These methods take us to the objectivity and univocity of symbolic language by purging the daily language of its polysemy (Eroğlu, 2012).

The critique against binary logic has brought forth the idea that situations between two extreme values should be taken into consideration. This critique also enabled the formation of fuzzy logic. Fuzzy logic as a concept was first coined by L.A. Zadeh in 1965 in his work titled *Fuzzy Sets.* The underlying philosophy of fuzzy logic is based on the assumption that a situation can have a continuous value between 'right' and 'wrong'.In other words, the value could be a reel number between 0 and 1 (Bostan, 2017). Fuzzy set theory emerged because Zadeh thought that the mathematical method of classical logic falls short in dealing with real-world problems (Avcı Öztürk, 2018; Elmas, 2003; Kaptanoğlu & Özok, 2006). The first application of fuzzy logic was in a steam engine designed by Mamdani in 1974. Zadeh introduced the theory of fuzzy logic to the world and Mamdani was the first person to put this theory into practice. Around the same time in Japan, practical application areas of fuzzy logic emerged (Özdağoğlu, 2016; Topçu, 2014).

In classical logic, the membership function for set A could be defined as follows:

Figure 1. *Function graph of membership in classical sets.*



$$\mu_{A(x)} = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$$

As can be seen in the function graph (Figure 1), µA membership function will assume values {0, 1} based on whether x members are in set A. In fuzzy logic, on the other hand, there are different membership functions such as triangular, trapezoidal, S and Z-shaped sigmoid, Cauchy, Gaussian, and monopulse (Baykal & Beyan, 2004; Cheng, 1996; Türe, 2006; Yen & Langari, 1999; Zimmermann, 2001). In practice, the most frequently used ones are the triangular, trapezoidal, curved, and Gaussian membership functions (Armağan, 2008). The core, support, and boundaries forming the membership function for a fuzzy set belonging to a universal set are shown in Figure 2 (Ross, 2010).

Figure 2. *Core, support and boundaries in a fuzzy set.*

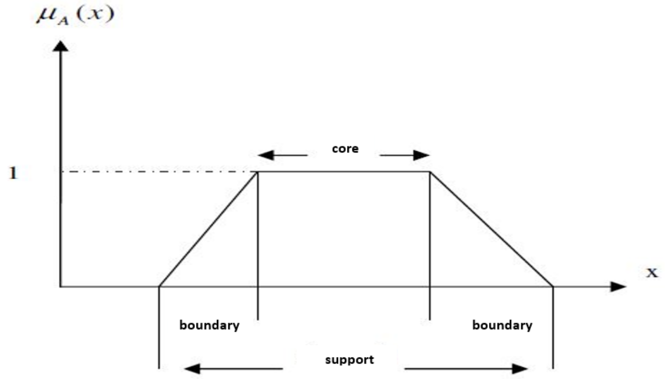


Figure 2 shows that in a fuzzy membership function, the core is a full member of set A in the universal set and contains elements the membership degrees of which are equal to 1. The support, on the other hand, is composed of elements whose membership degree is bigger than 0 in set A. In the fuzzy set A, boundaries indicate the area consisting of the elements, with degrees of membership different from zero, apart from full membership (Ross, 2010).

There are certain advantages and disadvantages of all systems used in decision-making depending on the area they are used. Among the advantages of fuzzy logic are the following: it requires fewer rules and decisions, assessments can be linguistically expressed, more observable variables can be assessed, the output can be related to the input, previously unsolved problems can be solved, quick prototyping is possible, it is more easily designed than traditional systems, it is cheaper, it can be used in the solution of complex problems, and it can be used in unstable and non-linear systems (Baykal & Beyan, 2014; Coşkunırmak, 2010; Elmas, 2003; McNeill & Thro, 1994; Özdağoğlu, 2016). Nevertheless, it has come with disadvantages in practice as well. Rules used in fuzzy logic are highly dependent on people's experience; variables of the membership function are specific to the application and are highly difficult to use in another application; while it is easy and fast to form a prototype it needs more simulation compared to the traditional control systems (Coşkunırmak, 2010; Elmas, 2003; McNeill & Thro, 1994; Özdağoğlu, 2016).

Multi-criteria decision-making methods, whether classical or fuzzy, can be used during decision-making processes. Since having too many criteria would complicate the decision-making process, multi-criteria decision-making methods ease the process and make it more objective (Cakar, 2020). Analytic Hierarchy Process (AHP), Vise Kriterijumska Optimizacija I Kompromisno Resenje (VIKOR), and Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) are some of the multi-criteria decision-making methods. These methods can be made fuzzy if necessary (Chen, 2000; Dündar, Ecer & Özdemir, 2010; Ertuğrul & Karakaşoğlu, 2008; Opricovic & Tzeng, 2004; Zimmermann, 1978).

Traditional approaches are used in making educational decisions, in line with classical logic. When determining student success, different types of tests (multiple choice, open-ended, true/false, matching, etc.) are utilized as a basis for educational decisions. Multiple-choice tests are one of the most frequently used methods to obtain valid and reliable results. While multiple-choice tests have certain advantages (being objective, having high content validity, easy scoring, easy application, etc.), they also have disadvantages when it comes to measuring students' advanced mental skills (such as problem-solving, creative thinking, critical thinking, and reasoning) (Bush, 2001; Klufa, 2018; McMillan, 2017; Miller, Linn & Gronlund, 2009; Popham, 1999; Tekin, 2010; Turgut & Baykul, 2012). To alleviate these disadvantages, open-ended items as well as in-class assessments are also utilised in assessing student success. Open-ended items are advantageous because they promote detailed learning, improve writing skills and alternative thinking, eliminate chance success, aim at improving advanced-level thinking skills, show the possibility of different correct answers as opposed to a single one, and enable students to structure their answers (Badger & Thomas, 1992; Cooney, Sanchez, Leatham & Mewborn, 2004; Geer, 1998; Karakaya, 2022; Öksüz & Güven Demir, 2019).

Ministry of Education (MEB) and Student Selection and Placement Centre (OSYM) (MEB, 2017; ÖSYM, 2017) have carried out trial applications using open-ended items. A total of 15 open-ended items in all fields were tested in an exam designed by OSYM. Answers were put on optic forms, and the scoring was done by a machine to ensure objectivity. In the first semester of the 2017-2018 school year, the Ministry of Education designed a TEOG (transition from primary to secondary education) exam with two open-ended items in Turkish, Mathematics, and Science. Items in this exam were open-ended and required long answers. Students were free to answer them as they liked; an answer sheet was used instead of an optic form, and the scoring was done by expert teachers. MEB prepared a structured answer key for the scoring of these items' answers; objectivity was ensured by asking the expert teachers to use this key when scoring the answers. Assessor-based objectivity has always been an issue when scoring the answers of especially open-ended questions, short-answer items, compositions, projects, and assignments. Using multiple assessors or the assessors scoring each item one by one are some of the methods used to alleviate this problem. Independent of the type of test, students' answers are scored in absolute numbers within the principles of classical logic. Scoring a student's answer to a multiple-choice question as 1 – 0 denotes certainty; scoring their composition 75 out of 100 also denotes certainty. In other words, these scores are certain, meaning they do not belong to a low or high-score group. This scoring takes place by employing the philosophy of classical logic systematised by Aristotle. In fuzzy logic, such concepts as certain and absolute are denoted by truth values, which are shown by membership degrees. These truth values are placed between completely true and completely false. One does not say that above a certain level is true or below a certain level is false. Using linguistic variables in assessment facilitates modeling operations (Elmas, 2011; Sarı, Murat, & Kırabalı, 2005).

Even though there are studies on the use of fuzzy logic in education (Hocalar, 2007; Kaptanoğlu & Özok, 2006; Bakanay, 2009), these studies are limited and none of them has tested fuzzy methods in scoring open-ended questions. It is believed that this present study will be one of the trials of using fuzzy logic methods in the field of assessment and evaluation. This study aimed to score open-ended items by Fuzzy TOPSIS, which is one of the multi-criteria decision-making methods. By doing so, a new method was tested; one in which students' answers did not have a certainty (0-1) and were scored based on different criteria weighted by experts, and one in which the experts scored the answers by linguistic expressions. To this end, scoring was done for the open-ended items developed for the mathematics classes, and students' gradation was compared to the classical method, TOPSIS, and fuzzy TOPSIS.

2. METHOD

In the study classical method, TOPSIS, and fuzzy TOPSIS methods were compared in the scoring of open-ended items. Carried out in the correlation research model, simulative data were used because the actual application was not possible as schools were closed due to COVID-19 restrictions. Moreover, since the study was a trial run to see whether fuzzy logic could be used in scoring open-ended items, only one item was used during the study so that the operations and the logic of gradation could be understood.

2.1. Simulative Design

Due to the global COVID-19 pandemic, schools were closed for face-to-face education and switched to online teaching in the 2020-2021 academic year. Simulated data were designed to exemplify a real-world application, as the study aimed to examine grading based on different methods. Number of students in a classroom may vary in different regions in Turkey. The MEB average for the 2019-2020 academic year was taken into account, and the data set was designed with 25 students (MEB, 2020, p. 24). Sub-criteria were devised to assess the open-ended mathematics item. Three field experts were consulted when devising the criteria and appointing significant weight to them; these experts also worked as scorers for the students' answers. Two of the experts in the study were maths teachers employed at MEB. The third expert was a maths teacher employed at the Evaluation and Assessment Centre at the Provincial Directorate of National Education. The teachers worked at the secondary education level and were included in the study by appropriate sampling (Altun, 2002; Damlar Demirci, 2019; Karadeniz, 2016; Van De Walle et al., 2014). A literature review was conducted when determining the criteria for the scoring of the open-ended mathematics items and different sub-criteria were determined. Then, these criteria were examined based on the separately gathered views of the experts and were reduced to seven, namely, (1) understanding the problem, (2) utilizing what is given in the problem, (3) using operations in the solution of the problem, (4) adapting the formula and the rules to the problem, (5) following the order of operations by making connections between operations, (6) making no mistakes in the operations, and (7) executing the operations clearly and in detail. The established criteria were emailed to the experts so that they could determine the sub-criteria of the scoring of the open-ended mathematics item. The experts assigned the values of "very low," "low," "somehow low," "medium," "somehow high," "high," and "very high," based on their personal views.

The student scores that would constitute the data of the study were randomly created between 1-7, keeping in mind the 7 criteria. During the fuzzification process, these scores were used as "very bad," "bad," "somehow bad," "medium," "somehow good," "good," and "very good" by converting them to linguistic variables. Students' scores and gradation were determined by fuzzy and classical methods by taking into account the scores obtained from students' answers to the items and the weight the experts have given to the sub-criteria.

2.2. Data Analysis

Students' scores for the mathematics item were first calculated according to the classical method. When doing this, the classical TOPSIS method was also used in addition to the classical scoring method. During the scoring, the weights of the sub-criteria were not used as the first method; instead, gradation was done by taking the average of the total scores given by the scorers. In the second method, on the other hand, TOPSIS was used as a multi-criteria decision-making method. The operational steps of the TOPSIS method were realized in the following order (Opricovic & Tzeng, 2004).

Step 1. A decision matrix is established by providing the criteria in the columns and alternatives in the lines.

According to each criterion in the study, scores given to the students are expressed as shown in Formula 1. The (C) in the columns symbolises the criteria, the (A) in the lines symbolise the alternatives, in other words, students, and the (W) symbolises the weight of the criterion.

$$D = \begin{array}{c|cccc|} & C_1 & C_1 & \cdots & C_n \\ \hline A_1 & x_{11} & x_{12} & \cdots & x_{1n} \\ A_2 & x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ A_m & x_{n1} & x_{n2} & \cdots & x_{mn} \end{array} \qquad w = [w_1, w_2, \ldots\ldots., w_n] \tag{1}$$

Step 2. A normalised decision matrix is established.

When forming the normalised decision matrix, the elements in the (D) decision matrix are used and the (r) matrix is formed by applying Formula 2. Each value in the decision matrix is divided by the square root of the sum of the squares of the x_{ij} values in the columns.

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{j=1}^{n} x_{ij}^{2}}}, \qquad i = 1, 2, ..., m \qquad j = 1, 2, ..., n \tag{2}$$

Step 3. A weighted normalised decision matrix is formed.

The weighted normalised decision matrix (V) is calculated by Formula 3. To carry out this operation, weight values of criteria (w) are first determined. Then, elements on each column of the R matrix are multiplied by the relevant criterion's weight value (w) thereby forming the (V) matrix.

$$v_{ij} = w_j . r_{ij} \qquad i = 1, 2, ..., m \qquad j = 1, 2, ..., n \tag{3}$$

Step 4. A positive ideal solution set and a negative ideal solution set are formed.

To establish ideal solution sets, a positive ideal solution set is formed by selecting the maximums of the weighted evaluation criteria in the (V) matrix, and a negative ideal solution set is formed by selecting the minimums. The minimum value is selected in the positive ideal solution set if the relevant criterion is minimization-oriented, and the maximum value in the negative ideal solution set is selected if it is maximization-oriented. These operations are shown below by Formula 4 and Formula 5, respectively.

$$A^{+} = \left\{ v_1^{+}, v_2^{+}, ..., v_n^{+} \right\} = \left\{ \left(\overset{Maksimum}{i} \; v_{ij} \mid j \in K \right), \left(\overset{Minimum}{i} \; v_{ij} \mid j \in K^{'} \right) \right\} \tag{4}$$

$$A^{-} = \left\{ v_1^{-}, v_2^{-}, ..., v_n^{-} \right\} = \left\{ \left(\overset{Minimum}{i} \; v_{ij} \mid j \in K \right), \left(\overset{Maksimum}{i} \; v_{ij} \mid j \in K^{'} \right) \right\} \tag{5}$$

Step 5. Ideal solution values are calculated.

Euclidean distances are used to find the distances of the evaluation criterion value for each Student (alternative) to the positive and negative ideal solution. Formulas concerning this calculation are given in Formula 6 and Formula 7.

$$D_i^{+} = \sqrt{\sum_{j=}^{n} \left(v_{ij} - v_j^{+} \right)^2} \qquad i = 1, 2, ..., m \qquad j = 1, 2, ..., n \tag{6}$$

$$D_i^{-} = \sqrt{\sum_{j=}^{n} \left(v_{ij} - v_j^{-} \right)^2} \qquad i = 1, 2, ..., m \qquad j = 1, 2, ..., n \tag{7}$$

Step 6. Alternative rankings are done based on ideal solution values.

When calculating each student's closeness to the ideal solution (CC$_i$), their distance to the positive and negative ideal solutions is used. As can be seen in Formula 8, the distance to the ideal solution is calculated with the ratio of the negative ideal solution to the total distance. This closeness value is between 0 and 1; when CC$_i$=0 it denotes absolute closeness to the negative ideal solution and when CC$_i$=1it denotes absolute closeness to the positive ideal solution.

$$C_i = \frac{D_i^{-}}{D_i^{+} + D_i^{-}} \qquad i = 1, 2, ..., m \qquad 0 \le C_i \le 1 \tag{8}$$

In the third method of the study, the following operation steps were carried out in the gradation of students' answers to the open-ended mathematics items with the fuzzy TOPSIS method (Chen, 2000).

Step 1. Decision-makers and criteria are selected.

Three experts were identified as the decision-makers in the study. Based on their expert opinion, the criteria were determined as (1) understanding the problem, (2) utilising what is given in the problem, (3) using operations in the solution of the problem, (4) adapting the formula and the rules to the problem, (5) following the order of operations by making connections between operations, (6) making no mistakes in the operations, and (7) executing the operations clearly and in detail.

Step 2. Appropriate linguistic variables are determined for the significance weights of the criteria; linguistic variables' levels are selected for alternatives according to the criteria.

The linguistic variables the scorers will assign to the criteria and students' answers are presented in Table 1.

Table 1. *Linguistic variables expressing the value weight for the criteria and the alternatives.*

Linguistic Variables for Criteria	Linguistic Variables for Alternatives
Very low (VL)	Very bad (VB)
Low (L)	Bad (B)
Somehow Low (SL)	Somehow Ba (SB)
Medium (M)	Medium (M)
Somehow High (SH)	Somehow Good (SG)
High (H)	Good (G)
Very High (VH)	Very Good (Very Good)

As can be seen in Table 1, seven options were identified for the sub-criteria and the alternatives. While it was thought that using fewer linguistic variables for the criteria and the alternatives would lessen the sensitivity of data, it was also believed that having more linguistic variables would not contribute to the study, either. In this respect, the number of linguistic variables was limited to seven to ensure an optimum sensitivity. The significance weights the scorers have given to the criteria of the scoring of the open-ended mathematics items are presented in Table 2.

Table 2. *Linguistic variables scorers provided for the significance brackets of decision criteria.*

Criteria	1st Scorer	2nd Scorer	3rd Scorer
Understanding the problem	VH	VH	VH
Using what is given in the problem	H	H	VH
Using operations in the solution of the problem	H	SH	VH
Adapting the formula and the rules to the problem	H	SH	H
Following the order of operations by making connections between operations	SH	M	H
Making no mistakes in the operations	SH	L	M
Executing the operations clearly and in detail	M	SH	VH

Formula 9 was used when calculating the significance levels the decision-makers assigned to the criteria. According to this formula, the operation is executed by taking the average of the weights the scorers gave to the criteria.

$$\tilde{w}_j = \frac{1}{K}\left[\tilde{w}_j^1 + \tilde{w}_j^2 + \cdots\cdots \tilde{w}_j^K\right] \tag{9}$$

Step 3. The linguistic variables determined by the decision-makers for the assessment of significance weights and alternatives are converted into triangular fuzzy numbers.

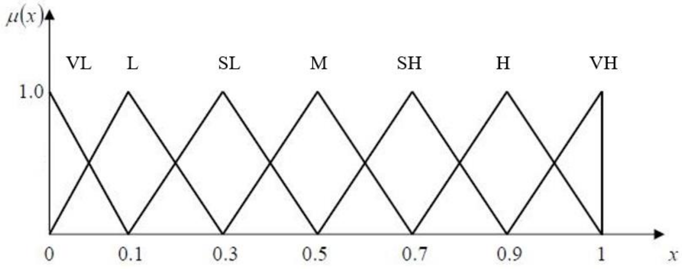
Within the scope of the story, triangular fuzzy numbers are preferred in the conversion of decisions to numbers. Triangular fuzzy number expressions of the criteria's significance weights are presented in Table 3.

Table 3. *Triangular fuzzy numerical expressions indicate the significance weights for the criteria.*

Linguistic Variables for Criteria	Triangular Fuzzy Numbers for Criteria
Very Bad (VB)	(0.0, 0.0, 0.1)
Bad (B)	(0.0, 0.1, 0.3)
Somehow Bad (SB)	(0.1, 0.3, 0.5)
Medium (M)	(0.3, 0.5, 0.7)
Somehow Good (SG)	(0.5, 0.7, 0.9)
Good (G)	(0.7, 0.9, 0.1)
Very Good (VG)	(0.9, 1.0, 1.0)

The graph of the significance weights in Table 3 is presented in Figure 3.

Figure 3. *Triangular fuzzy numbers show the significance weights for the criteria.*



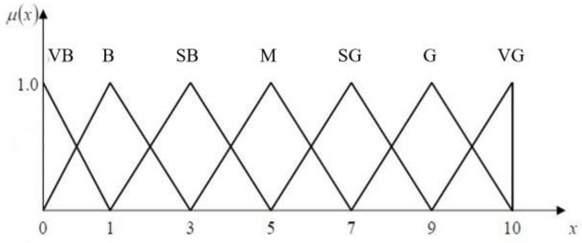
Fuzzy triangular numerical equivalents for the significance weights of alternatives are presented in Table 4.

Table 4. *Triangular fuzzy numbers show the significance weights for the alternatives.*

Linguistic Variables for Alternatives	Triangular Fuzzy Numbers for Alternatives
Very Low (VL)	(0, 0, 1)
Low (L)	(0, 1, 3)
Somehow Low (SL)	(1, 3, 5)
Medium (M)	(3, 5, 7)
Somehow High (SH)	(5, 7, 9)
High (H)	(7, 9, 10)
Very High (VH)	(9, 10, 10)

The graph of the significance weights of triangular fuzzy numbers in Table 4 is presented in Figure 4.

Figure 4. *Triangular fuzzy numbers show the significance weights for the alternatives.*



Values the scorers gave for the alternatives were calculated by Formula 10 and average values were thus obtained.

$$\tilde{x}_{ij} = \frac{1}{K}\left[\tilde{x}_{ij}^{1} + \tilde{x}_{ij}^{2} + \cdots\cdots \tilde{x}_{ij}^{K}\right] \tag{10}$$

Step 4. A fuzzy decision matrix and normalised fuzzy decision matrix are formed.

The fuzzy decision matrix shows the linguistic variables that each scorer assigned to the alternatives according to the criteria. Formula 11 was used for this operation.

$$\tilde{D} = \begin{array}{c} A_1 \\ A_2 \\ \cdots \\ A_m \end{array} \begin{pmatrix} \tilde{x}_{11} & \tilde{x}_{12} & \cdots & \tilde{x}_{1n} \\ \tilde{x}_{21} & \tilde{x}_{22} & \cdots & \tilde{x}_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} \tilde{W} = \left[\tilde{w}_1, \tilde{w}_2, \ldots, \tilde{w}_n\right] \tag{11}$$

The normalised decision matrix is devised out of the fuzzy decision matrix by dividing the fuzzy numbers in the column of each column to the largest upper limit in this column (Paksoy et al., 2013). Data in this study indicated that the highest fuzzy numbers for the 2nd, 4th, and 7th criteria was 9.7; and this value was used to establish the fuzzy decision matrix. The highest value of other criteria was 10 and it was left as it was. Then, all fuzzy numbers were normalised by dividing them by 10 thereby having the final version of the decision matrix. Formula 12 was used for the normalised fuzzy decision matrix.

$$\tilde{R} = \left[\tilde{r}_{ij}\right]_{mnx} \tag{12}$$

Since there are no negative criteria in this study, the benefit criterion was calculated by Formula 13.

$$\tilde{r}_{ij} = \left(\frac{a_{ij}}{c_j^{*}}, \frac{b_{ij}}{c_j^{*}}, \frac{c_{ij}}{c_j^{*}}\right), j \in B, c_j^{*} = \overset{\max}{i}\, c_{ij}, \tag{13}$$

Step 5. A weighted normalised fuzzy decision matrix is formed.

In this step Formula 14 was used to form the weighted normalised decision matrix by multiplying the normalised decision matrix with the criteria weights.

$$\tilde{V} = \left[\tilde{V}_{ij}\right]_{mxn} \tag{14}$$

In the weighted normalised fuzzy decision matrix, V_{ij} values are positive triangular fuzzy numbers, and their values vary between 0 and 1. Since each criterion has different significance degrees, a weighted normalised fuzzy decision matrix is calculated by Formula 15.

$$\tilde{V}_{ij} = \tilde{r}_{ij} \times \tilde{w}_j \tag{15}$$

Step 6. A fuzzy positive ideal solution and a fuzzy negative ideal solution are identified.

When determining the fuzzy positive and negative ideal solutions in this study, maximum values of the criteria were used for the fuzzy positive ideal solution and minimum values were used for the fuzzy negative ideal solution (Avcı Öztürk, 2018). The fuzzy positive ideal solution set for the normalised fuzzy decision matrix obtained by the triangular fuzzy numbers was calculated by Formula 16, and the negative ideal solution set was calculated by Formula 17.

$$A^* = \left(\tilde{V}_1^*, \tilde{V}_1^*, \ldots, \tilde{V}_n^*\right) \tag{16}$$

$$A^- = \left(\tilde{V}_1^-, \tilde{V}_1^-, \ldots, \tilde{V}_n^-\right) \tag{17}$$

The positive and negative ideal solution sets designed for the sub-criteria by using Formula 16 and Formula 17 are presented below.

$\tilde{A}^*$ = [(1.00, 1.00, 1.00), (1.00, 1.00 , 1.00), (0.97, 0.97, 0.97), (0.80, 0.80, 0.80), (0.87, 0.87, 0.87), (0.63, 0.63, 0.63), (0.87, 0.87, 0.87)]

$\tilde{A}^-$ = [(0.03, 0.03, 0.03) , (0.03, 0.03 , 0.03), (0.00, 0.00, 0.00), (0.01, 0.01, 0.01), (0.02, 0.02, 0.02), (0.00, 0.00, 0.00), (0.02, 0.02, 0.02)]

Step 7. The distance of each alternative first to the fuzzy positive ideal solution and then to the fuzzy negative ideal solution is calculated.

The distance of alternatives to the fuzzy positive ideal solution set was calculated by Formula 18 while their distance to the fuzzy negative ideal solution was calculated by Formula 19.

$$d_i^* = \sum_{j=1}^{n} d(\tilde{v}_{ij}, \tilde{v}_j^*), i = 1, 2, \ldots., m \tag{18}$$

$$d_i^- = \sum_{j=1}^{n} d\left(\tilde{v}_{ij}, \tilde{v}_j^-\right), i = 1, 2, \ldots, m \tag{19}$$

Formula 18 and Formula 19 show the distance between two fuzzy numbers. This distance is calculated by the Vertex method, which is developed to calculate the distance between fuzzy numbers.

$\tilde{A}$= (m_1, m_2, m_3) and $\tilde{B}$ = (n_1, n_2, n_3) are two fuzzy numbers, and Formula 20 was used to calculate the distance between $\tilde{A}$ and $\tilde{B}$ (Wang and Elhag, 2006; qtd. in Avcı Öztürk, 2018).

$$\mathrm{d}(\tilde{A}, \tilde{B}) = \sqrt{\frac{1}{3}[(m_1 - n_1)^2 + (m_2 - n_2)^2 + (m_3 - n_3)^2]} \tag{20}$$

Step 8. Closeness coefficients for each alternative are calculated.

Closeness coefficients were calculated by Formula 21 to rank the alternatives.

$$CC_i = \frac{d_i^-}{d_i^* + d_i^-} \tag{21}$$

Step 9. All alternatives are lined up according to closeness coefficients.

Students are ranked in descending order based on their closeness coefficient values; the student closest to 1 is considered the most successful and the student closest to 0 is considered the least successful.

Student ranking was done after obtaining the scores for the open-ended mathematics item via classical, TOPSIS, and fuzzy TOPSIS methods. To examine the correlation values between ranks Spearman Rank Correlation Coefficient was calculated.

3. RESULTS

Students' score averages and ranking according to the three scorers without using the criterion weights are presented in Table 5.

Table 5. *Ranking of students' scores based on the classical method.*

Rank	Students	Mean	Rank	Students	Mean
1	Student 16	36.3333	13	Student 15	28.6667
2	Student 17	35.3333	15	Student 22	28.0000
3	Student 12	34.0000	16	Student 3	27.3333
4	Student 20	33.3333	16	Student 7	27.3333
5	Student 6	33.0000	18	Student 13	26.6667
6	Student 24	31.0000	18	Student 19	26.6667
7	Student 1	30.6667	20	Student 8	25.6667
7	Student 4	30.6667	21	Student 10	25.3333
9	Student 9	30.0000	22	Student 14	25.0000
10	Student 2	29.3333	23	Student 18	23.6667
11	Student 21	29.0000	24	Student 25	22.6667
11	Student 23	29.0000	25	Student 11	20.3333
13	Student 5	28.6667			

As can be seen in Table 6, the highest mean is 36.33 and the lowest is 20.33. Since criterion weights were not used in the ranking, some students received the same score. When ranking these students, their student numbers were used in ascending order and there is no hierarchy among them. The ranking of the students' scores based on the classical TOPSIS method designed by Hwang and Yoon (1981) according to the closeness coefficient is presented in Table 6.

Table 6. *Students' ranking according to the closeness coefficient for the topsis application.*

Rank	Students	CC_i	Rank	Students	CC_i	Rank	Students	CC_i
1	Student 16	0.6826	10	Student 1	0.5408	19	Student 3	0.4611
2	Student 20	0.6628	11	Student 15	0.5385	20	Student 10	0.4428
3	Student 17	0.6280	12	Student 9	0.5336	21	Student 13	0.4014
4	Student 12	0.6276	13	Student 5	0.5221	22	Student 14	0.3585
5	Student 6	0.5779	14	Student 22	0.5160	23	Student 18	0.3453
6	Student 21	0.5677	15	Student 23	0.5146	24	Student 25	0.3400
7	Student 4	0.5576	16	Student 19	0.5097	25	Student 11	0.3057
8	Student 2	0.5491	17	Student 7	0.5021			
9	Student 24	0.5436	18	Student 8	0.4849			

Mean=0.5086

In Table 6 students' closeness coefficients are presented in descending order. According to this, Student 16 was in first place with 0.68 while Student 11 was last with 0.31. The mean of the class was $\bar{X}$ = 0.51. Students' scores were fuzzified according to the fuzzification steps suggested by Chen (2000). The ranking of students' closeness coefficient values after the operations for the fuzzy TOPSIS are presented in Table 7.

Table 7. *Students' ranking according to closeness coefficients for the fuzzy topsis application.*

Rank	Students	CC_i	Rank	Students	CC_i
1	Student 16	0.5676	13	Student 5	0.4510
2	Student 20	0.5594	15	Student 22	0.4438
3	Student 17	0.5532	16	Student 19	0.4342
4	Student 12	0.5514	16	Student 3	0.4151
5	Student 6	0.5150	18	Student 8	0.4081
6	Student 4	0.4879	18	Student 7	0.4032
7	Student 1	0.4833	20	Student 10	0.4010
7	Student 24	0.4830	21	Student 13	0.3903
9	Student 2	0.4784	22	Student 14	0.3645
10	Student 9	0.4756	23	Student 25	0.3456
11	Student 21	0.4752	24	Student 18	0.3383
11	Student 23	0.4640	25	Student 11	0.3030
13	Student 15	0.4612			

Mean=0.4501

The closeness coefficients in Table 7 show that Student 16 has the highest value with 0.57, which is followed by Student 20 with 0.56. The lowest value of the class, 0.30, belongs to Student 11. The mean of the class is $\bar{X}$ = 0.45.

The student rankings based on their scores obtained via classical, TOPIS, and fuzzy TOPSIS methods are presented in Table 8. Table 8 shows that Student 16 comes first in all methods and Student 11 comes last. While Student 17 comes second in the classical method, the same student comes third when ranked according to the multi-criteria decision-making methods, and Student 20 comes second.

Table 8. *Student rankings based on their scores obtained via classical, topis, and fuzzy topsis methods.*

Rank	Classical	TOPSIS	Fuzzy TOPSIS	Rank	Classical	TOPSIS	Fuzzy TOPSIS
1	Student 16	Student 16	Student 16	14	Student 15	Student 22	Student 5
2	Student 17	Student 20	Student 20	15	Student 22	Student 23	Student 22
3	Student 12	Student 17	Student 17	16	Student 3	Student 19	Student 19
4	Student 20	Student 12	Student 12	17	Student 7	Student 7	Student 3
5	Student 6	Student 6	Student 6	18	Student 13	Student 8	Student 8
6	Student 24	Student 21	Student 4	19	Student 19	Student 3	Student 7
7	Student 1	Student 4	Student 1	20	Student 8	Student 10	Student 10
8	Student 4	Student 2	Student 24	21	Student 10	Student 13	Student 13
9	Student 9	Student 24	Student 2	22	Student 14	Student 14	Student 14
10	Student 2	Student 1	Student 9	23	Student 18	Student 18	Student 25
11	Student 21	Student 15	Student 21	24	Student 25	Student 25	Student 18
12	Student 23	Student 9	Student 23	25	Student 11	Student 11	Student 11
13	Student 5	Student 5	Student 15				

When students' rank differences were examined, it was seen that most students were ranked in similar places no matter the method; the most obvious difference was with Student 21. Student 21 was ranked 6th in the TOPSIS method but was ranked 11th in the classical and fuzzy TOPSIS methods.

Spearman rank correlation coefficient values were obtained to test whether there is a relationship between student rankings and the Classical, TOPSIS, and fuzzy TOPSIS methods; these values can be found in Table 9.

Table 9. *Spearman rank correlation coefficient shows the relationship between student rankings and the employed method.*

Methods	Scoring via Classical Method	Scoring via TOPSIS	Scoring via Fuzzy TOPSIS
Scoring via Classical Method	1.000		
Scoring via TOPSIS	0.958*	1.000	
Scoring via Fuzzy TOPSIS	0.984*	0.975*	1.000

* p<0.01; n: 25

Table 9 shows that the highest ratio of similarity when it comes to student rankings was between fuzzy TOPSIS and Classical methods: r=0.984 and (p<0.01, r^2=0.968; n:25). The ratio of similarity between the two multi-criteria decision-making methods – TOPSIS and fuzzy TOPSIS – were found to be r=0.975 (p<0.01, r^2=0.951; n:25). All ranking methods used in this study have a positive high relationship.

4. DISCUSSION and CONCLUSION

Rigidly defined binary values such as yes/no, fast/slow, and good/bad are not always sufficient when making decisions in life. Some cases may contain qualities that fall under both the good and the bad. In such cases, the human mind makes a complex assessment by taking into account different conditions. Compared to classical logic, fuzzy logic is more compatible with the way humans think and it uses multi-level operations (Elmas, 2003; Yazırdağ, 2018). Fuzzy logic is a system of logic that overlaps with humans' ability to think in uncertain expressions (Ertuğrul, 2006). It indicates that assessment may have intermediate values as opposed to merely right and wrong results (Elmas, 2011; Uygunoğlu & Ünal, 2005). In decision-making, complex assessments are expressed in linguistic expressions. These linguistic expressions contain vagueness and variability (Yazırdağ, 2018). To alleviate this vagueness, linguistic expressions should be defined based on fuzzy sets and values that cannot be expressed clearly should be qualified approximately by using linguistic variables.

With the advancement of mathematical methods, different approaches to decision-making approaches have also emerged. Multi-criteria decision-making methods provide a more objective assessment alternative for the assessors along with classical and fuzzified ones. TOPSIS, which is one of the multi-criteria decision-making methods, is based on identifying the best alternative among the alternatives to be selected. The best alternative should geometrically have the shortest distance to the positive ideal solution and the longest distance to the negative ideal solution (Çakar, 2020; Tzeng & Huamg, 2011). In the Fuzzy TOPSIS method, fuzzy numbers are used to assign weight criteria, and linguistic scales are used in ranking alternatives (Madi, Garibaldi & Wagner, 2017). At the basis of the fuzzy TOPSIS method lies the fact that criteria used by the assessors may have different weights when assessing alternatives. This method eliminates the problems of subjectivity that emerge in making group decisions, and promotes more accurate decision-making (Ecer, 2007). The most significant point here is that different assessors can make different weightings and these weightings along with their numerical equivalents are included in the decision-making process. Fuzzy logic methods are more suitable for selecting the best among alternatives or classifying alternatives rather than a way of scoring. There are exemplary studies in the literature on this. In his 2006 study, Ertuğrul aimed to determine academics' performance by using the fuzzy logic method and categorised the results as "very inadequate," "inadequate," "normal," "successful," and "highly successful." Güler and Yücedağ (2017) developed a decision support system by using the fuzzy logic method to help vocational school students in selecting a field. Areas of the profession in which students may succeed were tried to be predicted by using the Self-concept scale. In their 2013 study, Çiçekli and Karaçizmeli aimed to determine students' ranking by using multiple criteria instead of merely evaluating their success based on their exam scores. A model was designed by using fuzzy AHP and students' rankings were examined. Wimatsari et al. (2013) aimed to help students at Udayan University in their selection of scholarships and determine the scholarship types according to established criteria. To this end, they combined Fuzzy TOPSIS and Fuzzy Multi-Criteria Decision-Making Methods to determine the functionality of scholarship selection. The study tried to determine the selection and rankings of students who would be given a scholarship.

Open-ended items play an important role in assessing advanced thinking skills, especially in in-class assessments; given the need for objectivity in scoring, testing the Fuzzy TOPSIS method's selection and ranking mechanism in the assessment of open-ended items was important. To this end, the criteria to be used in the study and their weights were determined by different experts, rankings were obtained by using both classical and fuzzy methods. Students' scores were not identical in the multi-criteria decision-making methods used in the study while some students received the same score when the classical method was used. This indicates that the classical method makes a less sensitive assessment even though it is easier to use. There was a strong and positive relationship among the rankings done by the Classical, TOPSIS, and Fuzzy TOPSIS methods. Weighted and fuzzified scores based on different criteria can be interpreted as not causing significant changes in the rankings of students compared to the classical method. Similar results have also been obtained in other studies in the literature. In a study conducted by Arslan in 2019, teacher performances were evaluated using fuzzy logic methods and the results were compared. In the study, the correlation value expressing the relationship between scores obtained via fuzzy and classical methods was determined, and it was concluded that there was a positive and high relationship between the two methods. In a study by Yılmaz in 2008, multi-criteria decision-making methods were used for selecting candidates applying for graduate studies. Within the scope of this study, criteria to be included in the assessment of student selection were determined, and the weighting of these criteria by pairwise comparison was carried out. Then, candidates were ranked using the AHP, TOPSIS, and Weighted Product methods, and the results were compared. Nursikuwagu et al. modeled and examined student competencies in vocational schools using the Fuzzy TOPSIS method in 2018. At the end of their study, they declared the Fuzzy TOPSIS as a simpler and more dynamic model that produces effective results in determining competencies compared to the traditional method which uses the average value.

The subjectivity of the scorer in scoring open-ended items affects the validity and reliability of scores (Haladyna, 1997; Nitko &Brookhart, 2014; Royal &Hecker, 2016). To prevent this, analytical or holistic scoring rubrics are used (Karakaya, 2022; Kutlu et al., 2014). When using these graded scoring rubrics, it is assumed that criteria weights are the same for each scorer, and the tools are designed accordingly. One wonders how rankings change when, rather than binary scoring, the weights of criteria change and when scores are considered with their intermediate values. The traditional method is undoubtedly the most common because it is practical for educators. Although there is a scoring key for scoring open-ended questions, evaluators have to evaluate according to these standard scores. Fuzzy logic, unlike classical logic, allows evaluators to weight and score criteria. This study focuses on how to apply fuzzification to the scores obtained from one of the most frequently used tools in the field of educational sciences and examines the resulting outcomes. Studies focusing on the differences of the methods can be done similarly. The results obtained from the study have focused more on ranking than on scoring, based on the preferred methods, and have shown that there were no significant changes in students' rankings. Given the increasing prevalence of fuzzy logic studies in the field of education, there will be a need to know the details of the algorithm, how the fuzzification mechanism differs from traditional methods, and how the selections and rankings yield results. The results of this study are expected to provide a cue for other studies. The study wanted to examine open-ended questions, an important component of assessment and evaluation, since it tested an example of especially the fuzzification process. On the other hand, external variables were kept at a minimum by limiting the scope. In this regard, conducting broader studies with both simulated and real data and examining their results would be beneficial. The fact that the algorithm can be created practically by using any coding language would make it easier for researchers to develop/test models in the future. Similar comparisons can be made not only by TOPSIS but by using other fuzzy methods, and the results of these comparisons can be examined.

Acknowledgments

A draft version of this study has been previously presented at the 7th International Congress on Measurement and Evaluation in Education and Psychology in 2021.

This study was based on Aykut Çitçi's master thesis titled "Using a fuzzy logic approach to scoring open-ended questions: example of the fuzzy topsis method" (2021), at Kocaeli University Institute of Social Sciences.

Declaration of Conflicting Interests and Ethics

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

Contribution of Authors

Aykut Çitçi: Investigation, Resources, Methodology, Visualization, Software, Formal Analysis, Validation, and Writing-original draft. **Fatih Kezer**: Supervision.

Orcid

Aykut Çitçi https://orcid.org/0000-0002-5473-7097
Fatih Kezer https://orcid.org/0000-0001-9640-3004

REFERENCES

Altun, M. (2002). *İlköğretim ikinci kademede matematik öğretimi [Mathematics teaching in the second level of primary education]*. Alfa Yayınevi.

Armağan, H. (2008). *A new approach for student academic performance evaluation* [Unpublished Master's thesis]. Süleyman Demirel University.

Arslan, M. (2019). *Evaluation of teacher performances with fuzzy logic method* [Unpublished Master's thesis]. Van Yüzüncü Yıl University.

Avcı Öztürk, B. (2018). *Analitik hiyerarşi süreci ve topsis: Bulanık uygulamaları ile [Analytic hierarchy process and topsis: With fuzzy applications]*. Dora Basın Yayın Dağıtım.

Badger, E., & Thomas, B. (1992). Open-ended questions in reading. *Practical Assessment, Research & Evaluation, 3*(4), 03. https://doi.org/10.7275/fryf-z044

Bakanay, D. (2009). *The assessment of micro-teaching performance by fuzzy logic*. [Unpublished Master's thesis]. Marmara University.

Baykal, N., Beyan, T. (2004). *Bulanık mantık ilke ve temelleri [Principles and fundamentals of fuzzy logic]*. Bıçaklar Kitabevi.

Bostan, A. (2017). *Measurement in criterion based tests with utilization of fuzzy logic and usage of this methodology in computerized adaptive tests* [Unpublished Doctoral thesis]. Gazi University.

Bush, M. (2001). A multiple choice test that rewards partial knowledge. *Journal of Further and Higher Education, 25*(2), 157–163. https://doi.org/10.1080/03098770120050828

Chen, T.C. (2000). Extensions of the topsis for group decision - making under fuzzy environment. *Fuzzy Sets And Systems*, *114*, 1-9. https://doi.org/10.1016/S0165-0114(97)00377-1

Cheng, C.H. (1996). Evaluating naval tactical missile systems by fuzzy ahp based on the grade value of membership function. *Europan Journal of Operational Research*, *96*(2), 343-350. https://doi.org/10.1016/S0377-2217(96)00026-4

Cooney, T.J., Sanchez, W.B., Leatham, K., & Mewborn, D.S. (2004). Open-ended assessment in math: A searchable collection of 450+ questions.

Çakar, T. (2020). *Bulanık çok ölçütlü karar verme yöntemleri [Multicriteria fuzzy decision making methods]*. İstanbul Gelişim Üniversitesi Yayınları.

Coşkunırmak, Y. (2010). *Fuzzy linear programming and an application of fuzzy goal programming in local governments* [Unpublished Master's thesis]. Çukurova University.

Damlar Demirci, P.(2019). *The investigation of open-ended items scoring methods by generalizability theory* [Unpublished Master's thesis]. Ege University.

Dündar, S., Fatih, E., & Özdemir, Ş. (2007). Fuzzy topsis yöntemi ile sanal mağazalarin web sitelerinin değerlendirilmesi [Evaluation of web sites of virtual stores with fuzzy topsis method]. *Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi*, *21*(1), 287-305. https://dergipark.org.tr/en/pub/atauniiibd/issue/2691/35409

Ecer, F. (2007). *Assessing candidates in human resource selection with the Fuzzy Topsis method and an application* [Unpublished Doctoral thesis]. Afyon Kocatepe University.

Elmas, Ç. (2003). *Bulanık mantık denetleyiciler: Kuram, uygulama, sinirsel bulanık mantık [Fuzzy logic controllers: Theory, application, neural fuzzy logic]*. Seçkin Yayıncılık.

Elmas, Ç. (2011). *Yapay zeka uygulamaları [Artificial intlligence applications]*. Seçkin Yayıncılık.

Erdin, C. (2007). *Fuzzy goal programming and an applied study in management* [Unpublished Doctoral thesis]. İstanbul University.

Eroğlu, G. (2012). The transition from classical logic to modern logic: some of the foundations grounding the rise of modern logic.*Hikmet Yurdu Düşünce - Yorum Sosyal Bilimler Araştırma Dergisi, 5*(9), 115-135. http://dx.doi.org/10.17540/hy.v5i9.167

Ertuğrul, İ. (2006). Akademik performans değerlendirmede bulanık mantık yaklaşımı [Fuzzy logic approach in academic performance evaluation]. *İktisadi ve İdari Bilimler Dergisi, 20*(1), 155-176. https://dergipark.org.tr/en/pub/atauniiibd/issue/2689/35353

Ertuğrul, İ., & Karakaşoğlu, N. (2008). Banka şube performasnlarının VIKOR yöntemi ile değerlendirilmesi [Evaluation of bank branch performances with VIKOR method]. *Endüstri Mühendisliği dergisi, 20*(1), 19-28. https://www.mmo.org.tr/sites/default/files/c4692732b25c1ee_ek.pdf

Geer, J.G. (1988). What do open-ended questions measure? *Public Opinion Quarterly, 52*(3), 365–367. https://doi.org/10.1086/269113

Güler, O., & Yücedağ, İ. (2017). Fuzzy logic-based approach to site selection problem of vocational secondary school students. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 32*(1), 111-122. https://doi.org/10.16986/HUJE.2016018727

Haladyna, T.M. (1997). *Writing tests items to evaluate higher-order thinking*. Allyn & Bacon.

Hasırcı, N. (2010). *İbn Teymiyye'nin mantık eleştirisi [İbn Teymiyye's criticism of logic]*. Araştırma Yayınları.

Hocalar, E. (2007). *An application of fuzzy balanced scorecard system for managing performance in higher education organizations* [Unpublished Master's thesis]. Sakarya University.

Hwang, C.L., & Yoon, P. (1981). *Multiple attribute decision making in: lecture notes in economics and mathematical systems*. Springer-Verlag, Berlin.

Kaptanoğlu, D., & Özok, A.F. (2006). Akademik performans değerlendirmesi için bir bulanık model [A fuzzy model for academic performance evaluation]. *İTÜ dergisi, 5* (1), 193-204. http://itudergi.itu.edu.tr/index.php/itudergisi_d/article/view/627

Karadeniz, A. (2016). *Design, evaluation and implementation of a system aimed at assessment of learning achievement through open-ended questions in massive, open and distance learning* [Unpublished Doctoral thesis] Anadolu University.

Karakaya, İ. (2022). *Açık uçlu soruların hazırlanması, uygulanması ve değerlendirilmesi [Preparation, implementation and evaluation of open-ended questions]*. Pegem Akademi.

Karataş, İ. (2018). Comparison of fuzzy logic, classic logic and symbolic logic. *European Journal of Educational & Social Sciences, 3*(2), 144-163. https://dergipark.org.tr/en/pub/ejees/issue/40157/477684

Köz, İ. (2002). Aristoteles mantığı ile felsefe-bilim ilişkisi [Relationship between philosophy-science and Aristotelian logic]. *Ankara Üniversitesi İlahiyat Fakültesi Dergisi, 43*(2), 55-37. https://dergipark.org.tr/en/download/article-file/583546

Klufa, J. (2018). Multiple choice question tests–advantages and disadvantages. *Recent Advances in Educational Technologies*. ISBN: 978-1-61804-322-1. https://www.inase.org/library/2015/zakynthos/bypaper/EDU/EDU-07.pdf

Kutlu, Ö., Doğan, C.D., & Karakaya, İ. (2014). *Öğrenci başarısının değerlendirilmesi: Performansa ve portfolyoya dayalı durum belirleme [Assessment of student achievement: Performance and portfolio-based assessment]*. Pegem Akademi.

Madi, E.N., Garibaldi, J.M., & Wagner, C. (2017, July). Exploring the use of type-2 fuzzy sets in multi-criteria decision-making based on TOPSIS. In *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1-6). IEEE. https://doi.org/10.1109/FUZZ-IEEE.2017.8015664

McMillan, J.H. (2017). *Classroom assessment: Principles and practice that enhance student learning and motivation*. Pearson.

McNeill, F.M., & Ellen T. (1994). *Fuzzy Logic: A practical approach*. Academic Press.

MEB. (2017). *8. sınıf merkezi ortak sınavları matematik dersi açık uçlu soru ve yapılandırlmış cevap anahtarı örnekleri [8th grade central common exams mathematics open-ended questions and structured answer key examples]*. MEB http://odsgm.meb.gov.tr/meb_iys_dosyalar/2017_09/15135732_Mat_acik_uclu.pdf

MEB. (2020). *Milli eğitim istatistikleri: Örgün eğitim 2019-20 [National education statistics: Formal education 2019-20]*. Türkiye İstatistik Kurumu.

MEB. (2020). *Ortaöğretim kurumlarına ilişkin merkezi sınav kılavuzu [Central exam guide for secondary education]*. Eğitim, Analiz ve Değerlendirme Raporları Serisi.

Miller, M.D., Linn, R., & Gronlund, N.E. (2009). *Measurement and assessment in teaching*. Merrill Prentice Hall.

Nitko, A.J., & Brookhart, S.M. (2014). *Education assessment of students*. Merrill Prentice Hall.

Opricovic, S., &Tzeng, G. H. (2004). Compromise solution by mcdm methods: A comparative analysis of vıkor and topsıs. *European Journal of Operational Research*, *156*(2), 445-455. https://doi.org/10.1016/S0377-2217(03)00020-1

Öksüz, Y., & Güven Demir, E. (2019). Comparison of open ended questions and multiple choice tests in terms of psychometric features and student performance. *Hacettepe University Journal of Education, 34(1*), 259-282. https://doi.org/10.16986/HUJE.2018040550

Öner, N. (1986). *Klasik mantık [Classical logic]*. Ankara Üniversitesi Basımevi.

ÖSYM. (2017). *Açık uçlu sorular hakkında bilgilendirme ve açık uçlu soru örnekleri [Information about open-ended questions and examples of open-ended questions]*. OSYM https://www.osym.gov.tr/TR,12909/2017-lisans-yerlestirme-sinavlari-2017-lys-acik-uclu-sorular-hakkinda-bilgilendirme-ve-acik-uclu-soru-ornekleri-05012017.html

ÖSYM. (2021). *2021 yılı yükseköğretim kurumları sınavı (YKS) kılavuzu [2021 higher education institutions exam guide]*. OSYM.

Özdağoğlu, A. (2016). *Bulanık işlemler, durulaştırma ve sözel eşikler: Bilgisayar uygulamalı örneklerle [Fuzzy operations, stabilization and verbal thresholds: With computer-implemented examples]*. Detay Yayıncılık.

Paksoy, T., Pehlivan, N.Y., & Özceylan, E. (2013). *Bulanık küme teorisi [Fuzzy set theory]*. Nobel Akademik Yayıncılık.

Popham, W. J. (1999). *Classroom assessment: What teachers need to know*. Allyn & Bacon.

Royal, K.D., & Hecker, K.G. (2016). Rater errors in clinical performance assessments. *Journal of Veterinary Medical Education*, *43*(1), 5-8. https://doi.org/10.3138/jvme.0715-112R

Ross, T.J. (2010). *Fuzzy logic with engineering applications*. John Wiley & Sons Ltd.

Sarı, M., Murat, Y.S., & Kırabalı, M. (2005).Fuzzy modeling approach and applications. *Dumlupınar Üniversitesi Fen Bilimleri Enstitüsü Dergisi 9*, 77-92. https://dergipark.org.tr/en/pub/pufbed/issue/36210/407845

Taylan, N. (2008). *Ana hatlarıyla mantık [Logic in outline]*. Ensar Neşriyat Yayıncılık.

Tekin, H. (2010). *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education]*. Yargı Yayınevi.

Topçu, H. (2014). *Examination of fuzzy AHP method and an application on the problem of reference book selection to KPSS preparations* [Unpublished Master's thesis]. Marmara University.

Turgut, M.F., & Baykul, Y. (2012). *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education]*. Pegem Akademi.

Türe, H. (2006). *Fuzzy linear programming and an application* [Unpublished Master's thesis]. Gazi University.

Tzeng, G.H., & Huang, J.J. (2011). *Multiple attribute decision making: methods and applications*. CRC press.

Uygunoğlu, T., & Osman, Ü. (2005). Fuzzy logic approach on the effect of seyitömer fly ash on compressive strength of concrete. *Yapı Teknolojileri Elektronik Dergisi*, *1*(1), 13-20. https://dergipark.org.tr/en/pub/yted/issue/22222/238558

Van de Walle, J.A., Karp, K.S., & Bay-Williams, J.M. (2014). *Elementary and middle school mathematics*. Pearson.

Wimatsari, G.A. Ketut, G.P., & Buana, P.W. (2013). Multi-attribute decision-making scholarship selection using a modified fuzzy topsis. *International Journal of Computer Science Issues*, *10*(1), 309-317. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=7f6d03cd62a7d6c4f80eca3b27c788c6a5000a5d

Yazırdağ, M. (2018). *Supply system with fuzzy AHP and fuzzy TOPSIS methods: An application in gendarmerie* [Unpublished Master's thesis]. Gazi University.

Yen, J., & Langari, R. (1999). *Fuzzy logic: intelligence, control, and information*. Prentice Hall.

Yılmaz, R. (2008). *Student selection for postgraduate education in Turkey: an empirical study at Turkish Military Academy Defense Sciences Institute* [Unpublished Master's thesis]. Kara Harp Okulu.

Zadeh, L.A. (1965). Fuzzy sets. *Information and Control*, *8*, 338-353. https://doi.org/10.1016/S0019-9958(65)90241-X

Zimmermann, H. J. (1978). Fuuzy programming and linear programming with several objetive functions. *Fuzzy Sets and Systems*, *1*(1), 45-55 https://doi.org/10.1016/0165-0114(78)90031-3

Zimmermann, H.J. (2001). *Fuzzy set theory-and its applications*. Springer Science.