Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN:1309-6575 *Yaz* 2022 *Summer* 2022 *Cilt: 13- Sayı: 2* *Volume: 13- Issue: 2*

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309 – 6575

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (EPOD) yılda dört kez yayınlanan hakemli ulusal bir dergidir. Yayımlanan yazıların tüm sorumluluğu ilgili yazarlara aittir.

Journal of Measurement and Evaluation in Education and Psychology (JMEEP) is a national refereed journal that is published four times a year. The responsibility lies with the authors of papers.

İletişim
e-posta: epodderdergi@gmail.com
Web: https://dergipark.org.tr/tr/pub/epod

Contact
e-mail: epodderdergi@gmail.com
Web: http://dergipark.org.tr/tr/pub/epod

Dizinleme / Abstracting & Indexing
Emerging Sources Citation Index (ESCI), DOAJ (Directory of Open Access Journals), SCOPUS, TÜBİTAK TR DIZIN Sosyal ve Beşeri Bilimler Veri Tabanı (ULAKBİM), Tei (Türk Eğitim İndeksi), EBSCO

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (Haziran 2022, Sayı: 13-2)

Journal of Measurement and Evaluation in Education and Psychology (June 2022, Issue: 13-2)

İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2022; 13(2); 105-116

Investigation of the Missing Data Imputation Methods on Characteristic Curve Transformation Methods Used in Test Equating*

Gülden ÖZDEMİR** Burcu ATAR***

Abstract

In this research, the aim was to evaluate the effect of zero imputation and multiple imputation missing data handling methods on item response theory (IRT) based test equating methods under different conditions. Data in this study was obtained from the administration of the TIMSS 2019 eighth-grade science test. Data sets were formed by randomly selecting a sample of 1000 students with full data from booklets 7 and 8. By deleting data under a completely random missing data mechanism within the scope of common-item nonequivalent groups (CINEG) design, four different data sets were obtained with the missing data rates of 10% or 20% in the new test or in both tests. The missing da ta problem was solved by using zero imputation and multiple imputation methods from these data sets. In this way, 8 different data sets were formed. Then, scaling transformation was performed by using characteristic curve transformation methods (Haebara, Stocking-Lord). Test equating results were reported in terms of observed scores. The root mean square error (RMSE) was used as the evaluation criterion to determine the error involved in test equating. As a result, it was determined that in the case of 10% missing data in both tests, generally lower RMSE values were obtained. It was observed that the multiple imputation method, one of the methods for handling missing data, was the method that produced RMSE values that were both the lowest and closer to the full data set as a reference value compared to the zero-imputation method. In addition, it was determined that, when compared to the Haebara method, Stocking-Lord method, one of the characteristic curve transformation methods, produced lower RMSE values and these values were closer to the full data set, which was taken as a reference value.

Keywords: Missing data, zero imputation method, multiple imputation method, test equating, characteristic curve transformation methods

Introduction

Exams play an important role in making some critical decisions in the lives of individuals. Selection of personnel for an institution, promotion, change of title, determination of level, selection of students for higher education, etc., are among those exams. Such exams are carried out at the national or international level for various purposes. There are some exams that can be administered multiple times a year (ALES, YDS, YÖKDİL, TOEFL, etc.) or in certain cycles (TIMSS, PISA, PIRLS, etc.). For these exams, alternative test forms consisting of different items are being developed to ensure the safety of the items (Cook & Eignor, 1991). Alternative test forms, which are also called parallel test forms, are very difficult to produce. There may be slight differences between the difficulty levels of the forms (Kolen & Brennan, 2004). It is important for validity that such exams treat all individuals who take different test forms equally and impartially (Kan, 2011). For this reason, in order to directly compare the performances of individuals who answered different items, their scores should be placed on a common scale. With this method, called test equating, different test forms are equated and the scores obtained become comparable (Cook & Eignor, 1991).

* This study was produced from the first author's doctoral thesis. Also, this study was presented at the 7th International Congress on Measurement and Evaluation in Education and Psychology (September 1-4, 2021, Ankara/Turkey).

** Teacher, Ministry of National Education, Ankara-Turkey, gulden.ozdemir@meb.gov.tr, ORCID ID: 0000-0002-3150-9438
*** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, burcua@hacettepe.edu.tr, ORCID ID: 0000-0003-3527-686X

To cite this article:
Özdemir, G., & Atar, B. (2022). Investigation of the missing data imputation methods on characteristic curve transformation methods used in test equating. *Journal of Measurement and Evaluation in Education and Psychology, 13*(2), 105-116. https://doi.org/10.21031/epod.1029044

Received: 27.11.2021
Accepted: 28.04.2022

Different data collection patterns (random groups design, single group design, and common-item nonequivalent groups design) can be used in test equating studies. In this study, the common-item nonequivalent groups (CINEG) design was used since the data set in the study was collected in accordance with this design. At CINEG, different groups take different forms of tests. These test forms have common items. Common items are used to reveal the equating relationship between the two groups by comparing the performances of each group (Hambleton et al., 1991; Kolen & Brennan, 2004). Common items include structure, item type, content, etc. of the entire test. In this respect, it is recommended to have a smaller version (representative) of the test (Angoff, 1971). In studies conducted in the related literature, it was aimed mainly to determine the test equating method that shows better performance under different conditions (such as sample size, number of items, item threshold parameter difference, item parameter drift, differential item functioning, guessing, mixed-format test, etc.), (Atalay Kabasakal, 2014; Aytekin Kazanç, 2019; Demirus, 2015; Han, 2008; Karagül, 2020; Kilmen, 2010; Mutluer, 2013; Tian, 2011; Uysal, 2014; Wolf, 2013) or it was aimed to compare the performances of test equating methods based on Classical Test Theory (CTT) and Item Response Theory (IRT) (Mutluer, 2021; Skaggs, 2005; Yang, 1997). Test equating methods are based on different theories such as CTT and IRT (Ryan & Brockmann, 2009). However, the research results show that test equating methods based on IRT generally give better results than methods based on CTT, depending on the sample size and the number of items (Hambleton & Jones, 1993; Jabrayilov et al., 2016; Mutluer, 2021; Yang, 1997).

Different test equating methods are used in IRT. Accordingly, this method can be examined under two headings: concurrent calibration method and separate calibration method. In the concurrent calibration method, item parameters are estimated together for both test forms. The estimated parameters are automatically on the same scale. In the separate calibration method, item parameters are estimated parameters on different scales, and linking or a scale transformation is needed. These transformation methods are referred to as moment methods (mean-mean, mean-sigma) and characteristic curve methods (Haebara and Stocking Lord; Kolen & Brennan, 2004). The mean-mean method described by Loyd and Hoover (1980) calculates by using the means of discrimination (*a*) and difficulty (*b*) parameters. Thus, A slope and B constant values are obtained, which help to determine the individual's ability levels in different test forms. Mean-sigma method described by Marco (1977) calculates by using the mean and standard deviation values of the *b* parameter. Thus, the coefficients A slope and B constant are determined. In the characteristic curve transformation methods developed by Haebara (1980) and Stocking and Lord (1983), parameters *a*, *b* and *c* (chance parameter) are estimated simultaneously. According to the Haebara (1980) approach, the difference between the item characteristic curves is a function that gives the sum of the squares of the differences between the item characteristic curves of each item for respondents at a given ability level. In the function developed by Stocking and Lord (1983), it is the square of the sum of the difference between the item characteristic curves of each item for respondents at a certain ability level. Whichever of these methods is used, IRT equating is performed after the item calibration and scale transformation steps. Test forms can be used interchangeably as a result of test equating, but proof of validity must be submitted for each alternative form used in national or international exams where important decisions about individuals will be made.

Missing data is an essential factor in making critical decisions about individuals, which may pose a question mark about test validity (Hohensinn & Kubinger, 2011). Missing data occurs as a result of not answering some of the items in the exams or leaving them blank. Missing data may cause a narrowing in the data set, as well as weakening the power of the estimations to be made (Rubin, 1987). On the other hand, there are also studies on missing data such as internal consistency, variance analysis parameters, model-data fit and item-data fit, psychometric properties of scales, measurement invariance, and changing item function affect (Akbaş, 2014; Bayhan, 2018; Enders, 2004; Hohensinn & Kubinger, 2011; Işıkoğlu, 2017, Öztemür, 2014; Tamcı, 2018). In addition, standard analysis methods are prepared according to the full data set and cannot be applied to missing data sets (Rubin, 1987).

Missing data can be on three different missing data mechanisms such as missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR), depending on whether the probability of missing data in a variable is related to other variables. MCAR data mechanism is the situation where there is no relationship between the probability of missing data in a variable and the values of this variable and the other variables; that is, it is completely random (Enders, 2010). In this case, the missing data can be negligible provided that the MCAR assumptions are met, but providing the MAR or MNAR assumptions does not provide sufficient evidence for the negligibility of the missing data. Therefore, when missing data is not negligible, it is necessary to use an appropriate method of handling missing data in the analyses to be made regarding the psychometric properties of the tests (Demir, 2013). For this reason, different methods for handling missing data have been developed. These methods were described by Little and Rubin (2002) as deletion methods (listwise deletion, partwise deletion), imputation methods (average imputation, regression imputation, hot/cold deck imputation, etc.), and model-based methods (expectation-maximization, multiple imputation method, Bayesian imputation methods, etc.).

The statistical software used in the researches offers treating as not administered or treating as incorrect as the default method for processing missing data (Ertoprak, 2017). In the method of treating as incorrect, also called the zero imputation method, if the value of 0 is among the values that can be obtained by observation in the data set, then the value of 0 is imputed instead of the missing data (McKnight et al., 2007). In the multiple imputation method, which is one of the newer and probability-based approaches, two or more values are imputed to replace the missing data, reflecting the distribution of possible values (Rubin, 1987). In studies comparing the performances of different methods, the method with the best performance; it has been determined that the rate of missing data varies according to different conditions such as missing data mechanism and sample size (Akbaş, 2014; Allison, 2003; Koçak, 2016; Wu et al., 2015).

Missing data influences the test equating results performed on forms that use different missing data handling methods. In order to equate different test forms with or without error, the data set should be analyzed using the most appropriate missing data handling method. Numerous studies have been encountered on methods of handling missing data or test equating, but it has been observed that studies that deal with both concepts are limited (Ertoprak, 2017; Kim, 2015; Ngudgratoke, 2009; Shin, 2009). When these studies are examined, it has been found that these studies are limited to the 3-parameter logistic model (3PLM), one of the characteristic curve transformation methods based on the Item Response Theory (IRT), and Stocking-Lord (SL) and root mean square error (RMSE) and equating bias (BIAS) values. Most of these studies used simulated data. Therefore, in this study, it is aimed to examine the impacts of the missing data imputation methods on characteristic curve transformation methods used in test equating under different conditions on the real data set.

In this regard, the research questions addressed in this study are:

1. When the test forms obtained by applying the zero imputation method are equated according to the characteristic curve transformation methods, how do the RMSE values change according to the location of missing data in the test forms (both tests, the new test) and the missing data rate (10%, 20%)?
2. When the test forms obtained by applying the multiple imputation method are equated according to the characteristic curve transformation methods, how do the RMSE values change according to the location of the missing data in the test forms (both tests, the new test) and the missing data rate (10%, 20%)?

Method

In this study, it was aimed to determine the effect of missing data coping methods on test equating methods under different conditions. For this purpose, data sets in which the method of dealing with missing data was applied according to the determined conditions were produced and it was planned to find the method that gave the least error. In the research, equating methods are compared with real data sets under different conditions in a controlled manner. The research that contributes to the theory is basic research in this respect (Karasar, 2009).

Data Set

In this study, the Trends in International Mathematics and Science Study (TIMSS) 2019 data were employed. For the study, the top ten countries (Singapore, Taiwan, South Korea, Russia, Finland, Lithuania, Hungary, United States, Sweden, and Portugal) that administer computer-based applications (eTIMSS) at the eighth-grade level in the TIMSS 2019 science achievement test were selected. Then, all the booklets were examined, and the booklets numbered 7 and 8, which had the highest number of items, were scored dichotomously, and the number of common items, and the number of respondents, were used. After the student answers containing missing data were removed, 2249 student data were obtained for booklet 7, and 2277 student data were obtained for booklet 8. The data set was formed by randomly selecting a sample of 1000 people from these booklets. The sample size of 1000 was selected as a generous number that would provide accurate results and a good baseline for comparison (Swaminathan & Gifford, 1983).

Data Collection Instruments

The data used in the study were obtained from the database (https://timss2019.org/international-database/) published by the International Association for the Evaluation of Educational Assessment (IEA). Half of the countries participating in TIMSS 2019 used the eTIMSS application for the first time. There were 14 student handbooks consisting of eighth-grade math and science items and common items to make connections between the booklets (Mullis et al., 2020). For each of the booklets numbered 7 and 8 used in this study, a total of 25 dichotomously scored science items were selected, 13 of which were common and 12 were non-common. Since the students who responded to the test forms in question were different, the equating pattern of the study was determined as the common test design in the unequal groups. According to Angoff (1971), in equating studies to be carried out in unequal groups, equating errors are to be minimized when the number of common items is equal to at least 20% of the total number of items.

Data Analysis

Since this research is based on IRT, the basic assumptions were tested first. Eigenvalues were calculated for Booklet 7 and Booklet 8 to test the unidimensionality assumption. In both booklets, it was determined that the eigenvalue of the first factor (6.38, 6.20, respectively) was more than three times the eigenvalue of the second factor (1.56, 1.83, respectively). This is an indication that the measured structure is one-dimensional (Büyüköztürk, 2011). Yen’s (1984) Q3 statistic was used to test the local independence assumption. It was determined that the Q3 values calculated for both booklets did not exceed .20. The fact that the Q3 value was calculated based on the correlation between residual values not exceeding .20 provides evidence for local independence (Zenisky et al., 2001). As a result of the preliminary analysis, it was seen that the assumptions of unidimensionality and local independence were supported. Another assumption of IRT is model-data fit. In order to perform a test equating based on IRT between the booklets belonging to the data sets, the model-data fit condition was checked. The purpose of evaluating this fit was to determine how well an IRT model fits the data (Hambleton & Swaminathan, 1985; DeMars, 2010). -2loglikelihood values and chi-square (X^2) statistics were used to determine which IRT model was compatible with the data.

Table 1
Determination of Model Data Fit

	Booklet 7		Booklet 8	
Model	2PLM	3PLM	2PLM	3PLM
-2loglikelihood	28595.60	28560.98	27775.54	27749.93
Number of parameters	74	99	74	99
Difference		34.62		25.61

According to Table 1, 2PLM (two-parameter logistic model) was preferred because the difference in the likelihood obtained from 2PLM and 3PLM (three-parameter logistic model) for both forms was not statistically significant ($p < .01$). For this reason, 2PLM was preferred for parameter estimation in the research.

After meeting the assumptions, data sets were created from the booklets according to the different conditions in the research. From the data sets, data were deleted under the missing completely at random (MCAR) data mechanism via R. Little's MCAR has been tested to see if the missing data in the data sets is completely random. According to the Little's MCAR test result, it was determined that the missing data was MCAR ($p > .05$). Four different data sets were obtained in the new test or in both tests of 10% or 20% missing data. In this study, booklet 7 was determined as "new test (NT)", booklet 8 as "old test (OT)", booklet 7 and booklet 8 as "both tests (BT)". Using the zero imputation and multiple imputation methods from these data sets, 8 different data sets were created to solve the missing data problem. Detailed information about the data sets formed within the scope of the research is given in Table 2.

Table 2

Data Sets Formed Within the Scope of the Research

Sample size	Missing data rate	Missing data location	Techniques of handling missing	
			Zero imputation	Multiple imputation
1000	10%	New test	DS1*	DS2
		Both test	DS3	DS4
	20%	New test	DS5	DS6
		Both test	DS7	DS8

* DS: Data Set

For item parameter estimation, the Expected A Posteriori (EAP) method (Embretson & Reise, 2000), which uses prior distribution information, was utilized. Analyses were performed with the "mirt" package (Chalmers et al., 2021) in the R software. Since the predicted item and ability parameters are in different scales, they should be placed on a common scale; that is, scale transformation should be performed (Kim & Hanson, 2000). In the research, scale transformation was performed by using characteristic curve transformation methods (Haebara *H*, Stocking-Lord *SL*), which is one of the test equating methods based on IRT. The scores obtained from the new form were equal to the scores obtained from the old form. Analyses were performed with the "equateIRT" package (Battauz, 2021) in R and test scores were reported in terms of observed scores. RMSE was used as the evaluation criterion to determine the error involved in test equating. The RMSE index provides a statistic based on the difference between the actual ability level and the predicted ability level. Equation 1 used to calculate the RMSE coefficient is given below. While writing the equation, Harris and Crouse (1993) and Keller and Keller (2011) were utilized.

$$\text{RMSE} = \sqrt{\frac{1}{f}\left(\sum_{i=1}^{f}\left(\hat{\Theta}_i - \Theta_i\right)^2\right.}, \qquad (1)$$

where $\hat{\Theta}_i$, predicted skill level; Θ_i, actual skill level; f, frequency.

Results

When the test forms obtained by using missing data imputation methods are equated according to test equating methods, the location of missing data in the test forms and RMSE values according to missing data rates are reported in Table 3 and Figure 1, respectively.

Table 3

RMSE Values in Test Equating with Zero Imputation and Multiple Imputation Methods

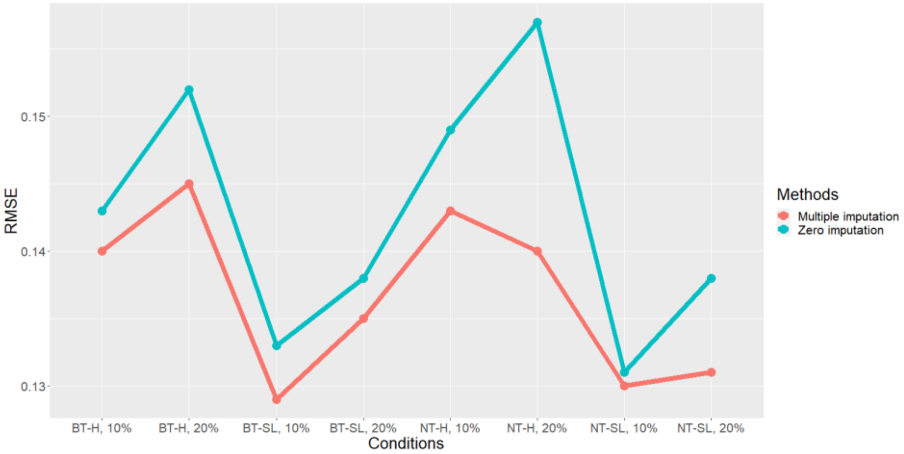
Missing data rate	Missing data imputation methods	Missing data location	Test equating method	Observed score RMSE
Full data set*			Haebara	0.141
			Stocking-Lord	0.130
10%	Zero imputation	New test	Haebara	0.149
			Stocking-Lord	0.131
		Both test	Haebara	0.143
			Stocking-Lord	0.133
	Multiple imputation	New test	Haebara	0.143
			Stocking-Lord	0.130
		Both test	Haebara	0.140
			Stocking-Lord	0.129
20%	Zero imputation	New test	Haebara	0.157
			Stocking-Lord	0.138
		Both test	Haebara	0.152
			Stocking-Lord	0.138
	Multiple imputation	New test	Haebara	0.140
			Stocking-Lord	0.131
		Both test	Haebara	0.145
			Stocking-Lord	0.135

* Taken as a reference value.

According to Table 3, the RMSE value was determined as 0.141 when the test forms that did not contain missing data at the beginning and had full data were equated according to the Haebara method, and the RMSE value was observed as 0.130 when they were equated according to the Stocking Lord method. These values are considered reference values.

Figure 1

RMSE Values in Test Equating with Zero Imputation and Multiple Imputation Methods



According to Figure 1, it is seen that the RMSE values are lower when the test forms obtained by using the multiple imputation method compared to the zero imputation method are equating under all conditions.

Change According to Zero Imputation and Characteristic Curve Transformation Methods

According to Table 3, if 10% missing data was only included in the new test, the RMSE value was determined as 0.149 for test forms with full data obtained by applying the zero imputation method to these missing data when equated according to the Haebara method, and 0.131 when equated according to the Stocking Lord method. In the case of 10% missing data that was included in both tests, the

RMSE value was observed as 0.143 for test forms with full data obtained by applying the zero imputation method to the missing data when equated according to the Haebara method, and 0.133 when equated according to the Stocking Lord method.

In the case of 20% missing data that was only included in the new test, the RMSE value was determined as 0.157 for test forms with full data obtained by applying the zero-imputation method to the missing data when equated according to the Haebara method, and 0.138 when equated according to the Stocking Lord method. In the case of 20% missing data that was included in both tests for test forms with full data obtained by applying the zero imputation method to the missing data, the RMSE value was found as 0.152 when equated according to the Haebara method, and 0.138 when equated according to the Stocking Lord method.

According to Figure 1, when the missing data rate increased, the performance of the zero imputation method decreased and it produced higher RMSE values. In general, it is seen that the missing data is found at the rate of 10% in the new test and the lowest RMSE value is obtained for the condition where the test equating is made according to the Stocking Lord method. In addition, it was determined that the missing data was found at the rate of 20% in the new test and the highest RMSE value was produced for the condition in which the test equating was made according to the Haebara method.

Change According to Multiple Imputation and Characteristic Curve Transformation Methods

According to Table 3, if 10% missing data was only included in the new test for test forms with full data obtained by applying the multiple imputation method to the missing data, the RMSE value was determined as 0.143 when equated according to the Haebara method, and 0.130 when equated according to the Stocking Lord method. In the case of 10% missing data that was included in both tests, the RMSE value was observed as 0.140 for test forms with full data obtained by applying the multiple imputation method to the missing data when equated according to the Haebara method, and 0.129 when equated according to the Stocking Lord method.

In the case of 20% missing data that was only included in the new test for test forms with full data obtained by applying the multiple imputation method to the missing data, the RMSE value was determined as 0.140 when equated according to the Haebara method, and 0.131 when equated according to the Stocking Lord method. When 20% missing data were included in both tests, the RMSE value was found as 0.145 for test forms with full data obtained by applying the multiple imputation method to the missing data when equated according to the Haebara method, and 0.135 when equated according to the Stocking Lord method.

According to Figure 1, when the missing data rate increases, the performance of the multiple imputation method decreases, and it produces higher RMSE values. It is seen that the missing data was found at the rate of 10% in both tests, and the lowest RMSE value was obtained for the condition where the test equating was made according to the Stocking Lord method. In addition, it was determined that the missing data was found at the rate of 20% in both tests, and the highest RMSE value was produced for the condition where the test equalization was made according to the Haebara method.

Discussion and Conclusion

In this research, when the test forms obtained by zero imputation and multiple imputation methods are equated according to characteristic curve transformation methods, one of the test equating methods based on IRT how the RMSE value changes according to different conditions (the rate of missing data, the location of missing data in the test forms) has been examined on the real data set. In the light of the findings obtained from the research, the impact of each missing data handling method on the test equating methods under different conditions was examined and discussed.

When the test forms obtained by applying the zero-imputation method were equated according to the characteristic curve transformation methods, the lowest RMSE value was obtained with 10% of missing data in both tests. In addition, it was observed that RMSE values increased when the missing data rate was 20%. When the test forms obtained by applying the multiple imputation method were equated according to the characteristic curve transformation methods, the lowest RMSE value was obtained with 10% of missing data in both tests, and it was obtained with 20% of missing data in the new test. These results indicate that lower RMSE values were generally obtained when missing data were present in both tests (new form, old form). These indicators support the results of the study conducted by Ertoprak (2017), which also covers the conditions in which the missing data was found in both tests, the new test and the joint test. According to the missing data rate condition discussed in the research, it was seen that the equating error generally increases as the amount of missing data increases. This result concurs with studies showing that the data sets handled under different conditions give more reliable results as the missing data rate decreases, RMSE and bias values decrease, and a closer parameter estimation can be made (Bayram, 2020; Finch, 2008; Zhu, 2014).

The multiple imputation method has been determined as the method that produces the lowest RMSE value among the methods of handling missing data discussed in the research. In addition, it was concluded that the multiple imputation method produced RMSE values closer to the full data set, which was considered the reference value in the research. This result coincides with the findings in the studies on the methods of handling missing data in the literature. It was emphasized that under the conditions discussed in the studies, the multiple imputation method came to the fore because it produced fewer error values (Bayram, 2020; Demir, 2013; Koçak, 2016; Zhu, 2014). Additionally, the results of the studies on methods of handling missing data and test equating methods together support this result (Ertoprak, 2017; Kim, 2015; Ngudgratoke, 2009; Shin, 2009).

The Stocking-Lord method, one of the characteristic curve transformation methods among the test equating methods examined in the research, produced both the lowest and the closest RMSE value to the full data set, which was considered as a reference value compared to the Haebara method. This result is consistent with the results of the study conducted by Karkee and Wright (2004), Kilmen (2010), Aksekioğlu (2017) and Mutluer (2021), which found that the Stocking-Lord method outperformed the Haebara method. However, in the study conducted by Lee and Ban (2010) using a random group design, they found that the Haebara method gave better results than the Stocking-Lord method. The reason for this can be explained by the difference in the selected test equating pattern.

Based on these results, in order to make an equation with the data set that is scored dichotomously and contains missing data, before determining the most appropriate techniques of handling the missing data, it is regarded as essential to examine the missing data rate and the location of the missing data in the test forms. As the missing data rate increases, the performance of the methods of dealing with missing data decreases and the RMSE values increase. According to the conditions discussed in the research, the multiple imputation method, one of the methods for dealing with missing data, and Stocking-Lord method, one of the test equating methods, came to the fore as less error-producing methods. However, it should also be noted that there is no single method that can be used in all conditions and gives the best results. This research is limited to a single data set as it has been obtained from the real data set. For this reason, it is recommended to conduct different studies to compare the results by replication. In addition, for further research, it can be suggested that this study should be conducted using different sample sizes, missing data mechanisms, methods of handling missing data, equating design, test equating methods, and/or evaluation criteria.

Declarations

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: Secondary data were used in this study. Therefore, ethical approval is not required.

References

Akbaş, U. (2014). *Farklı örneklem büyüklüklerinde ve kayıp veri örüntülerinde ölçeklerin psikometrik özelliklerinin kayıp veri baş etme teknikleri ile incelenmesi [Investigation of psychometric properties of scales with missing data techniques for different sample sizes and missing data patterns]* (Thesis No. 370326) [Doctoral dissertation, Ankara University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Aksekioğlu, B. (2017). *Madde tepki kuramına dayalı test eşitleme yöntemlerinin karşılaştırılması: PISA 2012 fen testi örneği [Comparison of test equating methods based on item response theory: PISA 2012 science test sample]* (Thesis No. 454879) [Master thesis, Akdeniz University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, *112*(4), 545-557. https://doi.org/10.1037/0021-843X.112.4.545

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508-600). American Council on Education.

Atalay Kabasakal, K. (2014). *Değişen madde fonksiyonunun test eşitlemeye etkisi [The effect of differential item functioning on test equating]* (Thesis No. 363206) [Doctoral dissertation, Hacettepe University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Aytekin Kazanç, E. (2019). *Şans başarısının test eşitlemeye etkisinin farklı eşitleme teknikleriyle araştırılması [Investigation of the effect of guessing on test equating with different equating methods]* (Thesis No. 584263) [Master thesis, Hacettepe University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Battauz, M. (2021). *Package 'equateIRT'*. https://cran.r-project.org/web/packages/equateIRT/equateIRT.pdf

Bayhan, A. (2018). *Farklı koşullardaki kayıp veri oranının iç tutarlığa etkisi [The effect of missing data rate on internal consistency within different conditions]* (Thesis No. 531022) [Master thesis, Hacettepe University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Bayram, İ. (2020). *Kayıp veri ile baş etme yöntemlerinin güvenirlik kestirimleri üzerine etkisi [Comparison of influence of the missing data handling methods on reliability estimation]* (Thesis No. 634087) [Master thesis, Hacettepe University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Büyüköztürk, Ş. (2011). *Sosyal bilimler için veri analizi el kitabı [Manual of data analysis for social sciences]*. Pegem A Publishing.

Chalmers, P., Pritikin, J., Robitzsch, A., Zoltak, M., Kim, K. H., Falk, C. F., Meade, A., Schneider, L., King, D., Liu, C. W., & Ogreten, O. (2021). *Package 'mirt'*. https://cran.r-project.org/web/packages/mirt/mirt.pdf

Cook, L. L., & Eignor, D. R. (1991). An NCME instructional module on IRT equating methods. *Educational Measurement: Issues and Practice, 10*(3), 37-45. http://www.edmeasurement.net/8225/Cook-1991-equating-methods.pdf

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

Demir, E. (2013). *Kayıp verilerin varlığında iki kategorili puanlanan maddelerden oluşan testlerin psikometrik özelliklerinin incelenmesi [Psychometric properties of tests composed of dichotomous items in the presence of missing data]* (Thesis No. 342477) [Doctoral dissertation, Ankara University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Demirus, K. B. (2015). *Ortak maddelerin değişen madde fonksiyonu gösterip göstermemesi durumunda test eşitlemeye etkisinin farklı yöntemlerle incelenmesi [The study of the effect of anchor items showing or not showing differantial item functioning to test equating using various methods]* (Thesis No. 399468) [Doctoral dissertation, Hacettepe University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Lawrence Erlbaum Associates Publishers.

Enders, C. K. (2004). The impact of missing data on sample reliability estimates: Implications for reliability reporting practices. *Educational and Psychological Measurement, 64*(3), 419-436. https://doi.org/10.1177/0013164403261050

Enders, C. K. (2010). *Applied missing data analysis.* The Guildford Press.

Ertoprak, D. G. (2017). *Kayıp verinin test eşitlemeye etkisinin incelenmesi [Investigating the effect of missing data on test equating]* (Thesis No. 470015) [Doctoral dissertation, Hacettepe University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement, 45*(3), 225-245. https://doi.org/10.1111/j.1745-3984.2008.00062.x

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*(3), 144-149. https://doi.org/10.4992/psycholres1954.22.144

Hambleton, R. K., & Jones, RW. (1993). 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.

Han, K. T. (2008). *Impact of item parameter drift on test equating and proficiency estimates* (UMI No. 3325324) [Doctoral dissertation, University of Massachusetts Amherst]. Available from ProQuest Dissertations and Theses Global database.

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6*(3), 195-240. https://doi.org/10.1207/s15324818ame0603_3

Hohensinn, C., & Kubinger, K.D. (2011). On the impact of missing values on the item fit and the model validness of the rasch model. *Psychological Test and Assessment Modeling*, *53*(3), 380-393. https://www.researchgate.net/publication/263655931

Işıkoğlu, M. A. (2017). *Kayıp veri ile baş etme yöntemlerinin ölçme değişmezliğine etkisi açısından karşılaştırılması [Comparison of influence of the missing data handling methods on measurement invariance]* (Thesis No. 484106) [Master thesis, Hacettepe University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement, 40*(8), 559-572. https://doi.org/10.1177/0146621616664046

Kan, A. (2011). Test eşitleme: OKS testlerinin istatistiksel eşitliğinin sınanması [Test equating: Checking statistical equivalance of OKS test edition]. *Education and Science, 36*(160), 38-51. http://eb.ted.org.tr/index.php/EB/article/view/310/258

Karagül, A. E. (2020). *Küçük örneklemelerde çok kategorili puanlanan maddelerden oluşan testlerde klasik test eşitleme yöntemlerinin karşılaştırılması [Comparison of classical test equating methods with polytomously scored tests and small samples]* (Thesis No. 610685) [Master thesis, Ankara University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

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

Karkee, T. B., & Wright, K. R. (2004, April 16). *Evaluation of linking methods for placing three-parameter logistic item parameter estimates onto a one-parameter scale* [Conference presentation]. Annual Meeting of the American Educational Research Association in San Diego, California. https://files.eric.ed.gov/fulltext/ED491663.pdf

Keller, L. A., & Keller, R. R. (2011). The long-term sustainability of different Item Response Theory scaling methods. *Educational and Psychological Measurement, 71*(2), 362-379. https://doi.org/10.1177/0013164410375111

Kilmen, S. (2010). *Madde Tepki Kuramı'na dayalı test eşitleme yöntemlerinden kestirilen eşitleme hatalarının örneklem büyüklüğü ve yetenek dağılımına göre karşılaştırılması [Comparison of equating errors estimated from test equation methods based on Item Response Theory according to the sample size and ability distribution]* (Thesis No. 279926) [Doctoral dissertation, Ankara University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Kim, J. S., & Hanson B. A. (2000). Test equating under the multiple-choice model. *Applied Psychological Measurement, 26*(3), 255-270. https://doi.org/10.1177/0146621602026003002

Kim, M. S. (2015). *Linking with planned missing data: Concurrent calibration with multiple imputation* [Doctoral dissertation, University of Kansas]. KU Scholar Works. https://kuscholarworks.ku.edu/handle/1808/20985

Koçak, D. (2016). *Kayıp veriyle baş etme yöntemlerinin madde tepki kuramı bir parametreli lojistik modelinde model veri uyumuna ve standart hataya etkisi [The effect of missing data tecniques in one parameter logistic model of item response theory on model fit and standard error]* (Thesis No. 456676) (Doctoral dissertation, Ankara University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). Springer.

Lee, W. C., & Ban, J. C. (2010). A comparison of IRT linking procedures. *Applied Measurement in Education, 23*(1), 23-48. https://doi.org/10.1080/08957340903423537

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Wiley Publishing.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*(3), 179-193. https://doi.org/10.1111/j.1745-3984.1980.tb00825.x

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, *14*(2), 139-160. https://doi.org/10.1111/j.1745-3984.1977.tb00033.x

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. Guilford Press.

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). TIMSS 2019 *international results in mathematics and science*. Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/international-results/

Mutluer, C. (2013). Y*ıl içinde farklı akademik personel ve lisansüstü eğitimi giriş sınavı (ALES) puanlarına ilişkin bir test eşitleme çalışması [A test equating study concerning to ALES (Academic Personnel and Postgraduate Education Entrance Exam) scores obtained at different times in a year]* (Thesis No. 336323) [Master thesis, Abant İzzet Baysal University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Mutluer, C. (2021). *Klasik Test Kuramı'na ve Madde Tepki Kuramı'na dayalı test eşitleme yöntemlerinin karşılaştırılması: Uluslararası öğrenci değerlendirme programı (PISA) 2012 matematik testi örneği [Comparison of test equating methods based on Classical Test Theory and Item Response Theory: International Student Assessment Program (PISA) 2012 mathematics test case]* (Thesis No. 658052) [Doctoral dissertation, Gazi University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Ngudgratoke, S. (2009). *An investigation of using collateral information to reduce equating biases of the post-stratification equating method* (UMI No. 3381312) [Doctoral dissertation, Michigan State University]. Available from ProQuest Dissertations and Theses Global database.

Öztemür, B. (2014). *Kayıp veri yöntemlerinin farklı değişkenler altında varyans analizi (t- testi, anova) parametreleri üzerine etkisinin incelenmesi [Examining the effect of missing data methods on variance analysis (t-Test, ANOVA) parameters under different variables]* (Thesis No. 357738) [Master thesis, Abant İzzet Baysal University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* John Wiley & Sons.

Ryan, J., & Brockmann, F. (2009). *A practitioner's introduction to equating with primers on Classical Test Theory and Item Response Theory.* CCSSO.

Shin, S. H. (2009). How to treat omitted responses in Rasch model-based equating. *Practical Assessment Research & Evaluation, 14*(1), 1-8. https://doi.org/10.7275/x9vv-xg85

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement, 42*(4), 309-330. https://doi.org/10.1111/j.1745-3984.2005.00018.x

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201-210. https://doi.org/10.1177/014662168300700208

Swaminathan, J., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 13-30). Academic.

Tamcı, P. (2018). *Kayıp veriyle başa çıkma yöntemlerinin değişen madde fonksiyonu üzerindeki etkisinin incelenmesi [Investigation of the impact of techniques of handling missing data on differential item functioning]* (Thesis No. 517260) [Master thesis, Hacettepe University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Tian, F. (2011). *A comparison of equating/linking using the Stocking-Lord method and concurrent calibration with mixed-format tests in the non-equivalent groups commonitem design under IRT* [Doctoral dissertation, Boston College]. Boston College Libraries. http://hdl.handle.net/2345/2370

Uysal, İ. (2014). *Madde Tepki Kuramı'na dayalı test eşitleme yöntemlerinin karma modeller üzerinde karşılaştırılması [Comparison of irt test equating methods for mixed format tests]* (Thesis No. 370226) [Master thesis, Abant İzzet Baysal University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Wolf, R. (2013). *Assessing the impact of characteristics of the test, common-items, and examinees on the preservation of equity properties in mixed-format test equating* [Doctoral dissertation, University of Pittsburgh]. D-Scholarship. http://d-scholarship.pitt.edu/id/eprint/20130

Wu, W., Jia, F., & Enders, C. (2015). A comparison of imputation strategies for ordinal missing data on likert scale variables. *Multivariate Behavioral Research, 50*(5), 484-503. https://doi.org/10.1080/00273171.2015.1022644

Yang, W. L. (1997). *The effects of content homogeneity and equating method on the accuracy of common-item test equating* (UMI No. 9839718) [Doctoral dissertation, Michigan State University]. Available from ProQuest Dissertations and Theses Global database.

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*(2), 125-145. https://doi.org/10.1177/014662168400800201

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2001). Effects of local item dependence on the validity of IRT item, test and ability statistics [Monograph]. *MCAT*, 3-30. https://eric.ed.gov/?id=ED462426

Zhu, X. (2014). Comparison of four methods for handing missing data in longitudinal data analysis through a simulation study. *Open Journal of Statistics, 4*(11), 933-944. http://dx.doi.org/10.4236/ojs.2014.411088

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2022; 13(2); 117-133

Reliability Generalization Meta-Analysis of Mathematics Anxiety Scale for Primary School Students

Ömer Şahin ASLAN * Salih GÖCEN ** Sedat ŞEN ***

Abstract
The purpose of this study was to estimate the mean value for the reliability coefficients reported by the studies using the Mathematics Anxiety Scale for Primary School Students (MASPSS) and to examine the sources of the variation of the reliability coefficients reported in each study. A reliability generalization meta-analysis study was conducted by combining the Cronbach's alpha values of 34 studies that met the inclusion criteria. The Cronbach's alpha values used in the studies were converted to the transformed coefficient values by applying the Bonett transformation, and the analyses were carried out under the random-effects model. The mean Cronbach's alpha value of the MASPSS across 34 studies was found to be .855 (95% CI: .841-.869), and this result was statistically significant (p<.01). According to the results, there was a lack of publication bias in this meta-analysis study. Moderator analyses were conducted to explain the possible sources of heterogeneity across the individual studies. Findings revealed that the Cronbach's alpha estimates did not show any statistical differences based on publication year, female percentage, publication type and research method variables. It was found that the sample type affected the estimation of Cronbach's alpha coefficient. In addition, suggestions were made for psychometric studies that would use the MASPSS in the future.

Keywords: mathematics anxiety, reliability generalization, meta-analysis

Introduction

The concept of anxiety has been included in different areas as an indispensable element of daily life for many communities in the world. Our formal schooling period, which covers a certain part of many people's lives, is one of these areas. It is known that some people are more anxious during these periods. Although the concept of anxiety is attributed to different meanings by people, the concept that is most confused is the fear (Manav, 2011). So what does the concept of anxiety really mean? While the anxiety was previously accepted as a biological concept, it has entered the psychological literature with its definition by Freud as a function of the ego (Manav, 2011). Anxiety is a reflection of the fear of any danger and is defined as a state of uneasiness or irrational fear that manifests itself in people and differs from fear as it is objectless (Budak, 2000, p. 437).

Throughout the formal education life, mathematics has been one of the most encountered fields, because it is universal and penetrates many areas of life. During their formal education years, many people may have encountered people with anxiety in math classes. This state of anxiety, which we see in mathematics lessons, is called mathematics anxiety. Different definitions of this concept have been made in previous studies (Newstead, 1998; Richardson & Suinn, 1972; Tobias & Weissbrod, 1980). Richardson and Suinn (1972) defined mathematics anxiety as "a feeling of tension and anxiety that interferes with the manipulation of numbers and solving math problems in a variety of life and academic situations" (p. 551).

Students' mathematics anxiety is among the important factors affecting math achievement (Bozkurt, 2012; Dursun & Bindak, 2011; İlhan & Öner Sünkür, 2012; Kutluca et al., 2015; Kuzu, 2021; Mutlu et

* Graduate student., Harran University, Faculty of Education, Şanlıurfa-Turkey, omersahinaslan63@gmail.com, ORCID ID: 0000-0003-1330-1251
** Graduate student., Harran University, Faculty of Education, Şanlıurfa-Turkey, salihgocen48@gmail.com, ORCID ID: 0000-0001-6594-4967
*** Assoc. Prof., Harran University, Faculty of Education, Şanlıurfa-Turkey, sedatsen@harran.edu.tr, ORCID ID: 0000-0001-6962-4960

To cite this article:
Aslan, Ö.Ş., Göcen, S., & Şen, S. (2022). Reliability generalization meta-analysis of Mathematics Anxiety Scale for primary school students, *13*(2), 117-133. https://doi.org/10.21031/epod.1119308

Received: 21.05.2022
Accepted: 26.06.2022

al., 2017; Şad et al., 2016). Previous studies have stated that as the mathematics anxiety level increases, students' math achievement decreases and it affects learning negatively (Bozkurt, 2012; Dursun & Bindak, 2011; İnci Kuzu, 2021; Kutluca et al., 2015; Mutlu et al., 2017; Tooke & Leonard, 1998; Wadlington & Wadlington, 2008). A recent meta-analysis study examining the relationship between mathematics anxiety and math achievement has concluded that students with higher mathematics anxiety levels had lower math achievement levels (Şad et al., 2016). In a different study, a negative and high-level correlation was also found between the students' readiness for mathematics lesson and their mathematics anxiety (Ergenç, 2011). The concept of mathematics anxiety that develops in students can be caused by four main reasons: teachers, students themselves, their families and friends (Alkan, 2011). It has also been stated that mathematics anxiety can occur due to more than one reason (Alkan, 2011). According to this information, it can be interpreted that anxiety can be caused by more than one variable and different variables can affect anxiety at the same time. For instance, when we consider the studies examining the effect of the gender variable on mathematics anxiety, statistically significant differences were observed in some studies (Arı et al, 2010; Şahin, 2008), while in others, no statistically significant difference was observed (Dede & Dursun, 2008; Dursun & Bindak, 2011; Gündüz Çetin, 2020; Kandal & Baş, 2021; Kutluca et al., 2015; Mutlu et al., 2017; Şimşek et al., 2017; Tan, 2015; Taşdemir, 2015; Yetgin, 2017). When we look at the studies examining the effects of the grade level variable on mathematics anxiety, some studies have stated that grade level has a significant effect on mathematics anxiety (Dursun & Bindak, 2011; Taşdemir, 2015), while some others have stated that grade level has no significant effect (Bozkurt, 2012; Dede & Dursun, 2008; Kandal & Baş, 2021). Some studies showed that families could also be the cause of mathematics anxiety in students (Kesici, 2018b; Maloney et al., 2015). According to Yetgin (2017), students who received private lessons or study training center support had less mathematics anxiety than students who did not receive support.

Many scale development and adaptation studies have been conducted to determine students' mathematics anxiety levels and to explain which variables would cause anxiety (Alexander & Martray, 1989; Bai et al., 2009; Fennema & Sherman, 1976; Hopko et al., 2003; Hunt et al., 2011; Ikegulu, 1998; Plake & Parker, 1982; Richardson & Suinn, 1972; Sandman, 1980; Suinn & Edwards, 1982; Suinn et al., 1988). Some of these scales were adapted into Turkish (Akçakın et al., 2015; Baloğlu & Balgalmış, 2010; Baloğlu, 2005).

In Turkey, some scale development and adaptation studies were also carried out in order to measure the mathematics anxiety of students, teachers and parents (Akçakın et al., 2015; Akın et al., 2011; Baloğlu, 2005; Baloğlu & Balgamış, 2010; Bindak, 2005; Mutlu & Söylemez, 2018; Mutlu et al., 2018; Peker, 2006; Sarı, 2014; Şan & Akdağ, 2017; Üldaş, 2005; Yıldırım & Gürbüz, 2017). For instance, Akçakın et al. (2015) adapted the Mathematics Anxiety Scale (MAS; Bai et al., 2009) into Turkish. Similarly, Akın et al. (2011) also adapted the Revised Mathematics Anxiety Rating Scale (RMARS; Plake and Parker, 1982) into Turkish. In another study, Baloğlu and Balgalmış (2010) adapted the Mathematics Anxiety Rating Scale Primary Education Form (MARS-E; Suinn et al., 1988) scale into Turkish. The Mathematics Exam Anxiety Scale (MEAS) was developed by Şan and Akdağ (2017) to determine the mathematics test anxiety of middle school students. The Mathematics Anxiety Scale for Primary School Students, developed by Bindak (2005), was also developed to determine the mathematics anxiety levels of primary school students. When we look at the scale adaptation and development studies to determine students' mathematics anxiety levels, it was seen that the scale developed by Bindak (2005) was used in more studies compared to other Turkish scales.

In this meta-analysis study, Mathematics Anxiety Scale for Primary School Students (MASPSS), developed by Bindak (2005), was examined as it is the most cited scale among Turkish scales measuring mathematics anxiety. Bindak (2005) developed a 10-item mathematics anxiety scale as a result of his analysis. With this scale developed, it was aimed to determine the levels of mathematics anxiety in primary school students (Bindak, 2005). In the first draft of the scale, there were four items to obtain students' personal information and 16 items in a 5-point Likert type format that can express mathematics anxiety. Each student responding to the scale selects one of the five categories (*always, most of the time, sometimes, almost never* and *never*). Positive items for anxiety in the scale were scored as 5-4-3-2-1 and negative items for anxiety were scored as 1-2-3-4-5. Thus, an anxiety score is obtained for the whole scale. Higher scores obtained from the scale indicate higher mathematics anxiety. Based on the item-total correlation (two items) and factor analysis (four items) with the preliminary scale, six items were eliminated and a final scale of 10 items was constructed in Bindak (2005). In this case, the anxiety score formed by the scale ranges from 10 to 50 points.

As a result of the factor analysis, the explained total variance of the scale with 10 items was 51.7% (Bindak, 2005). In addition, Bindak (2005) reported that the internal consistency coefficient (i.e., Cronbach's alpha) of the scale was .84. After this original scale development study, the scale (MASPSS) has been used in several studies. It has been observed that studies using this scale have reported different Cronbach's alpha values. Using the scale developed by Bindak (2005), several studies have been carried out in different cities and districts, at different grade levels, in different years and with different research methods (Aydın & Keskin, 2017; Berber, 2021; Dede & Dursun, 2008; Küçük, 2019; Yurt & Kurnaz, 2015). However, while some studies found the Cronbach's alpha coefficient to be around .70 (Erdik, 2018; Şimşek et al., 2017; Tuncer & Yılmaz, 2016) some studies had values of .80 and above (Akgül & Nuhoğlu, 2020; Aydın & Keskin, 2017; Küçük, 2019; Şahin, 2018). For this reason, a number of different Cronbach's alpha coefficient has been reported in these studies. Some studies also reported the reliability value obtained from previous studies that used that scale before, rather than calculating based on their own sample. In this case, it is assumed that reliability is a fixed and stable characteristic of the scale itself, not the measurement results. This is called reliability induction (Vacha-Haase, et al., 2000). Since the reliability values reported in previous studies vary, it may not be appropriate to generalize a reliability value over the studies without necessary analyses. As a result, the reliability of the scores should be confirmed due to the variability in reported reliability values with its widespread use in different environments and populations. A reliability generalization study was needed to learn the general condition of the reliability coefficient obtained from the MASPSS and to help researchers who may want to induce a reliability value. In this study, a "reliability generalization (RG)" (Vacha-Haase, 1998), which is a meta-analytic technique, was conducted for the review, integration and analysis of research results. Our aim in this study is to obtain an overall reliability coefficient inference of the MASPSS developed by Bindak (2005) and to examine how the reliability coefficients change between the uses of the scales in different samples. More precisely, the aims of this study are: a) to examine the generalizability of reliability estimates in studies using the MASPSS and to investigate the variables that may explain this heterogeneity if heterogeneity is found in the estimates of this parameter. In order to achieve these aims, studies reporting alpha coefficients were determined using the aforementioned scale, and subgroup analyses were carried out by examining some variables as well.

Method

Research Design

Glass (1976) defined meta-analysis as an analysis method for summarizing the results obtained from individual studies as a single result. In other words, meta-analysis is a type of quantitative study that combines the findings of more than one study and presents it as a single finding (Şen & Yıldırım, 2020). The meta-analysis study conducted on the reliability values of a specified scale is known as the reliability generalization meta-analysis (Vacha-Haase, 1998). Reliability generalization is known as a meta-analysis study that investigates the reliability values of the scores obtained from the scales and helps to determine what causes measurement error. This meta-analysis study for the reliability generalization of the primary school mathematics anxiety scale was presented following the REGEMA guidelines (Sánchez-Meca et al., 2021).

Data Collection

For the research data used in the study, National Thesis Center of the Council of Higher Education, Web of Science, and ERIC databases and Google Scholar search engine were scanned extensively. While searching the Turkish terms "matematik kaygı ölçeği" and "matematik kaygısı" and combinations of the English equivalents of these terms "Math Anxiety Scale", "Math Anxiety", "Mathematics Anxiety

Scale" and "Mathematics Anxiety" were searched throughout the text. In addition, via the "cited by" option in Google Scholar, references to the primary school mathematics anxiety scale (Bindak, 2005) were also reviewed in this study.

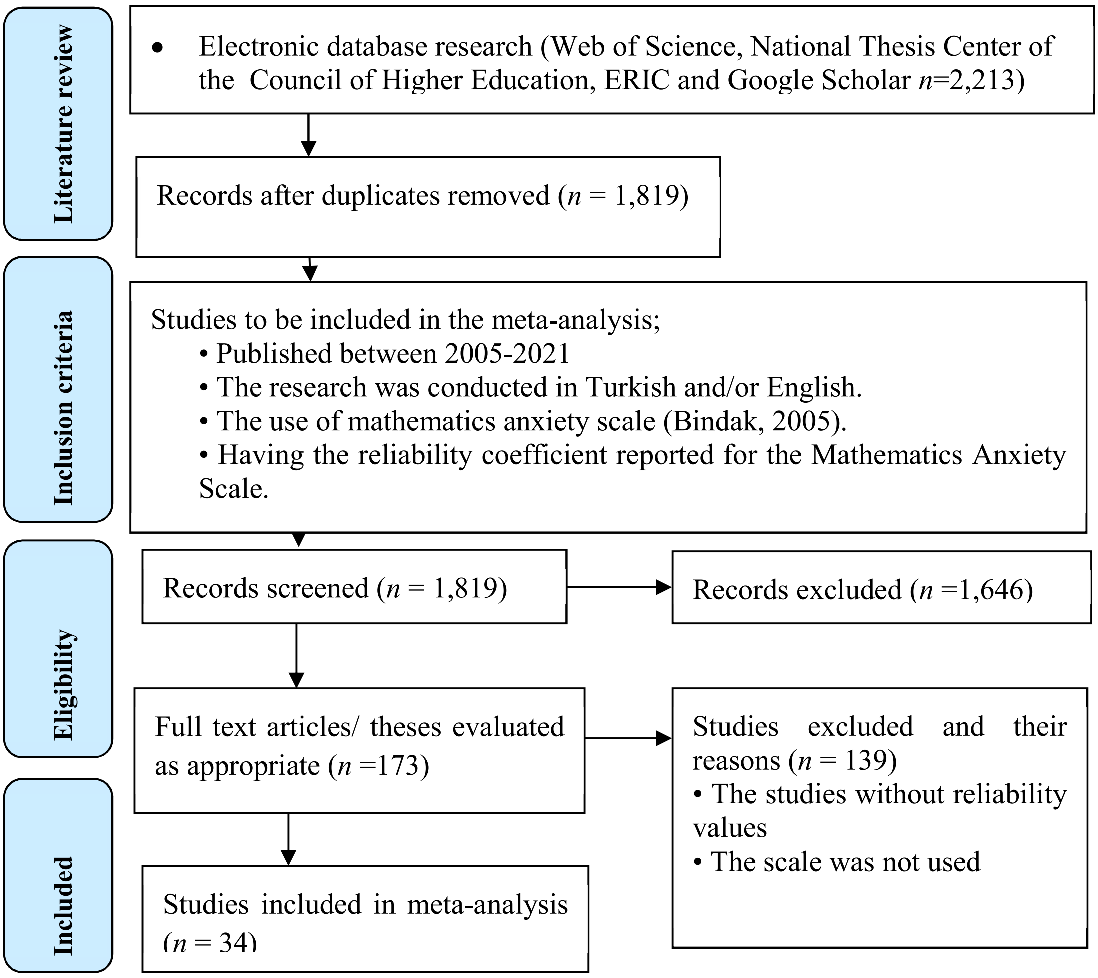
Inclusion/Exclusion Criteria

The following criteria were considered in determining the individual studies included in this study;

1) The articles must be written in Turkish and/or English.
2) The Mathematics Anxiety Scale for Primary School Students (Bindak, 2005) must be used.
3) Since the Mathematics Anxiety Scale for Primary School Students (Bindak, 2005) used in this study was published in 2005, the studies to be included can be either published or unpublished studies after 2005 until December 2021.
4) Having a reliability coefficient (e.g., Cronbach's alpha) reported on the sample in the study for the Mathematics Anxiety Scale for Primary School Students used in the studies.

The criteria listed above have been used in this study for selecting the possible studies. The inclusion/exclusion process is illustrated by the PRISMA flowchart (Figure 1). The number of studies found in searches was 2,213. Of these studies, 2,179 were excluded from the study due to the reasons such as being a duplicate study, not including the reliability (Cronbach's alpha) coefficient. After excluding the studies that do not meet the inclusion criteria, 34 studies remained.

Figure 1
Prisma Flow Chart: Studies Included in the Research



Coding of Data

In addition to the reliability coefficient and the number of items reported by the individual studies, the year they were published, the types of publications, the sample sizes, the school level in which the study was conducted, the percentage of female participants, the mean score and the standard deviation of MASPSS, the research method used in the studies were saved in an Excel file and coded accordingly. Information on the recorded variables is summarized in Table 1.

Table 1
Coding Method of Studies Included in the Meta-Analysis

Variable	Type	Coding Method
ID	Categorical	A unique number assigned to each study.
Year of publication	Continuous	Year of publication or report (for unpublished studies)
Publication type	Categorical	0= Thesis, 1= Article
Sample size	Continuous	The sample sizes specified in the studies were recorded.
Alpha	Continuous	The Cronbach's alpha values presented in the studies were recorded.
Number of items	Continuous	The number of scale items presented in the studies was recorded.
The average score	Continuous	The average anxiety scores presented in the studies were recorded.
Standard deviation	Continuous	The standard deviation scores of the means presented in the studies were recorded.
Sample type	Categorical	0= Primary School, 1=Middle School, 2= High School
Research method	Categorical	0= Experimental, 1= Non-experimental studies

Two researchers independently coded the data from individual studies. The inter-rater reliability was examined using the agreement index, which is a relatively simple way of checking the inter-rater reliability (Şen & Yıldırım, 2020). The percentage of agreement between the coders was calculated as 95%. Inconsistencies in coding were discussed and corrected by consensus. Then, the final data file created was transferred to the R software environment (R Core Development Team, 2021) for the statistical analyses.

Effect Size Calculation and Statistical Analyses

The focus of this study was the reliability generalization analysis used to estimate the mean reliability coefficient of the Mathematics Anxiety Scale for Primary School Students. An average value was calculated over the reliability coefficients reported in individual studies that used this scale. Since Cronbach's alpha value was predominantly used in these individual studies, only alpha coefficient was considered in the current study. While conducting reliability generalization studies, it is necessary to consider the transformation and weighting the alpha coefficients (Şen, 2021). Since the typical Cronbach's alpha values appeared to be skewed (Semma et al., 2019), Bonett's transformation formula (Bonett, 2002) was used to normalize the sample distributions and stabilize the variance. The calculation of the mean effect size in the meta-analysis literature is carried out with the either a fixed-effect model or the random-effects model (Borenstein et al., 2009; Şen & Yıldırım, 2020). It would be a more accurate approach to use the random-effects model for the studies conducted in social sciences (Borenstein et al., 2009). The random-effects model is needed in cases where results are desired to be generalized to the population (Schmid et al., 2021). In this respect, the mean effect size was calculated using the random-effects model in this study.

Whether there was heterogeneity among the studies included in the meta-analysis was examined by calculating Cochran's Q-test and I^2 value (Higgins & Thompson, 2002). A significant Q-statistic and an I^2 value of more than 75% can be taken as the evidence of the heterogeneity (Higgins et al., 2003). In cases where heterogeneity was detected, the relationship between Cronbach's alpha values and moderator variables would be examined using metaregression for continuous variables and weighted analysis of variance (analog to the ANOVA) models for categorical variables. All of the analyses in this study were carried out using the metafor (Viechtbauer, 2010) package in the R software environment.

Publication Bias Analyses

Publication bias term is used to describe that statistically significant results are more likely to be presented and published than nonsignificant results (Petitti, 2000). Since researchers generally tend to publish large effect sizes rather than small effect sizes, including only studies with large effect sizes in the meta-analyses raises a problem referred to as publication bias (Göçen & Şen, 2021). Publication bias is seen as a possible threat for meta-analysis studies (Rothstein et al., 2005). Moreover, publication bias, which is a widespread problem, can skew the effect size to be estimated (Thornton & Lee, 2000), and this might distort results of meta-analysis (Yumuşak & Korkmaz, 2021). In this study, publication bias was assessed using a funnel plot, Rosenthal's (1979) fail safe *N* and two statistical tests based on rank correlations (Begg & Mazumdar, 1994) and Egger's (Egger et al., 1997) regression method.

Results

Characteristics of Individual Studies

This reliability generalization meta-analysis includes 34 individual studies conducted between 2005 and 2021 that have reported reliability coefficients based on their own sample. Seventy-four percent of the studies were published (N =25) and 26% were unpublished (N =9). The total population in individual studies consisted of 10,855 individuals. Sixteen percent of the studies were carried out at primary school level, 75% at middle school level and 9% at high school level. Some of the studies using this scale at different school levels have conducted confirmatory factor analysis (Erdik, 2018; Gündüz Çetin, 2020; Yetgin, 2017), however, most of them have used the scale without any validation analyses. The Cronbach's alpha coefficient as a reliability index in the studies was reported between .737 and .920. In addition, 79% of the studies were conducted with experimental design and 21% with non-experimental design. Summary information about the studies is shown in Table 2.

Table 2

Information about the Studies

ID	Author(Year)	Publication Type	*N*	*α*	Sample Level	Female (%)	Publication Year	Method
1	Bindak (2005)	Published	122	.840	MS		2005	Non-experimental
2	Dede and Dursun (2008)	Published	204	.800	MS	42.0	2008	Non-experimental
3	Dursun and Bindak (2011)	Published	266	.888	MS	45.5	2011	Non-experimental
4	Küçük (2019)	Unpublished	52	.870	MS		2019	Experimental
5	Yurt and Kurnaz (2015)	Published	260	.800	MS	51.5	2015	Non-experimental
6	Aydın and Keskin (2017)	Published	619	.860	MS	50.4	2017	Non-experimental
7	Kutluca et al. (2015)	Published	158	.800	MS	49.4	2015	Non-experimental
8	Kesici (2018a)	Published	463	.884	HS		2018	Non-experimental
9	Çoruk and Çakır (2017)	Published	31	.880	PS	59.0	2017	Experimental
10	Duran et al. (2017)	Published	51	.820	MS		2017	Non-experimental
11	Kandal and Baş (2021)	Published	124	.840	MS	49.2	2021	Non-experimental
12	Şimsek et al. (2017)	Published	437	.780		44.1	2017	Non-experimental
13	İlhan and Öner Sünkür (2012)	Published	201	.830	MS	50.7	2012	Non-experimental
14	Doruk et al. (2016)	Published	246	.870	MS	54.8	2016	Non-experimental
15	Tuncer and Şimşek (2019)	Published	72	.840	MS		2019	Experimental
16	Akgül and Nuhoğlu (2020)	Published	121	.910	PS	37.2	2020	Non-experimental
17	Yetgin (2017)	Unpublished	860	.910	HS	37.0	2017	Non-experimental
18	Gündüz Çetin (2020)	Unpublished	555	.890	HS	49.4	2020	Non-experimental
19	Erdik (2018)	Published	1563	.737	MS	51.4	2018	Non-experimental
20	Baklacı (2017)	Unpublished	204	.850	MS	42.2	2017	Non-experimental
21	Kesici (2018b)	Published	132	.879	MS		2018	Non-experimental
22	İlhan and Öner Sünkür (2013)	Published	348	.860	MS	49.7	2013	Non-experimental
23	Berber (2021)	Unpublished	40	.920	MS	57.5	2021	Experimental
24	Kuzu and Çalışkan (2018)	Published	375	.876		78.4	2018	Non-experimental
25	Tuncer and Yılmaz (2016)	Published	225	.795	MS	48.0	2016	Non-experimental
26	Taşdemir (2015)	Published	280	.850	MS	51.4	2015	Non-experimental
27	Ergenç (2011)	Unpublished	526	.890	MS	49.8	2011	Non-experimental
28	Çağırgan and Soytürk (2021)	Published	568	.865	MS	51.1	2021	Non-experimental
29	Gevrek (2009)	Unpublished	932	.800	MS	50.9	2009	Non-experimental
30	Tabakçı (2018)	Unpublished	415	.840	PS	52.0	2018	Non-experimental
31	Borlat (2018)	Unpublished	18	.860	PS	38.9	2018	Experimental
32	Şahin (2018)	Published	30	.830	PS		2018	Experimental
33	Birgin et al. (2010)	Published	220	.910	MS	51.4	2010	Non-experimental
34	Tok (2013)	Published	137	.860	MS	52.7	2013	Experimental

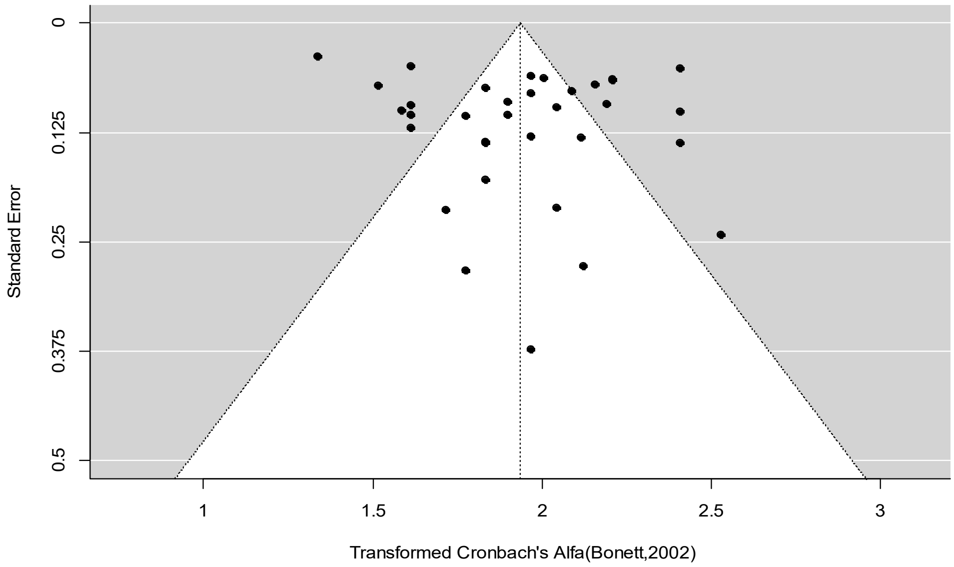
Notes. N=sample size, α=Cronbach's alpha, PS=Primary school, MS=Middle school, HS=High school. All studies had 10 items on the scale except that Tok (2013) had 9.

Results of Publication Bias

In the present study, the possibility of publication bias was investigated using the funnel plot, Rosenthal's (1979) fail safe *N* value and two statistical tests: Begg and Mazumdar (1994) rank correlations, and Egger's (Egger et al., 1997) linear regression method. An asymmetrical shape observed in the funnel plot indicates a possible publication bias (Borenstein et al., 2009). As seen in the funnel plot presented in Figure 2, Cronbach's alpha values of the studies appear to be symmetrically distributed according to the mean transformed alpha value. Therefore, it was found that the present study did not have publication bias. However, since it cannot be said that an asymmetrical funnel plot is formed as a result of publication bias, care should be taken while interpreting the chart (Üstün & Eryılmaz, 2014). When the publication bias was examined according to Rosenthal's classical fail safe *N* method, 7081 studies were required to turn the mean effect size value into statistically non-significant (p>.05) situation. If the Rosenthal's fail safe *N* value is NR >5k+10 (180 for this study), the possibility of publication bias is low (Şen & Yıldırım, 2020). Kendall's tau b statistic was observed to be nonsignificant (Tau b = -0.05; $p_{two-tails}$=.6565) according to Begg and Mazumdar's rank correlations. Finally, it was observed that the t-value was not statistically significant in Egger's linear regression test ($t_{(32)}$=0.7792, p =.4416). These findings showed that there was no indication of publication bias.

Figure 2

Funnel Plot Examining the Relationship between Transformed Alpha (Bonett, 2002) and Standard Error



Mean reliability

The mean value of raw reliability coefficient values reported in 34 studies, without weighting, was .851 (SD=0.04, Median=.86). A mean of .851 indicates good internal consistency for the mathematics anxiety scale among studies. A stem and leaf plot of the raw reliability coefficients is presented in Figure 3.

Figure 3

Distribution of Raw Alpha Values

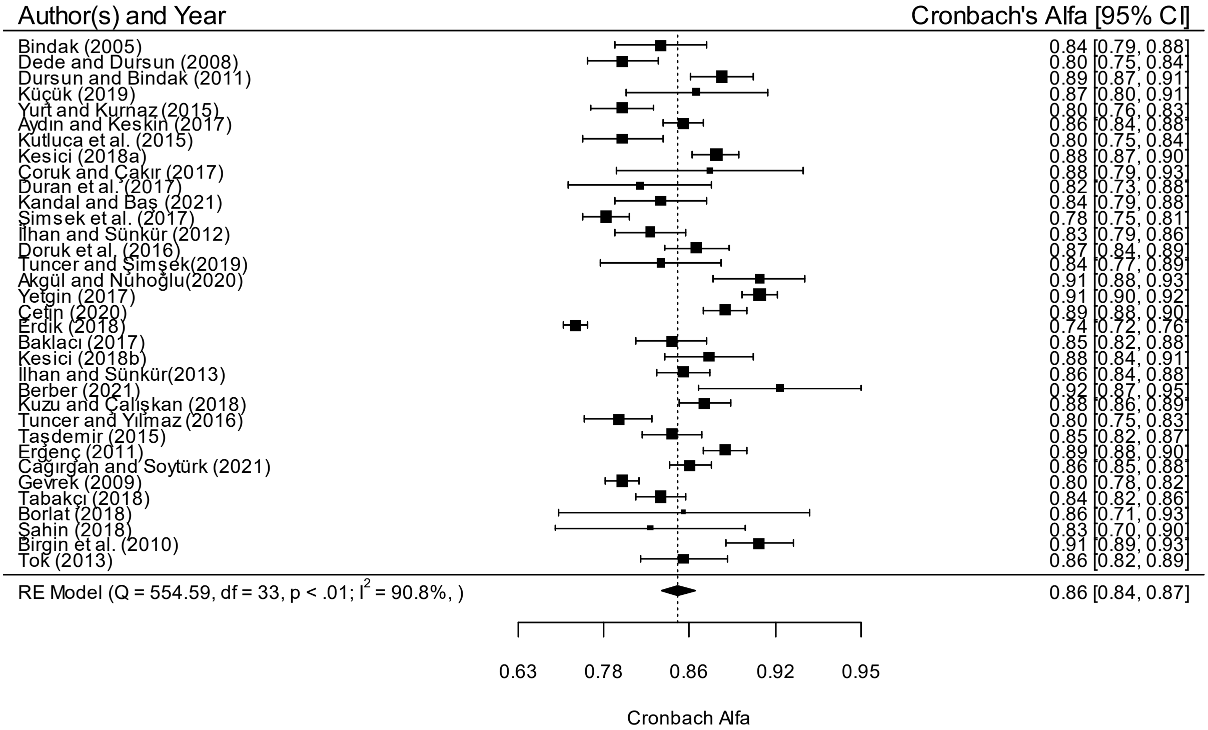
.73 | 7
.78 | 0
.79 | 5
.80 | 0000
.82 | 0
.83 | 00
.84 | 0000
.85 | 00
.86 | 00005
.87 | 0069
.88 | 048
.89 | 00
.91 | 000
.92 | 0

As can be seen in Figure 3, the reliability coefficient values reported in studies vary between .737 and .920. According to these values, the reliability coefficients appear to be at a sufficient level (>.70) in all samples.

The reliability coefficients are known to have a skewed distribution and cannot be directly used in reliability generalization studies. Thus, the reliability coefficient values of thirty-four studies were transformed using the Bonett (2002) transformation method in order to normalize the distribution of the coefficients and stabilize the variance. The mean reliability coefficient was obtained by analyzing them with the random-effects model. The pooled reliability coefficient value in this study was found to be .855 (95% CI: .841-.869), which was statistically significant (p<.001). This mean value was close to the first published reliability coefficient value of the scale (Bindak, 2005). Q-test value in this study was also statistically significant ($Q_{(33)}$ =554,588; p<.001). In addition, I^2 value of 90.80 showed that the heterogeneity among the studies was high. The variability in the reliability coefficients can also be observed in the forest plot (see Figure 4).

Figure 4

Forest Plot of Studies



Results of the Moderator Analyses

Considering the heterogeneity that emerged in this meta-analysis study, it is important to identify the possible sources of this heterogeneity. Therefore, the variables of publication type and publication year, sample level, female percentage, and research method given in Table 2 were determined as moderator variables. Three of the variables were categorical (type of publication, sample level and research method), and the remaining variables were continuous (year of publication and female percentage). Descriptive statistics of moderator variables are presented in Table 3. Categorical variables were analyzed with analog to the ANOVA approach and continuous variables were analyzed with meta-regression method to reveal whether there was a relationship between alpha values and the moderator variables.

Table 3
Characteristics of the moderators

Variable	*k*	%
Publication Type (*k*= 34)		
Published	25	73.53
Unpublished	9	26.47
Sample Level (*k* =32)		
Primary School	5	15.63
Middle School	24	75
High school	3	9.37
Research Method (*k* =34)		
Experimental	7	20.58
Non-experimental	27	79.42
	M	*S*
Year of publication	2016	3.95
Female percentage	49.84	7.77

M = Mean, *S* = Standard Deviation

The results of the analog to the ANOVA analyses of three moderators are presented in Table 4. As shown in Table 4, no statistically significant difference was found between the sub-categories of publication type and research method variables (p>.05). However, a statistically significant difference was found among the subcategories of the sample level variable (p<.05). While the reliability coefficient was found to be higher at high school level (.8958) than primary school level (.8694), the lowest value was found at middle school level (.8483) (See Table 4).

Table 4
Mixed Effects Analog to the ANOVA Results

Variable	**Category**	***k***	***α***	***95% CI*** Lower Limit	Upper Limit	Q_B	***df***	***p***
Sample level	Primary school	5	0.8694	0.8266	0.9016			
	Middle school	24	0.8483	0.8308	0.8642	6.3690	2	**0.0414**
	High school	3	0.8958	0.8784	0.9108			
Publication type	Published	25	0.8490	0.8310	0.8652	2.3345	1	0.1265
	Unpublished	9	0.8731	0.8462	0.8954			
Research method	Experimental	7	0.8684	0.8286	0.8988	0.5457	1	0.4601
	Non-experimental	27	0.8533	0.8366	0.8684			

Notes. k = Number of studies, CI = Confidence interval.

Two continuous moderators, the year of publication and the female percentage, of the reliability coefficient were analyzed with meta-regression. Meta-regression results are given in Table 5. As can be seen in Table 5, none of the continuous variables were statistically significant predictors of the reliability coefficient (p>.05).

Table 5
Meta-Regression Results According to Moderator Variables

Moderator	*k*	b_j	*SE*	*p*	Q_E
Year	34	-21.6115	26,1454	0.3678	552.3145**
Female %	25	1.9227	0.4007	0.9754	516.9344

k=number of studies, b_j =Unstandardized regression coefficient, *SE* =Standard error, Q_E =Heterogeneity statistics, ** $p<.001$.

Discussion and Conclusion

The aim of the present reliability generalization meta-analysis study was to obtain the general reliability for the Mathematics Anxiety Scale for Primary School Students and to examine the moderator variables that would reveal the variability between studies. For this purpose, individual studies using the Mathematics Anxiety Scale for Primary School Students and reporting the alpha coefficient of the sample were examined. The pooled reliability coefficient from 34 studies was .855. Based on this value, the overall estimate of Cronbach's alpha can be said to be within reasonable limits (>.70) for exploratory research (Clark & Watson, 1995; DeVellis, 1991; Nunnally & Bernstein, 1994).

Heterogeneity was observed between studies (I^2 = 90.80). This shows that it would not be appropriate to generalize the reliability coefficients of the Mathematics Anxiety Scale for Primary School Students since they vary in different samples. Therefore, it is not recommended for researchers who will use this scale to apply reliability induction.

The moderator variables that could be the source of the resulting heterogeneity in the reliability coefficients were examined. Three categorical (publication type, sample level and research method) and two continuous (year of publication and female percentage) variables were examined. It was concluded that the mean estimate of Cronbach's alpha coefficient did not show statistical differences according to the subcategories of publication type and research method and there was no statistically significant relationship between alpha coefficient and two continuous moderators: publication year and female percentage. It was concluded that only the sample level variable statistically affected the estimates of Cronbach's alpha coefficient. When Cronbach's alpha reliability coefficients were examined for the sample level, it was observed that the highest estimate was observed in the studies applied at the high school level. Although the name of the scale is Mathematics Anxiety Scale for Primary School Students, this scale was applied to high school students and it was observed that higher reliability coefficients were obtained. One of the reasons for obtaining a higher reliability coefficient in high school students is that these students are older and may have a better understanding of what the items in the scale mean.

Reporting the reliability findings of the sample in the studies conducted is very important to increase the validity, generalization and quality of the results (Wilkinson, 1999). Despite this, studies that did not report the reliability coefficient were encountered during the search process of this study. In some of these studies, reliability coefficients were not included, while in others, the reliability coefficient reported in the original article (Bindak, 2005) was reported. Assuming that reliability is a fixed and unchanging property of the scale itself, not the results of the measurement, is called reliability induction (Vacha-Haase et al., 2000). It is not appropriate to apply reliability induction except in special cases where it can be applied (Crocker & Algina, 1986; Vacha-Haase et al., 2000).

In the meta-analysis study, only studies reporting the widely used Cronbach's alpha reliability coefficient for the Mathematics Anxiety Scale for Primary School Students were included. However, it is known that Cronbach's alpha coefficient has some unrealistic assumptions (McNeish, 2018). Although Cronbach's alpha is widely used, there are different reliability methods for scales: test-retest, parallel (equivalent) tests and split-half methods. Some studies showed that the coefficients of Omega (McDonald, 1970), H-coefficient (Hancock & Mueller, 2001), maximal reliability (Hancock & Mueller, 2001), and greatest lower bound" (Jackson & Agunwamba, 1977) may be better option than alpha in terms of examining the reliability of the measure produced by a scale in different situations. In this context, different reliability coefficients could be used to give better results for the study. Therefore, it would be useful to report different reliability coefficients in future studies to obtain better results. The use of only alpha coefficients in this study can be considered a limitation.

In the reliability generalization study, care was taken to include all studies conducted with the Mathematics Anxiety Scale for Primary School Students. The obtained Cronbach's alpha values were limited to the scope of the literature review. Therefore, it is possible that this reliability generalization study may have not included all studies using the Mathematics Anxiety Scale for Primary School Students. In addition, the inability to publish articles with low reliability values may have led to the underrepresentation of studies with lower reliability values. Besides, the fact that some studies do not report the sample level and female student percentages can be another limitation. Since the reliability coefficients may differ in each sample, it is important for researchers who use any scale in their own study to present detailed demographic and descriptive information about the sample from which this value was obtained, while reporting their reliability findings.

Considering the effect of mathematics anxiety on children in daily life, it is thought that the results obtained from current study would be useful for the researchers and practitioners. A properly performed reliability generalization meta-analysis can provide more precise estimates of score reliability. The findings of the current study would be useful for researchers who want to study mathematics anxiety and make informed decisions.

As a result, it has been observed that this study and studies using the Mathematics Anxiety Scale for Primary School Students produced Cronbach's alpha coefficients at an acceptable level. In addition, the scale provides reliable results at different sample levels. Results of the study revealed that the reliability coefficients produced from the scale used differ in terms of the sample level variable. We also believe that it would be useful to make some suggestions based on the experiences we have gained as a result of current study. It was observed that the reliability coefficients used in the study were not reported in all individual studies. Therefore, there is a need to strengthen reliability reporting studies. We think that researchers should be more careful when reporting reliability coefficient and characteristic information such as age, gender, ethnicity, and sample level. It would be beneficial to use not only Cronbach's alpha values but also consider other reliability indicators (e.g. Omega, composite reliability) in future research studies.

Declarations

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: Reliability generalization meta-analysis using secondary data was carried out in the study. Therefore, ethical approval is not required.

Studies included in the current reliability generalization meta-analysis are marked with an asterisk (*).

References

Akçakın, V., Cebesoy, Ü. B., & İnel, Y. (2015). İki Boyutlu Matematik Kaygısı Ölçeğinin Türkçe formunun geçerlik ve güvenirlik çalışması. *GEFAD / GUJGEF 35*(2), 283-301. https://dergipark.org.tr/en/download/article-file/77532

*Akgül, S., & Nuhoğlu, H. (2020). Üstün yetenekli öğrencilerin matematik kaygısı ve mükemmeliyetçilik düzeylerinin incelenmesi. *Yaşadıkça Eğitim, 34*(2), 299-312. https://doi.org/10.33308/26674874.2020342204

Akın, A., Kurbanoğlu, İ., & Takunyacı, M. (2011). Revize edilmiş Matematik Kaygısı Değerlendirme Ölçeği: Doğrulayıcı faktör analizi çalışması. *Necatibey Eğitim Fakültesi Elektronik Fen ve Matematik Eğitimi Dergisi*, *5*(1), 163-180. https://dergipark.org.tr/en/pub/balikesirnef/issue/3372/46541

Alexander, L., & Martray, C. (1989). The development of an abbreviated version of the Mathematics Anxiety Rating Scale. *Measurement and Evaluation in Counseling and Development, 22*, 143-150. https://doi.org/10.1080/07481756.1989.12022923

Alkan, V. (2011). Etkili matematik öğretiminin gerçekleştirilmesindeki engellerden biri: Kaygı ve nedenleri. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi*, *29*, 89-107. https://dergipark.org.tr/en/download/article-file/114598

Arı, K., Savaş, E., & Konca, Ş. (2010). İlköğretim 7. sınıf öğrencilerinin matematik kaygısının nedenlerinin incelenmesi. *Selçuk Üniversitesi Ahmet Keleşoğlu Eğitim Fakültesi Dergisi*, *29*, 211-230. https://docplayer.biz.tr/8305492-Ilkogretim-7-sinif-ogrencilerinin-matematik-kaygisinin-nedenlerinin-incelenmesi.html

*Aydın, M., & Keskin, İ. (2017). 8. sınıf öğrencilerinin matematik kaygı düzeylerinin bazı değişkenler açısından incelenmesi. *Kastamonu Eğitim Dergisi, 25*(5), 1801-1818. https://dergipark.org.tr/en/download/article-file/349010

Bai, H., Wang, L., Pan, W., & Frey, M. (2009). Measuring mathematics anxiety: Psychometric analysis of a bidimensional affective scale. *Journal of Instructional Psychology*, *36*, 185-193. https://web.s.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=0&sid=5d119e25-be27-45f6-98ea-2de5e749dc5a%40redis

*Baklacı, E. (2017). *Ortaokul öğrencilerinin matematik dersine yönelik kaygı ve umutsuzluk düzeyleri* [The levels of secondary school students' anxiety and despite for the mathematics course] [Unpublished master thesis, Pamukkale University]. http://acikerisim.pau.edu.tr/xmlui/handle/11499/2505

Baloğlu, M. (2005). Matematik Kaygısını Derecelendirme Ölçeği'nin Türkçe'ye uyarlanması, dil geçerliği ve ön psikometrik incelemesi. *Kuram ve Uygulamada Eğitim Bilimleri*, *5*(1), 7-30. https://toad.halileksi.net/sites/default/files/pdf/matematik-kaygisini-derecelendirme-olcegi-toad.pdf

Baloğlu, M., & Balgamış, E. (2010). Matematik Kaygısını Derecelendirme Ölçeği İlköğretim Formu'nun Türkçe'ye uyarlanması, dil geçerliği ve psikometrik incelemesi. *Kuram ve Uygulamada Eğitim Bilimleri*, *10*(1), 77-110. https://www.researchgate.net/profile/Mustafa-Baloglu/publication/286925904_MARS-E_adaptation_and_validation/links/56701d5b08aececfd5530e4a/MARS-E-adaptation-and-validation.pdf

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50* (4), 1088–1101. https://www.jstor.org/stable/2533446

*Berber, M. (2021). *Rasyonel sayılar konusunun farklılaştırılmış öğretim yöntemi ile öğretiminin farklı öğrenme stillerine sahip olan yedinci sınıf öğrencilerinin akademik başarılarına ve matematik kaygılarına etkisi [The effect of teaching rational numbers with a differentiated teaching method on the academic achievement and math anxiety of seventh grade students who have different learning styles]* [Unpublished master thesis, Bursa Uludağ University]. http://acikerisim.uludag.edu.tr/jspui/bitstream/11452/18998/1/MERVE_BERBER.pdf

*Bindak, R. (2005). İlköğretim öğrencileri için matematik kaygı ölçeği. *Fırat Üniversitesi Fen ve Mühendislik Bilimleri Dergisi*, *17*(2), 442-448. https://toad.halileksi.net/sites/default/files/pdf/ilkogretim-ogrencileri-icin-matematik-kaygi-olcegi-toad.pdf

*Birgin, O., Baloğlu, M., Çatlıoğlu, H., & Gürbüz, R. (2010). An investigation of mathematics anxiety among sixth through eighth grade students in Turkey. *Learning and Individual Differences, 20*(6), 654-658. https://doi.org/10.1016/j.lindif.2010.04.006

Bonett, D. G. (2002). Sample size requirements for estimating intraclass correlations with desired precision. *Statistics in Medicine, 21*(9), 1331–1335. https://doi.org/10.1002/sim.1108

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

*Borlat, G. (2018). *Yaratıcı drama yönteminin matematik kaygısı ve motivasyonuna etkisi [The effect of creative drama method on math anxiety and math motivation]* (Thesis No. 524662) [Unpublished master thesis, Çanakkale Onsekiz Mart University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Bozkurt, S. (2012). *İlköğretim ikinci kademe öğrencilerinde sınav kaygısı, matematik kaygısı, genel başarı ve matematik başarısı arasındaki ilişkilerin incelenmesi [An investigation into the relationship between test anxiety, mathematics anxiety, academic achievement and mathematics achievement of the seventh and eight grade primary school students]* (Thesis No. 314728) [Unpublished master thesis, İstanbul University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Budak, S. (2000). *Psikoloji Sözlüğü*. (Birinci basım). Ankara: Bilim ve Sanat Yayınları

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment,* *7*(3), 309–319. http://www.bwgriffin.com/gsu/courses/edur9131/content/Clark_validity_scaledevelopment.pdf

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Holt, Rinehart and Winston

*Çağırgan, D., & Soytürk, İ. (2021). The relationship between mathematics anxiety, student engagement in mathematics and responsibilities towards learning among middle school students. *Ilkogretim Online, 20*(1), 456-467. https://doi.org/10.17051/ilkonline.2021.01.040

*Çoruk, H., & Çakır, R. (2017). Çoklu ortam kullanımının ilkokul öğrencilerinin akademik başarılarına ve kaygılarına etkisi. *Türk Bilgisayar ve Matematik Eğitimi Dergisi, 8*(1), 1-27. doi: 10.16949/turkbilmat.286655

*Dede, Y., & Dursun, Ş. (2008). İlköğretim II. kademe öğrencilerinin matematik kaygı düzeylerinin incelenmesi. *Uludağ Üniversitesi Eğitim Fakültesi Dergisi, 21*(2), 295-312. https://dergipark.org.tr/en/download/article-file/153341

DeVellis, R. F. (1991). *Scale development: Theory and applications*. Sage Publications.

*Doruk, M., Öztürk, M., & Kaplan, A. (2016). Investigation of the self-efficacy perceptions of middle school students towards mathematics: Anxiety and attitude factors. *Adıyaman University Journal of Educational Sciences, 6*(2), 283-302. https://doi.org/10.17984/adyuebd.306387

*Duran, M., Doruk, M., & Kaplan, A. (2017). An examination of the effectiveness of argumentation-based probability teaching on middle school students' achievements and anxieties. *Eğitimde Kuram ve Uygulama, 13*(1), 55-87. https://dergipark.org.tr/tr/download/article-file/275203

*Dursun, Ş., & Bindak, R. (2011). İlköğretim II. kademe öğrencilerinin matematik kaygılarının incelenmesi. C*umhuriyet Üniversitesi Edebiyat Fakültesi Sosyal Bilimler Dergisi, 35*(1), 18-21. http://cujos.cumhuriyet.edu.tr/tr/download/article-file/49853

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*, 629–634. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2127453/pdf/9310563.pdf

*Erdik, C. (2018). The relationship between teacher autonomy support and the attitude and anxiety toward mathematics course. *International Education Studies, 11*(6), 109-118. https://doi.org/10.5539/ies.v11n6p109

*Ergenç, T. S. (2011). *İlköğretim yedinci sınıf öğrencilerinin matematik dersi bilişsel hazır bulunuşluk düzeyleri ile matematik kaygı düzeyleri arasındaki ilişkinin incelenmesi [The investigation of the relationship between the level of readness at Maths and the level of Maths anxiety of the 7th grade students]* (Thesis No. 296472) [Unpublished master thesis, Eskişehir Osman Gazi University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Fennema, E., & Sherman, J. A. (1976). Fennema-Sherman mathematics attitudes scales: Instruments designed to measure attitudes toward the learning of mathematics by females and males. *Journal for research in Mathematics Education*, *7*(5), 324-326. https://www.jstor.org/stable/748467

*Gevrek, L. (2009). *İlköğretim ikinci kademe öğrencilerinin öğrenilmiş çaresizlik düzeylerinin çeşitli değişkenler açısından incelenmesi [Analysing the level of learning helpnessness of secondary school students according to some variables]* (Thesis No. 244466) [Unpublished master thesis, Eskişehir Osman Gazi University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Glass, G. V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher, 5*, 3–8. https://www.jstor.org/stable/1174772

Göçen, A., & Şen, S. (2021). Spiritual leadership and organizational citizenship behavior: A meta-analysis. *SAGE Open*, *11*(3), 1-16. https://doi.org/10.1177/21582440211040777

*Gündüz Çetin, İ. (2020). *Ortaöğretim öğrencilerinin matematik umutsuzluğunu yordayan değişkenler: matematik kaygısı, matematiğe yönelik motivasyonel inançlar, matematik başarısı (Köşk ilçesi örneği) [Variables affecting the mathematics hopelessness of secondary school students: Mathematic anxiety, motivational beliefs towards mathematics, mathematics success (Köşk district example)]* (Thesis No. 637929) [Unpublished master thesis, Aydın Adnan Menderes University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. D. Toit, ve D. Sörbom (Eds.), *Structural equation modeling: Present and future* (pp. 195-216). Lincolnwood, IL: Scientific Software International

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539-1558. https://doi.org/10.1002/sim.1186

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *Bmj, 327* (7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Hopko, D. R., Mahadevan, R., Bare, R. L., & Hunt, M. A. (2003). The Abbreviated Math Anxiety Scale (AMAS): Construction, validity, and reliability. *Assessment*, *10*, 178-182. https://doi.org/10.1177/1073191103010002008

Hunt, T. E., Clark-Carter, D., & Sheffield, D. (2011). The development and part validation of a U.K. Scale for mathematics anxiety. *Journal of Psychoeducational Assessment*, *29*, 455-466. https://doi.org/10.1177/0734282910392892

Ikegulu, T. N. (1998). An empirical development of an instrument to assess mathematics anxiety and apprehension. https://files.eric.ed.gov/fulltext/ED433235.pdf

*İlhan, M., & Öner Sünkür, M. (2012). Matematik kaygısı ile olumlu ve olumsuz mükemmeliyetçiliğin matematik başarısını yordama gücü. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 8*(1), 178-188. https://dergipark.org.tr/en/download/article-file/160812

*İlhan, M., & Öner Sünkür, M. (2013). Matematik kaygısının matematik başarısını yordama gücünün cinsiyet ve sınıf değişkenleri açısından incelenmesi. *Gaziantep University Journal of Social Sciences, 12*(3), 427-441. https://web.p.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=0&sid=98d217b3-4f15-4163-9bc3-345e90d900ab%40redis

Jackson, P.H., & Agunwamba, C.C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I: Algebraic lower bounds. *Psychometrika* 42, 567–578 (1977). https://link.springer.com/content/pdf/10.1007/BF02295979.pdf

*Kandal, R., & Baş, F. (2021). The prediction status of secondary school students' metacognitive awareness, self-regulatory learning strategies, anxiety and attitude levels towards mathematics on mathematics achievement. *International Journal of Educational Studies in Mathematics, 8*(1), 27-43. https://doi.org/10.17278/ijesim.834851

*Kesici, A. (2018a). Lise öğrencilerinin matematik motivasyonunun matematik başarısına etkisinin incelenmesi. *Ondokuz Mayıs Üniversitesi Eğitim Fakültesi Dergisi, 37*(2), 177-194. doi: 10.7822/omuefd.438550

*Kesici, A. (2018b). Matematik kaygısı ebeveynlerden çocuklara aktarılan kültürel bir miras mı? *Dicle Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 10*(20), 304-313. https://dergipark.org.tr/en/download/article-file/1716476

*Kutluca, T., Alpay, F. N., & Kutluca, S. (2015). 8. sınıf öğrencilerinin matematik kaygı düzeylerine etki eden faktörlerin incelenmesi. *Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi Dergisi*, *25*(2015), 202-214. https://dergipark.org.tr/en/download/article-file/786706

Kuzu, Ç. İ. (2021). Aile ve öğretmen yaklaşımlarının matematik kaygı düzeyine etkisi ile ilgili öğrenci görüşleri. *EKEV Akademi Dergisi,* *25*(85), 113–128. http://www.ekevakademi.org/Makaleler/2132927058_07%20Cigdem%20INCI%20KUZU.pdf

*Kuzu, O., & Çalışkan, N. (2018). *Öğretmen adaylarının motivasyon ve matematik kaygı düzeylerinin çeşitli değişkenler açısından incelenmesi*. C.Arslan, E. Hamarta, S. Çiftçi, M. Uslu, & O. Köksal (Eds.), Eğitim Bilimleri Çalışmaları (ss. 5–11). https://www.researchgate.net/publication/330967670

*Küçük, M. (2019). *Yazma etkinliklerinin matematik öğretiminde problem çözme becerisine, tutum ve kaygıya etkisi [The effects of writing activities on problem solving skills, attitudes and anxiety in mathematics teaching]* (Thesis No. 551550) [Unpublished master thesis, Pamukkale University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Maloney, E. A., Ramirez, G., Gunderson, E. A., Levine, S. C. & Beilock, S. L. (2015). Intergenerational effects of parents' math anxiety on children's math achievement and anxiety. *Psychological Science, 26*(9), 1480-1488. https://doi.org/10.1177/0956797615592630

Manav, F. (2011). Kaygı kavramı. *Toplum Bilimleri Dergisi*, *5*(9), 201-211. http://bnposta.bingol.edu.tr/bitstream/handle/20.500.12898/253/Kayg%C4%B1%20Kavram%C4%B1.pdf?sequence=1&isAllowed=y

McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology, 23*, 1–21. https://doi.org/10.1111/j.2044-8317.1970.tb00432.x

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*, 412–433. https://doi.org/10.1037/met0000144

Mutlu, Y. & Söylemez, İ. (2018). İlkokul 3. ve 4. sınıf çocukları için matematik kaygı ölçeği; Güvenirlik ve geçerlik çalışması. *EKEV Akademi Dergisi,* *22*(73), 429-440. https://app.trdizin.gov.tr/publication/paper/detail/TXpJME5EZzFOUT09

Mutlu, Y., Sarı, M. H., & Çam, Z. (2018). Ebeveyn Matematik Kaygısı Ölçeği geliştirme çalışması. *Anemon Muş Alparslan Üniversitesi Sosyal Bilimler Dergisi*, *6*(STEMES'18), 139-145. http://dx.doi.org/10.18506/anemon.463756

Mutlu, Y., Söylemez İ., & Yasul, A. F. (2017). İlkokul öğrencilerinin matematik kaygısı ile matematik başarıları arasındaki ilişkinin incelenmesi. *Journal of Human Sciences*, *14*(4), 4425-4434. doi:10.14687/jhs.v14i4.5019

Newstead, K. (1998). Aspects of children's mathematics anxiety. *Educational Studies in Mathematics, 36*(1), 53-71. https://link.springer.com/content/pdf/10.1023/a:1003177809664.pdf

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Peker, M. (2006). Matematik Öğretmeye Yönelik Kaygı Ölçeğinin geliştirilmesi. *Eğitim Bilimleri ve Uygulama*, *5*(9), 73-92. https://toad.halileksi.net/sites/default/files/pdf/matematik-ogretmeye-yonelik-kaygi-olcegi-toad.pdf

Petitti, D. B. (2000). *Meta-analysis, decision analysis and cost-effectiveness analysis: methods for quantitative synthesis in medicine.* New York: Oxford University Press.

Plake, B. S., & Parker, C. S. (1982). The development and validation of a revised version of the Mathematics Anxiety Rating Scale. *Educational and Psychological Measurement*, *42*, 551-557. https://doi.org/10.1177/001316448204200218

R Core Team. (2021). R: A language and environment for statistical computing R Foundation for Statistical Computing https://www.R-project.org/.

Richardson, F. C., & Suinn, R. M. (1972). The Mathematics Anxiety Rating Scale: Psychometric data. *Journal of Counseling Psychology*, *19*(6), 551-554. https://doi.org/10.1037/h0033456

Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://pages.ucsd.edu/~cmckenzie/Rosenthal1979PsychBulletin.pdf

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). Publication bias in meta-analysis. H. R., Rothstein (Ed.), A. J., Sutton, (Ed.), & M., Borenstein(Ed.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (1-7). Hoboken, NJ: John Wiley & Sons Inc.

Sánchez-Meca, J., Marín-Martínez, F., López-López, J. A., Núñez-Núñez, R. M., Rubio-Aparicio, M., López-García, J. J., & López-Nicolás, R. (2021). Improving the reporting quality of reliability generalization meta-analyses: The REGEMA checklist. *Research Synthesis Methods, 12*(4), 516-536. https://doi.org/10.1002/jrsm.1487

Sandman, R. S. (1980). The mathematics attitude inventory: Instrument and user’s manual. *Journal for Research in Mathematics Education*, *11*, 148-149. https://doi.org/10.5951/jresematheduc.11.2.0148

Sarı, M. H. (2014). Sınıf öğretmenlerine yönelik Matematik Öğretimi Kaygı Ölçeği geliştirme. *İlköğretim Online*, *13*(4), 1296-1310. https://web.s.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=0&sid=44676963-898a-4f52-9c56-80f4ee45a010%40redis

Semma, B., Henri, M., Luo, W., & Thompson, C. G. (2019). Reliability generalization of the meaning in life questionnaire subscales. *Journal of Psychoeducational Assessment, 37*(7), 837-851. https://doi.org/10.1177/0734282918800739

Schmid, C.H., Stijnen,T. & White, I. R. (2021). General themes in meta-analysis. C.H. Schmid (Ed.), T. Stijnen (Ed.), & I. R. White(Ed.), *Handbook of meta-analysis* (19-26). Boca Raton: CRC Press.

Suinn, R. M., & Edwards, R. (1982). The measurement of mathematics anxiety: The mathematics anxiety rating scale for adolescents: MARS-A. *Journal of Clinical Psychology*, *38*(3), 576-580. https://doi.org/10.1002/1097-4679(198207)38:3<576::AID-JCLP2270380317>3.0.CO;2-V

Suinn, R. M., Taylor, S., & Edwards, R. W. (1988). Suinn Mathematics Anxiety Rating Scale for Elementary School Students (MARS-E): Psychometric and normative data. *Educational and Psychological Measurement*, *48*(4), 979-986. https://doi.org/10.1177/0013164488484013

Şad, A.N., Kış, Demir, M., & Özer, N. (2016). Meta-analysis of the relationship between mathematics anxiety and mathematics achievement. *Pegem Eğitim ve Öğretim Dergisi*, *6*(3), 371-392. https://doi.org/10.14527/pegegog.2016.019

Şahin, F.Y. (2008). Mathematics anxiety among 4th and 5th grade turkish elementary school students. *International Electronic Journal of Mathematics Education*, *3*(3), 179-192. https://www.iejme.com/download/mathematics-anxiety-among-4th-and-5th-grade-turkish-elementary-school-students.pdf

*Şahin, S. (2018). Investigation of the effect of enhanced education programmes on mathematics anxiety of students. *Online submission, 4*(11), 255-274. https://doi.org/10.5281/zenodo.1328249

Şan, İ., & Akdağ, M. (2017). Ortaokul Öğrencileri İçin Matematik Sınavı Kaygısı Ölçeği: bir ölçek geliştirme çalışması. *Necatibey Eğitim Fakültesi Elektronik Fen ve Matematik Eğitimi Dergisi*, *11*(1), 128-159. https://doi.org/10.17522/balikesirnef.356142

Şen, S. (2021). A reliability generalization meta-analysis of Runco Ideational Behavior Scale. *Creativity Research Journal*, *34*(2), 178-194. https://doi.org/10.1080/10400419.2021.1960719

Şen, S., & Yıldırım, İ. (2020). *CMA ile meta-analiz uygulamaları*. Ankara: Anı Yayıncılık

*Şimşek, H., Şahinkaya, N., & Aytekin, C. (2017). Investigation of the anxieties and attitudes of elementary school students towards mathematics with various variables. *Necatibey Faculty of Education Electronic Journal of Science & Mathematics Education, 11*(2), 82-108. https://dergipark.org.tr/tr/download/article-file/395970

*Tabakçı, S. (2018). *Matematik kaygısı ile çocuklarda öğrenilmiş çaresizlik arasındaki ilişkinin incelenmesi [An investigation of the mathematics anxiety and learned helplessness levels of primary school 4th grade* *students in terms of various variables and relation between them]* (Thesis No. 510986) [Unpublished master thesis, Toros University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Tan, M. N. (2015). *Ortaokul öğrencilerinin matematik kaygısı öğrenilmiş çaresizlik ve matematiğe yönelik tutum düzeyleri arasındaki ilişkilerin incelenmesi [The investigation of relationships among the levels of secondary school students'mathematics anxiety, learned helplessness and attitudes toward mathematics]* (Thesis No. 407580) [Unpublished master thesis, Necmettin Erbakan University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

*Taşdemir, C. (2015). Ortaokul öğrencilerinin matematik kaygı düzeylerinin incelenmesi. *Batman Üniversitesi Yaşam Bilimleri Dergisi, 5*(1), 1-12. https://dergipark.org.tr/tr/download/article-file/313384

Thornton, A., & Lee, P. (2000). Publication bias in meta-analysis: Its causes and consequences. *Journal of Clinical Epidemiology, 53*(2), 207-216. https://doi.org/10.1016/S0895-4356(99)00161-4

Tobias, S., & Weissbrod, C. (1980). Anxiety and mathematics: An update. *Harvard Educational Review, 50*(1), 63-70. https://doi.org/10.17763/haer.50.1.xw483257j6035084

*Tok, Ş. (2013). Effects of the know-want-learn strategy on students' mathematics achievement, anxiety and metacognitive skills. *Metacognition Learning, 8*(2), 193-212. https://doi.org/10.1007/s11409-013-9101-z

Tooke, D.J.L. & Leonard, C. (1998). Effectiveness of a mathematics methods course in reducing mathematics anxiety of preservice elementary teachers. *School Science and Mathematics*, *98*(3) 136-142. https://doi.org/10.1111/j.1949-8594.1998.tb17406.x

*Tuncer, M., & Şimşek, M. (2019). Ortaokul 5. sınıf matematik dersi bölme işlemi konusunda plickers uygulamasının matematik kaygısına ve matematik başarısına etkisi. *OPUS Uluslararası Toplum Araştırmaları Dergisi, 13*(19), 281-310. https://doi.org/10.26466/opus.578149

*Tuncer, M., & Yılmaz, Ö. (2016). Ortaokul öğrencilerinin matematik dersine yönelik tutum ve kaygılarına ilişkin görüşlerinin değerlendirilmesi. *Kahramanmaraş Sütçü İmam Üniversitesi Sosyal Bilimler Dergisi, 13*(2),47-64. https://dergipark.org.tr/en/download/article-file/222175

Üldaş, İ. (2005). Öğretmen ve öğretmen adaylarına yönelik Matematik Kaygı Ölçeği (MKÖ-Ö)'nin geliştirilmesi ve matematik kaygısına ilişkin bir değerlendirme *[Development of a mathematics anxiety scale towards teachers and teacher candidates (MAS-T) and an assessment on mathematics anxiety]* (Thesis No. 188782) [Unpublished master thesis, Marmara University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Üstün, U., & Eryılmaz, A. (2014). Etkili araştırma sentezleri yapabilmek için bir araştırma yöntemi: Meta-analiz. *Eğitim ve Bilim, 39*(174), 1-32. http://dx.doi.org/10.15390/EB.2014.3379

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

Viechtbauer, W. (2010). Metafor meta-analysis package for R, Version 1.4-0 [Computer software]. Retrieved from http:// CRAN.R-project.org/packag=metafor

Wadlington, E., & Wadlington, P.L. (2008). Helping students with mathematical disabilities to succeed. *Preventing School Failure*, *53*(1), 2-7. https://doi.org/10.3200/PSFL.53.1.2-7

Wilkinson, L. (1999). The Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594-604. https://doi.org/10.1037/0003-066X.54.8.594

*Yetgin, O. (2017). *Ortaöğretim öğrencilerinin matematik kaygısı ve öğrenmeye ilişkin tutumlarının incelenmesi [İnvestigation of secondary school students'mathematical anxiety and the attitudes towards learning mathematics]* (Thesis No. 460478) [Unpublished master thesis, Adıyaman University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Yıldırım, K., & Gürbüz, R. (2017). Öğretmenlere Yönelik Matematik Kaygı Ölçeği (Ö-MKÖ) geliştirme çalışması. *Eğitimde Kuram ve Uygulama*, *13*(3), 392-410. https://doi.org/10.17244/eku.331901

Yumuşak, H., & Korkmaz, M. (2021). Liderlik ve okul etkililiği düzeyleri arasındaki ilişki: Bir meta-analiz çalışması. *Erciyes Journal of Education*, *5*(2), 123-149. https://doi.org/10.32433/eje.980997

*Yurt, E., & Kurnaz, A. (2015). An investigation of the effects of the mathematics sources of self-efficacy on talented students' mathematics anxiety. *Pegem Journal of Education and Instruction*, *5*(4), 347-360. https://doi.org/10.14527/pegegog.2015.019

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2022; 13(2); 134-144

Measurement Equivalence of the Turkish Version of the Satisfaction with Life Scale across Age

Devrim ERDEM *

Abstract
The Satisfaction with Life Scale is a widely used measurement tool. In this study, it was aimed to examine the measurement invariance of the Turkish version of the life satisfaction scale across ages. Of the 483 people participating in the study, 198 were men and 285 were women. Participants were divided into two age ranges, 18-24 and 25-43. A first-order single factor solution was provided for both the entire group and all age groups. The findings showed that the configural invariance was achieved by age groups. However, metric invariance could not be provided for age groups. The absence of equivalent factor loadings between the ages of 18-24 and 25-43 means that the latent structure measured by life satisfaction according to age groups does not have the same meaning. Partial metric invariance was obtained when the constraint of the 4th item parameter was freely estimated in further analysis. Subsequent analysis showed that scalar invariance was supported. On the other hand, full strict invariance could not be obtained, but only partially when the parameter constraint of item 1 was released. In summary, the results of this study revealed that comparison of age groups is possible with invariant items. It is hoped that this research will help us to clarify and deepen our inferences about life satisfaction and lifespan.

Keywords: Life satisfaction, age, invariance, partial metric, partial strict

Introduction

Studies on examining psychological variables in a cultural context and comparing them between cultures have intensified in the last three decades. One of the main reasons for this is to determine whether the psychological construct of interest is specific to the developed culture or a structure that has intercultural validity. With such studies, our understanding of these psychological constructs also improves (Dimitrov, 2010; Leong et al., 2010).

As of their origins, psychological theories and scales are of western origin. The validity of these scales and theories in different languages and cultures is investigated through adaptation. However, following the routine adaptation processes does not mean that the scores of the relevant psychological structure are comparable across cultures and between subgroups (Hambleton, 2005; Sireci, 2005). It has been suggested that taxonomic equivalence should be examined in evaluating the comparability of translated or adapted tests (van de Vijver & Tanzer, 2004). The most important thing in the adaptation process is to be able to figure out whether the scores obtained from the adapted scale are at a common scale level with the original scale (Sireci, 2005). Provided that this is achieved, the comparability of the scores becomes meaningful. In other words, to the extent that equivalence is not achieved, the comparability of scores becomes limited (van de Vijver & Poortinga, 2005). The impairment of equivalence may mean that test scores may be affected by cultural bias.

A series of incremental processes are required to ensure score comparability across cultures. The first stage is to test the configural equivalence (providing the same factor pattern between the comparison groups), the second stage requires testing the metric equivalence (providing the same factor loadings between the groups), and the third stage requires testing the scalar equivalence (providing the same intercepts and same measurement unit between the groups) (Dimitrov, 2010; van de Vijver &

* Assoc. Prof., Niğde Ömer Halisdemir University, Faculty of Education, Niğde-Turkey, erdem_devrim@yahoo.com, ORCID ID: 0000-0003-1810-2454

To cite this article:
Erdem, D. (2022). Measurement equivalence of the Turkish version of the satisfaction with life scale across age. *Journal of Measurement and Evaluation in Education and Psychology, 13*(2), 134-144. https://doi.org//10.21031epod.923785

Received: 20.04.2021
Accepted: 25.03.2022

Poortinga, 2005). However, if an equivalence at the scalar level can be achieved, it is able to compare the characteristics of individuals included in diverse groups in a valid and direct manner. When scalar equivalence is mentioned, it is assumed that measurements are made completely independent of bias (van de Vijver & Poortinga, 2005). Strict equivalence is considered the top-order equivalence and refers to the residual variances equivalence between comparison groups. However, invariance testing "across groups is often loosely applied (or not at all) in studies that deal with validation of assessment instruments in counseling and education" (Dimitrov, 2010, p. 121).

Subjective well-being (SWB) refers to individuals' judgments about evaluating their lives which "include people's emotional reactions to events, their moods, and judgments they form about their life satisfaction, fulfillment, and satisfaction with domains such as marriage and work" (Diener et al., 2003, p. 404). The cognitive dimension of subjective well-being, which is presented as a two-dimensional construct, has been defined to express life satisfaction (Diener et al., 1985). The reason why the cognitive dimension is important is that it is a more robust construct "typically not susceptible to change due to short-term emotional reactions to life events" (Proctor et al., 2009, p. 129).

Life satisfaction has been an issue studied in numerous areas such as education, health, psychology, social sciences, and economics (Diener et al., 1999). This is because life satisfaction is related to many variables (Tomás et al., 2015), including the individual's personality (Diener et al., 2003; Meléndez et al., 2019), self-esteem and self-perception (Miller et al., 2019), attitudes (Crowe & Kim, 2020), social support (Hansson et al., 2005), job satisfaction (Ilies et al., 2019), financial situation (Steckermeier, 2021), mental health (Fergusson et al., 2015; Huebner et al., 2000), psychiatric disorders (in Goldbeck et al., 2007) and health behavior (Grant et al., 2009).

Life satisfaction is also highly correlated with educational variables. Researches have indicated that life satisfaction is connected with academic achievement (Areepattamannil & Bano, 2020), student engagement (Hakimzadeh et al., 2016), achievement goals (Antaramian, 2017; Diseth et al., 2012), academic competence (Leung et al., 2004), academic self-efficacy (Kandemir, 2014; O'Sullivan, 2011), achievement motivation, academic stress and locus of control (Karaman & Watson, 2017), school climate (Suldo et al., 2008) and academic procrastination (Balkıs, 2013). On that account, it is very valuable to scrutinize life satisfaction in understanding the characteristics of human and social welfare in behavioral sciences as well as in education (Diener et al., 2003). Besides, life satisfaction has become accepted as an indicator of social progress and development for policymakers and social scientists (Stiglitz et al., 2009). The use of such statistical indicators is important and necessary as they reflect "modern economies and the widespread use of information technology" (Stiglitz et al., 2009, p. 7) as well as the developments in education. Besides being an educational indicator, due to the importance and functionality of the critical role it plays in decision-making mechanisms, the tools employed to ascertain the level of life satisfaction should be well tested psychometrically.

The Satisfaction with Life Scale (SWLS; Diener et al., 1985) is a short and easy-to-use self-report tool. Therefore, it has been adapted to many cultures and languages. In addition, it is widely used in comparing life satisfaction between cross-cultural and socio-demographic groups. If the metrics used are not equivalent for the comparison groups, inferences based on these metrics may be flawed or biased. Therefore, measurement invariance studies should be conducted for life satisfaction to be valid for both the relationships with other variables and the scores obtained for the comparison groups.

The measurement invariance of life satisfaction (LS) according to many socio-demographic variables was examined. One of these variables was age; since it was stated in the studies that life satisfaction could change with age depending on factors such as attitude, health, and social-economic status (Suh et al., 2012). Life satisfaction is sensitive to changing life conditions; therefore, life satisfaction is expected to change depending on age (Hartung et al., 2021). Cross-sectional and longitudinal studies have shown that life satisfaction scores change as life periods change (Bittmann, 2021; Chen, 2001; Jovanović & Lazić, 2020)

Since life satisfaction is an age-related variable, life satisfaction in different age groups was examined. Inconsistent findings on life satisfaction regarding age in the literature necessitated the continuity of these studies. In some studies, measurement invariance was obtained among age groups. For instance, Checa et al. (2019) obtained scalar invariance in LS between the groups aged 18-24 years and aged 25-47 years. Durak et al. (2010) also reported no statistically significant difference between the constrained and non-constrained models among the ages. Tomás et al. (2015) found strict invariance across age (among 14–65-year-old). Likewise, Ortuño-Sierra et al. (2019) found scalar invariance for participants' ages ranged from 13 to 19. Esnaola et al. (2017) found scalar invariance in LS among adolescents. Similarly, Bacro et al. (2020) found strict invariance across students aged 8 to 16 years.

On the other hand, different findings regarding the equivalence of life satisfaction scores in age groups were also reported in some studies. Clench-Aas et al. (2011) found partial scalar invariance was met, but strict invariance couldn't be achieved across age groups. Pons et al. (2000) observed factor structure and factor loading non-invariance between adolescent and elderly groups. Hultell and Gustavsson (2008) stated in their study that two items were sensitive to the three age groups (24 years old or younger, 25 to 34 years, and 35 years old or older). The researchers stated that age differences in life satisfaction might be "the result of adaptation strategies, cohort effects or age-specific life circumstances" (Westerhof et al., 2001, p. 183). Likewise, in the study conducted by Chen (2001), it was stated that age groups might have a cohort effect on life satisfaction. The previous research findings yielded traces of measurement invariance for particular age groups. Therefore, measurement invariance should be inspected to provide valid inferences on age-based life satisfaction latent mean comparisons.

In the literature, there are different results regarding the form of the relationship between age and life satisfaction (Pinquart & Sörensen, 2000). A considerable amount of research from large-scale life satisfaction surveys shows a U-shaped pattern of life satisfaction with age. (Blanchflower, 2020, Blanchflower & Oswald, 2008; Frijters & Beatton 2012; Park et al., 2020). Hudomiet et al. (2021) stated "average life satisfaction is high at younger ages, reaches a minimum at about age 40, which is sometimes called the 'midlife crisis,' after which it monotonically increases" (p. 1). Checa et al. (2019) examined measurement invariance in two groups aged 24 years or younger and older than 24 years in their study. The researchers declared that "these two groups were used as life-change references in previous SWLS invariance studies" (Checa et al., 2019, p. 267). Clench-Aas et al. (2011), on the other hand, examined the measurement invariance of life satisfaction according to the 16-24 and 25-44 age groups. Tomás et al. (2015) also separated two of the age groups as 18-24 and 25-34 in their study in which they examined the measurement invariance of life satisfaction according to age. In the current study, age groups were determined in accordance with the literature. Participants aged 24 and under are still defined as students, and participants aged 25 and over are defined as a graduate, non-working or working. In addition, it was deemed appropriate to take the cut-off age as 24 in order to provide a sufficient number to examine the measurement invariance in the older age group.

In short, cultural and contextual studies have been going on for a long since in life satisfaction studies. However, most of these studies focus on western and individualistic cultures. There are relatively a few studies drawn from eastern and collectivist cultural settings. However, a deep analysis of each psychological construct in a particular culture provides important information about both that variable and the culture under which it is studied (Cheung et al., 2011). Many studies conducted in Turkey have benefited from the satisfaction with life scale. However, life satisfaction measurement invariance has rarely been studied in the Turkish sample (e.g., Arıkan & Zorbaz, 2020). Furthermore, exploring age differences in life satisfaction can raise awareness to have more care for the psychological development of individuals throughout life. It can also help guide our assessments of educational and quality of life to advance the life satisfaction of individuals regarding age.

Purpose of the Study

Therefore, in this study, it was aimed to examine the measurement invariance of the Turkish version of the life satisfaction scale by age.

Method

This study is a cross-sectional study. At the same time, it is an explanatory study as it examines the measurement invariance of the Life Satisfaction Scale in this culture.

Participants

Data were collected with convenience sampling. There were 483 people in the sample. All individuals were volunteers. The 198 of the participants were men, and the 285 were women. The 309 participants were between the ages of 18-24, and 174 were between the ages of 25-43.

Data Collection Instruments

The SWLS developed by Diener et al. (1985) was used. There are five items in the SWLS scale. These items are as follows: (1) "In most ways my life is close to my ideal." (2) "The conditions of my life are excellent." (3) "I am satisfied with my life." (4) "So far I have gotten the important things I want in life." (5) "If I could live my life over, I would change almost nothing." The response set is taken from a 7-point Likert-type scale from 1 (*strongly disagree*) to 7 (*strongly agree*). Therefore, the total score can take values between 5 and 35. The scale is a one-dimensional self-reporting scale. Items measure perceived general life satisfaction. SWLS was adapted into Turkish first by Köker (1991) and then by Yetim (1993). Köker (1991) adapted the life satisfaction scale on 150 university students aged between 17 and 24 within the scope of his/her thesis. The researcher stated that the test-retest reliability of the scale was .85 and the item-test correlations were .71 to .80. The Turkish version of the scale, which was later adapted to Turkish culture by Yetim (1993), was reported to have .86 Cronbach alpha and .73 test-retest reliability. This current study revealed that Cronbach's Alpha reliability was .87 in the 18-24 age group and .82 in the 25-43 age group.

Data Analysis

Analyses were conducted with AMOS version 22. Multi-group confirmatory factor analyses (MGCFA) were performed to examine the measurement invariance. Following the suggestion of Hu and Bentler (1999), multiple fit indices such as the root mean square of error of approximation (RMSEA), standardized root mean square residual (SRMR), comparative fit index (CFI), Tucker–Lewis Index (TLI), Goodness of fit (GFI), Incremental Fit Index (IFI), chi-square (χ^2), and the chi-square to degrees of freedom ratio (χ^2/df) were taken into account in the evaluation. RMSEA and SRMR values less than .08, CFI, TLI, GFI, IFI values higher than .90, a non-significant p-value of the chi-square test and χ^2/df ratio of 3 or less indicate an acceptable fit to the data (Byrne, 2008; Hu & Bentler, 1999).

Measurement Invariance Tests

When examining measurement invariance, a taxonomic order is followed. Starting with the simplest/unconstrained model (configural), an upper constrained model (metric, scalar, and strict, respectively) is evaluated with gradually increased constraints. Before conducting a multi-group analysis, the model-data fit should be evaluated separately for both the whole group and each comparison group (Milfont & Fischer, 2010). In the invariance analysis, "a baseline model needs to be established prior to any invariance constraints" (Wu & Yao, 2006, p. 1263). Therefore, firstly two CFAs are conducted separately for participants in each age group. At this stage, after examining the model data fit indices and deciding that the baseline model is the same for both age groups, the configural model is tested. What is tested here is whether each of the age groups being compared has the same factor pattern. Model fit indices are taken into account when evaluating the configural model. If the configural model is achieved, it is time to test the more restricted metric model. In order to test for metric invariance, factor loadings are constrained to be equal across the groups (Wu & Yao, 2006). Thus, the equivalence of factor loadings is tested for age groups compared in the metric model. At this stage, difference tests are applied to test equivalence. In difference tests, the difference between an upper (more restricted) model and a lower (less restricted) model is tested. If the inequivalence is determined, partial equivalence is examined. In this study, each newly tested model is sequentially numbered (such as model 1, model 2, etc.). If metric or partial metric invariance is achieved, scalar invariance is tested in the next step. In scalar invariance process, the equivalence of latent intercepts for age groups is investigated. If scalar or partial scalar invariance is achieved, then the highest level of strict invariance is tested. The equivalence of residual variances is tested in strict invariance.

Difference tests are applied to evaluate these nested models (Cheung & Rensvold, 2002). In this study $\Delta\chi^2$, ΔCFI and ΔRMSEA were applied. The fact that the chi-square difference ($\Delta\chi^2$) between the models is not statistically significant means that invariance is provided (Dimitrov, 2010). Considering the suggestion of Chen (2007), a ΔCFI of ≥-.01 was used to indicate invariance, and ΔRMSEA of ≥ .015 was used to indicate of non-invariance between nested models (Chen, 2007).

In addition, the Akaike information criterion (AIC) value was also taken into account in the decision of model selection of nested models since the AIC is "one of the more popular methods of comparing multiple models, taking both descriptive accuracy and parsimony into account" (Wagenmakers & Farrell, 2004, p. 192). The smaller the AIC, the better the fit of the model. When comparing models, the lower AIC value is preferred. "However, the process of model evaluation is complicated by the fact that a model with many free parameters is more flexible than a model with only a few parameters" (Wagenmakers & Farrell, 2004, p. 192). Therefore, AIC has been evaluated together with the other indicators mentioned above when comparing nested models.

Results

Initial analysis displayed that there were no extreme responses or missing values. Means, standard deviations, skewness, and kurtosis for SWLS items were examined. The skewness and the kurtosis values were found to be within the range of (-2.00, 2.00). The values of means and standard deviations were given in Table 2. The reason why it is given in Table 2 is so that it can be evaluated together with other values. As a result of examining these values, no findings that violate the assumptions under CFA analysis were found.

Before examining the measurement invariance, a series of CFAs were performed. First of all, the model-data fit of the one-factor structure was evaluated for the entire sample. As seen in the Table 1, the overall model-data fit of the Turkish version of SWLS was quite well except for the chi-square value (χ^2 / df = 3.91). However, it has been noticed that the RMSEA (.079) value was very close to the cut-off value. Then, the baseline model was examined separately for age groups. The CFAs tests for separate groups revealed that model data fit was better in the older group than in the younger group (in Table 1).

Table 1

Goodness-of-fit Indexes for the Full Sample and the Baseline Model among Age Groups

Group	χ^2	*df*	*p*	χ^2/df	CFI	GFI	TLI	SRMR	RMSEA	90% CI for RMSEA L	90% CI for RMSEA U
Whole	19.58	5	.000	3.91	.98	.99	.97	.032	.079	.049	.111
18-24 aged	17.65	5	.003	3.53	.97	.98	.94	.053	.089	.051	.147
25-43 aged	6.03	5	.302	1.21	.99	.98	.99	.060	.045	.000	.150

As seen in Table 1, the 25-43 aged were provided excellent model-data fit. Although the incremental fit index values were very well for the 18-24 aged group, absolute fit indices such as chi-square and RMSEA indicated model misfit. Since chi-square is a statistic sensitive to the sample size, it is not surprising that it indicates a model mismatch. The literature states that when the degree of freedom (df) of the model is small, "the RMSEA too often falsely indicates a poor fitting model" (Kenny et al., 2015, p. 486). The fact that the chi-square and RMSEA could produce misleading results in assessing model data fit under current conditions, other indices were taken into account. According to these other indices, it was observed that the model-data fit was achieved across age groups.

Standardized factor loadings and error terms for the baseline model across age presented in Table 2. Standardized factor loadings varied from .51 and .82 for the 18-24 aged group, and .65 and .89 for the 25-43 aged group. Since no justification for modification was found in the analysis outputs, the unmodified model was used for multi-group tests. After the suitability of the baseline model was supported, it was time to examine a series of models in which the constraints were gradually increased in order to examine equivalence between age groups.

Measurement Invariance

Table 3 represents the results of comparison fit indexes between the 18-24 aged and 25-43 aged respondent groups. Model 1 fitted well across age groups, so configural invariance was met across the age groups. The change in chi-square between model 1 (configural) and model 2 (metric) was not statistically significant at the .01 level, but with regard to ΔCFI there was being present a significant diminish in model fit (ΔCFI = -.014 < -.01). The lack of support for metric invariance provided evidence of non-invariance factor loadings across age groups. At this stage, the difference between the factor loadings of the item is the most; the loading of that item is released. As given in Table 2, the difference between factor loadings for item 4 was found to be the highest. Additionally, tests showed that if factor loading of the item 4 (SWLS-4) is freely estimated across the age groups, partial metric invariance is achieved by assessing the differentiation in chi-square, ΔCFI and ΔRMSEA for model 1 and model 3 (partial metric) (Table 3). In the consequent process, scalar invariance was met. However, the comparison of model 4 (scalar) and model 5 (strict) pointed out significant decrease in model fit both statistically ($p < .01$) and practically (ΔCFI = -.034 > -.01). Since strict invariance was rejected, partial invariance was examined. After investigating which item or items caused the largest difference between the error terms in the unconstrained model across the age groups (Table 2), item 1 was detected the source of discrepancy. When the error parameter of item 1 freely estimated across age groups, non-significant model decrease was observed between the model 4 and model 6 (p = .092, ΔCFI = -.008 > -.01 and ΔRMSEA = -.001 < .015). Thus, partial strict invariance was met.

Table 2

Standardized Factor Loadings and Error Terms (for the baseline model) Across Age Groups

	Standardized factor loadings		Error variances		Mean (SD)	
Items	18 - 24	25 - 43	18 - 24	25 - 43	18 - 24	25 - 43
1	.72	.65	.66	1.35	4.85 (1.10)	4.71 (1.38)
2	.70	.76	.83	1.14	4.32 (1.26)	4.31 (1.44)
3	.82	.77	.54	.96	5.08 (1.29)	5.06 (1.38)
4	.68	.89	.91	.44	4.88 (1.18)	4.60 (1.47)
5	.51	.67	2.35	1.98	3.70 (1.70)	3.32 (1.91)

Table 3

Fit Indices for MGCFA Models and Difference Tests

Model	χ^2	df	χ^2/df	CFI	TLI	IFI	AIC	RMSEA
Model 1 Configural	26.172	10	2.617	.970	.94	.97	66.17	.072
Model 2 Metric	37.825	14	2.702	.956	.94	.96	69.83	.074
Model 3 Partial metric-I4	28.357	13	2.181	.972	.96	.97	62.36	.062
Model 4 Scalar	37.340	17	2.196	.963	.96	.96	83.34	.062
Model 5 Strict	60.433	22	2.747	.929	.94	.93	96.43	.075
Model 6 Partial strict-I1	45.327	21	2.158	.955	.96	.96	83.33	.061

Difference models	$\Delta\chi^2$	Δdf	p	ΔCFI	ΔRMSEA
Model 2 – Model 1	1.653	4	.020	-.014	.002
Model 3 – Model 1	2.185	3	.534	.002	-.01
Model 4 – Model 3	8.983	4	.062	-.009	.00
Model 5 – Model 4	23.093	5	.000	-.034	.013
Model 6 – Model 4	7.987	4	.092	-.008	-.001

Discussion and Conclusion

This study investigated measurement invariance for the Turkish version of life satisfaction items across the age groups. Moreover, its dimensionality in such a middle-eastern culture has been examined. The solution for the first-order one-factor model indicated an acceptable fit for the entire sample. Therefore, the one-dimensional structure of SWLS has been verified without modification. These results showed that the one-factor structure was supported in the entire group. This result was consistent with the other findings in the literature that the one-dimensional structure of SWLS is quite common (Emerson et al., 2017).

With regard to measurement invariance, configural invariance was observed, suggesting that the SWLS uni-dimensional model operates similarly between age groups. That means, a similar latent factor structure was affirmed with respect to age groups. Therewithal, among the age groups, the concept of life satisfaction was framed similarly. In this research, metric invariance was not supported according to age groups. The fact that metric invariance is not reached shows that the factor loadings connecting life satisfaction construct and items are not equivalent between age periods. The factor loadings were not invariant across samples aged 18-24 years and aged 25-43 years. The further tests indicated that individuals with aged 18-24 years and aged 25-43 years were assigned a different weight to item 4 “So far, I have gotten the important things I want in life” while evaluating their life satisfaction. Concerning the loading of item 4, it is .68 in the younger age group and .89 in the older age group. Item 4 had more weight in the older group. This implies item 4 more strongly related to life satisfaction for aged 25-43 years. Meanwhile, van de Vijver and Poortinga (2005) mentioned that differences in factor loadings on an item basis might indicate item bias. After the constraint on item 4 was set to free, partial metric invariance has achieved throughout age groups. In other words, it implies that item 4 does not have the same meaning for the age periods. This finding is consistent with Hultell and Gustavsson (2008), who pointed out that item 4 was sensitive to age. In their study, the findings did not demonstrate any invariance for three age groups (24 or younger, 25-34 years, and aged 35 or elder).

These results are in accordance with the findings of Pavot and Diener (2008), who stated that the item 4 might be linked with aging. Consistent with the results of this present study, it has been reported that item 4 is not invariant with respect to age groups in other studies (Bai et al., 2011). Esnaola et al. (2017) pointed out that “still there was only partial scalar invariance, with the intercept for item 4 varying across countries” (p. 597).

After obtaining partial invariance by releasing factor loadings’ constraint of item 4, the scalar invariance was tested. The scalar invariance was supported, which indicated no age differences occurred at the level of item intercepts. However, strict invariance has not been achieved throughout the age groups. Partial strict invariance was reached after the residual variance of item 1 was freely estimated. That is, the residual of item 1 (In most ways my life is close to my ideal) varied throughout the age groups. Emerson et al. (2017) pointed out that the interpretation of the words such as the word ideal in the item content might differ in various languages and cultural contexts. The word ideal is perhaps more questioned over time and may be an ambiguous concept for older individuals.

Differences in response patterns in item 4 and item 1 indicate that there may be conceptual differences between age groups. Previous studies reported similar results (Pavot & Diener, 2008). For instance, Clench-Aas et al. (2011) obtained partial metric invariance by freeing constraints on factor loadings for item 1. Likewise, Pons et al. (2000) found the error variances for the observed item 4 to be different between adolescents and elderly age groups. Similarly, Jovanović (2019) achieved partial scalar invariance by age. Emerson et al. (2017) declared that item 4 might have functioned differently between age groups because of differences in time conceptualization. The finding that item 4 is not invariant in age groups indicates that there is sensitivity to this item in the context of age subgroups.

Considering the age group, it was observed that the younger participants had higher -but not significant- mean scores on the LS items than the elderly. In parallel with this research result, Diener and Diener (1995) emphasized, in many parts of the world, it has been stated that university students are mostly satisfied with their lives. Similarly, Jovanović (2019) found that older adults reported the lowest life satisfaction. Likewise, in the study conducted by Chen (2001), it was stated that "life satisfaction decreased as age advanced" (p. 74).

The results confirmed a first-order single-factor solution in both for the whole group and each of the age group members. Briefly, these results mean that the single-factor latent structure of the Turkish version of SWLS is valid. The findings suggested that the configural invariance was achieved with regard to age groups. This represents the same factor pattern is valid among age groups. Metric invariance was not supported for age groups. The fact that there were not equivalent factor loadings across 18-24 years and aged 25-43 years suggested that the latent construct measured by the life satisfaction regarding age groups does not attribute the same meaning. However, invariant factor loadings for age were partially confirmed when the restriction of item 4 parameter was freely estimated. The forward analysis showed that scalar invariance was supported. On the other hand, full strict invariance could not be obtained, but only partially when the parameter constraint of item 1 was released. Summing up, the results of this study reveal that comparison of age groups is possible through invariant items.

This study was conducted in only two age categories. This limits the fulfillment of generalizability. It is recommended to repeat the future studies by expanding the diverse age span. Since such cross-sectional studies may limit our understanding of life satisfaction, it is recommended that future studies be conducted in such a way that researchers can examine all segments of society more comprehensively. For example, longitudinal panel studies may provide more valid results for age groups. Researchers are recommended to interpret with caution when comparing SWLS scores by age, as the results of the study show that partial invariance has been achieved.

Despite its limitations, this study has some implications. The knowledge that life satisfaction is a factor related to age gives hints to professionals on how to improve the physiological, mental and psychological well-being of the individuals. In addition, the emergence of age-specific non-invariance on life satisfaction items indicates that we should be careful when comparing between age groups to achieve valid inferences. Despite the variety of findings in the literature, the continuation of such research will help us to clarify and deepen our inferences regarding life satisfaction and life period/span.

References

Antaramian, S. (2017). The importance of very high life satisfaction for students' academic success. *Cogent Education*, *4*(1), 1307622. https://doi.org/10.1080/2331186X.2017.1307622

Areepattamannil, S., & Bano, S. (2020). Psychometric properties of the satisfaction with life scale (SWLS) among middle adolescents in a collectivist cultural setting. *Psychological Studies, 65*(4), 497-503. https://doi.org/10.1007/s12646-020-00578-4

Arıkan, Ç. A., & Zorbaz, S. (2020). Measurement invariance of the satisfaction with life scale across gender and time. *Turkish Journal of Education, 9*(4), 260-272. https://doi.org/10.19128/turje.774452

Bacro, F., Coudronnière, C., Gaudonville, T., Galharret, J. M., Ferrière, S., Florin, A., & Guimard, P. (2020). The French adaptation of the satisfaction with life scale (SWLS): Factorial structure, age, gender and time-related invariance in children and adolescents. *European Journal of Developmental Psychology, 17*(2), 307-316. https://doi.org/10.1080/17405629.2019.1680359

Bai, X., Wu, C., Zheng, R., & Ren, X. (2011). The psychometric evaluation of the satisfaction with life scale using a nationally representative sample of China. *Journal of Happiness Studies, 12*(2), 183-197. https://doi.org/10.1007/s10902-010-9186-x

Balkıs, M. (2013). Academic procrastination, academic life satisfaction and academic achievement: The mediation role of rational beliefs about studying. *Journal of Cognitive & Behavioral Psychotherapies, 13*(1), 57-74. https://www.researchgate.net/profile/Murat-Balkis-2/publication/236173872_Academic_procrastination_academic_life_satisfaction_and_academic_achievement_The_mediation_role_of_rational_beliefs_about_studying/links/0c960533a98186b6b9000000/Academic-procrastination-academic-life-satisfaction-and-academic-achievement-The-mediation-role-of-rational-beliefs-about-studying.pdf

Bittmann, F. (2021). Beyond the U-shape: Mapping the functional form between age and life satisfaction for 81 countries utilizing a cluster procedure. *Journal of Happiness Studies, 22*(5), 2343-2359. https://link.springer.com/article/10.1007/s10902-020-00316-7

Blanchflower, D. G. (2020). Unhappiness and age. *Journal of Economic Behavior & Organization, 176*, 461-488. https://www.nber.org/papers/w26642

Blanchflower, D. G., & Oswald, A. J. (2008). Is well-being U-shaped over the life cycle? *Social Science and Medicine, 66*(8), 1733-1749. https://www.nber.org/system/files/working_papers/w12935/w12935.pdf

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema, 20*(4), 872-882. Retrieved from https://reunido.uniovi.es/index.php/PST/article/view/8744

Checa, I., Perales, J., & Espejo, B. (2019). Measurement invariance of the satisfaction with life scale by gender, age, marital status and educational level. *Quality of Life Research, 28*(4), 963-968. https://doi.org/10.1007/s11136-018-2066-2

Chen, C. (2001). Aging and life satisfaction. *Social Indicators Research, 54*(1), 57-79. https://link.springer.com/content/pdf/10.1023/A:1007260728792.pdf

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464-504. https://doi.org/10.1080/10705510701301834

Cheung, F. M., van de Vijver, F. J., & Leong, F. T. (2011). Toward a new approach to the study of personality in culture. *American Psychologist, 66*(7), 593-603. https://doi.org/10.1037/a0022389

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness- of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5

Clench-Aas, J., Nes, R. B., Dalgard, O. S., & Aarø, L. E. (2011). Dimensionality and measurement invariance in the satisfaction with life scale in Norway. *Quality of Life Research, 20*(8), 1307-1317. https://doi.org/10.1007/s11136-011-9859-x

Crowe, A., & Kim, T. (2020). Ethnicity, life satisfaction, stress, familiarity, and stigma toward mental health treatment. *Journal of Counseling & Development, 98*(1), 83-93. https://doi.org/10.1002/jcad.12302

Diener, E., & Diener, M. (1995). Cross-cultural correlates of life satisfaction and self-esteem. *Journal of Personality and Social Psychology, 68*(4), 653-663. http://slatestarcodex.com/Stuff/diener2.pdf

Diener, E. D., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment, 49*(1), 71-75. https://emmons.faculty.ucdavis.edu/wp-content/uploads/sites/90/2015/08/1985_5-SWLS.pdf

Diener, E., Oishi, S., & Lucas, R. E. (2003). Personality, culture, and subjective well-being: Emotional and cognitive evaluations of life. *Annual Review of Psychology, 54*(1), 403-425. https://doi.org/10.1146/annurev.psych.54.101601.145056

Diener, E., Suh, E. M., Lucas, R. E., & Smith, H. L. (1999). Subjective well-being: Three decades of progress. *Psychological Bulletin, 125*(2), 276-302. https://doi.org/10.1037/0033-2909.125.2.276

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development, 43*(2), 121-149. https://doi.org/10.1177/0748175610373459

Diseth, A., Danielsen, A. G., & Samdal, O. (2012). A path analysis of basic need support, self-efficacy, achievement goals, life satisfaction and academic achievement level among secondary school students. *Educational Psychology, 32*(3), 335-354. https://doi.org/10.1080/01443410.2012.657159

Durak, M., Senol-Durak, E., & Gencoz, T. (2010). Psychometric properties of the Satisfaction with life scale among Turkish university students, correctional officers, and elderly adults. *Social Indicators Research, 99*(3), 413-429. https://doi.org/10.1007/s11205-010-9589-4

Emerson, S. D., Guhn, M., & Gadermann, A. M. (2017). Measurement invariance of the satisfaction with life scale: Reviewing three decades of research. *Quality of Life Research, 26*(9), 2251-2264. https://doi.org/10.1007/s11136-017-1552-2

Esnaola, I., Benito, M., Agirre, I. A., Freeman, J., & Sarasa, M. (2017). Measurement invariance of the satisfaction with life scale (SWLS) by country, gender and age. *Psicothema, 29*(4), 596-601. https://doi.org/10.7334/psicothema2016.394

Fergusson, D. M., McLeod, G. F. H., Horwood, L. J., Swain, N. R., Chapple, S., & Poulton, R. (2015). Life satisfaction and mental health problems (18 to 35 years). *Psychol Med, 45*(11), 2427-2436. https://doi.org/10.1017/S0033291715000422

Frijters, P., & Beatton, T. (2012). The mystery of the U-shaped relationship between happiness and age. *Journal of Economic Behavior & Organization, 82*(2-3), 525-542. https://doi.org/10.1016/j.jebo.2012.03.008

Goldbeck, L., Schmitz, T. G., Besier, T., Herschbach, P., & Henrich, G. (2007). Life satisfaction decreases during adolescence. *Quality of Life Research, 16*(6), 969-979. https://doi.org/10.1007/s11136-007-9205-5

Grant, N., Wardle, J., & Steptoe, A. (2009). The relationship between life satisfaction and health behavior: A cross-cultural analysis of young adults. *International Journal of Behavioral Medicine, 16*(3), 259-268. https://doi.org/10.1007/s12529-009-9032-x

Hakimzadeh, R., Besharat, M. A., Khaleghinezhad, S. A., & Ghorban Jahromi, R. (2016). Peers' perceived support, student engagement in academic activities and life satisfaction: A structural equation modeling approach. *School Psychology International, 37*(3), 240-254. https://doi.org/10.1177/0143034316630020
Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In: R. K. Hambleton, P. F. Merenda & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3-38). Lawrence Erlbaum Associates.
Hansson, A., Hillerås, P., & Forsell, Y. (2005). Well-being in an adult Swedish population. *Social Indicators Research, 74*, 313-325. https://doi.org/10.1007/s11205-004-6168-6
Hartung, J., Spormann, S. S., Moshagen, M., & Wilhelm, O. (2021). Structural differences in life satisfaction in a US adult sample across age. *Journal of Personality, 89(*6), 1232-1251. https://doi.org/10.1111/jopy.12654
Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55. https://doi.org/10.1080/10705519909540118
Hudomiet, P., Hurd, M. D., & Rohwedder, S. (2021). The age profile of life satisfaction after age 65 in the US. *Journal of Economic Behavior & Organization, 189*, 431-442. https://www.nber.org/system/files/working_papers/w28037/w28037.pdf
Huebner, E. S., Funk III, B. A., & Gilman, R. (2000). Cross-sectional and longitudinal psychosocial correlates of adolescent life satisfaction reports. *Canadian journal of school psychology, 16*(1), 53-64. https://doi.org/10.1177/082957350001600104
Hultell, D., & Gustavsson, J. P. (2008). A psychometric evaluation of the Satisfaction with life scale in a Swedish nationwide sample of university students. *Personality and Individual Differences, 44*(5), 1070-1079. https://doi.org/10.1016/j.paid.2007.10.030
Ilies, R., Yao, J., Curseu, P. L., & Liang, A. X. (2019). Educated and happy: A four-year study explaining the links between education, job fit, and life satisfaction. *Applied Psychology, 68*(1), 150-176. https://doi.org/10.1111/apps.12158
Jovanović, V. (2019). Measurement invariance of the Serbian version of the satisfaction with life scale across age, gender, and time. *European Journal of Psychological Assessment, 35*(4), 555-563. https://doi.org/10.1027/1015-5759/a000410
Jovanović, V., & Lazić, M. (2020). Is longer always better? A comparison of the validity of single-item versus multiple-item measures of life satisfaction. *Applied Research in Quality of Life, 15*(3), 675-692. https://doi.org/10.1007/s11482-018-9680-6
Kandemir, M. (2014). Reasons of academic procrastination: Self-regulation, academic self-efficacy, life satisfaction and demographics variables. *Procedia-Social and Behavioral Sciences, 152*, 188-193. https://doi.org/10.1016/j.sbspro.2014.09.179
Karaman, M. A., & Watson, J. C. (2017). Examining associations among achievement motivation, locus of control, academic stress, and life satisfaction: A comparison of US and international undergraduate students. *Personality and Individual Differences, 111*, 106-110. https://doi.org/10.1016/j.paid.2017.02.006
Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of free- dom. *Sociological Methods & Research, 44*(3), 486-507. https://doi.org/10.1177/0049124114543236
Köker, S. (1991). *Normal ve sorunlu ergenlerin yaşam doyumu düzeylerinin karşılaştırılması [Comparison of life satisfaction levels of normal and problematic adolescents]* (Thesis No. 16802) [Unpublished master's thesis, Ankara University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezDetay.jsp?id=hCZd2ZCwUfr1UUf132SIyw&no=hCZd2ZCwUfr1UUf132SIyw
Leong, F. T., Leung, K., & Cheung, F. M. (2010). Integrating cross-cultural psychology research methods into ethnic minority psychology. *Cultural Diversity and Ethnic Minority Psychology, 16*(4), 590-597. https://doi.org/10.1037/a0020127
Leung, C. Y., McBride-Chang, C., & Lai, B. P. (2004). Relations among maternal parenting style, academic competence, and life satisfaction in Chinese early adolescents. *Journal of Early Adolescence, 24*(2), 113-143. https://doi.org/10.1177/0272431603262678
Meléndez, J. C., Satorres, E., Cujiño, M. A., & Reyes, M. F. (2019). Big five and psychological and subjective well-being in Colombian older adults. *Archives of Gerontology and Geriatrics, 82*, 88-93. https://doi.org/10.1016/j.archger.2019.01.016
Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research, 3*(1), 111-130. https://doi.org/10.21500/20112084.857

Miller, B. K., Zivnuska, S., & Kacmar, K. M. (2019). Self-perception and life satisfaction. *Personality and Individual Differences*, *139*, 321-325. https://doi.org/10.1016/j.paid.2018.12.003

O'Sullivan, G. (2011). The relationship between hope, eustress, self-efficacy, and life satisfaction among undergraduates. *Social Indicators Research, 101*(1), 155-172. https://doi.org/10.1007/s11205-010-9662-z

Ortuño-Sierra, J., Aritio-Solana, R., Chocarro de Luis, E., Nalda, F. N., & Fonseca-Pedrero, E. (2019). Subjective well-being in adolescence: New psychometric evidences on the satisfaction with life scale. *European Journal of Developmental Psychology, 16*(2), 236-244. https://doi.org/10.1080/17405629.2017.1360179

Park, J., Joshanloo, M., & Scheifinger, H. (2020). Predictors of life satisfaction in Australia: A study drawing upon annual data from the Gallup World Poll. *Australian Psychologist, 55*(4), 375-388. https://doi.org/10.1111/ap.12441

Pavot, W., & Diener, E. (2008). The satisfaction with life scale and the emerging construct of life satisfaction. *The Journal of Positive Psychology*, *3*(2), 137-152. https://doi.org/10.1080/17439760701756946

Pinquart, M., & Sörensen, S. (2000). Influences of socioeconomic status, social network, and competence on subjective well-being in later life: A meta-analysis. *Psychology and Aging, 15*(2), 187-224. https://doi.org/10.1037/0882-7974.15.2.187

Pons, D., Atienza, F. L., Balaguer, I., & Garcia-Merita, M. L. (2000). Satisfaction with life scale: Analysis of factorial invariance for adolescents and elderly persons. *Perceptual and Motor Skills, 91*(1), 62-68. https://doi.org/10.2466/pms.2000.91.1.62

Proctor, C., Alex Linley, P., & Maltby, J. (2009). Youth life satisfaction measures: A review. *The Journal of Positive Psychology*, *4*(2), 128-144. https://doi.org/10.1080/17439760802650816

Sireci, S. G. (2005). Using bilinguals to evaluate the comparability of different language versions of a test. In R. K. Hambleton, P. F. Merenda & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 117-138). Lawrence Erlbaum Associates.

Steckermeier, L. C. (2021). The value of autonomy for the good life: An empirical investigation of autonomy and life satisfaction in Europe. *Social Indicators Research, 154*(2), 693-723. https://link.springer.com/content/pdf/10.1007/s11205-020-02565-8.pdf

Stiglitz, J. E., Sen, A., & Fitoussi, J. P. (2009). *Report by the commission on the measurement of economic performance and social progress*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.215.58&rep=rep1&type=pdf

Suh, S., Choi, H., Lee, C., Cha, M., & Jo, I. (2012). Association between knowledge and attitude about aging and life satisfaction among older Koreans. *Asian Nursing Research, 6*(3), 96-101. https://doi.org/10.1016/j.anr.2012.07.002

Suldo, S. M., Shaffer, E. J., & Riley, K. N. (2008). A social-cognitive-behavioral model of academic predictors of adolescents' life satisfaction. *School Psychology Quarterly*, *23*(1), 56-69. https://doi.org/10.1037/1045-3830.23.1.56

Tomás, J. M., Gutiérrez, M., Sancho, P., & Romero, I. (2015). Measurement invariance of the satisfaction with life scale (SWLS) by gender and age in Angola. *Personality and Individual Differences, 85*, 182-186. https://doi.org/10.1016/j.paid.2015.05.008

Van de Vijver, F. J., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39-63). Lawrence Erlbaum Associates.

Van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *54*(2), 119-135. https://doi.org/10.1016/j.erap.2003.12.004

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11*(1), 192-196. https://link.springer.com/content/pdf/10.3758/BF03206482.pdf

Westerhof, G. J., Dittmann-Kohli, F., & Thissen, T. (2001). Beyond life satisfaction: Lay conceptions of well-being among middle-aged and elderly adults. *Social Indicators Research, 56*(2), 179-203. https://link.springer.com/content/pdf/10.1023/A:1012455124295.pdf

Wu, C. H., & Yao, G. (2006). Analysis of factorial invariance across gender in the Taiwan version of the satisfaction with life scale. *Personality and Individual Differences, 40*(6), 1259-1268. https://scholars.lib.ntu.edu.tw/bitstream/123456789/42993/1/37.pdf

Yetim, U. (1993). Life satisfaction: A study based on the organization of personal projects. *Social Indicators Research, 29*(3), 277-289. https://link.springer.com/content/pdf/10.1007/BF01079516.pdf

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2022; 13(2); 145-163

Investigation of Variables Explaining Science Literacy in PISA 2015 Turkey Sample *

Şerife ZEYBEKOĞLU ** Hakan KOĞAR ***

Abstract

The purpose of this study was to examine the variables that explain science literacy with the answers given by Turkish students to the Program for International Student Assessment (PISA) 2015 student questionnaire. The Turkish sample of the research, which was conducted in a relational scanning model, is composed of 5895 students selected through a stratified sampling design. The sample of the study consists of 3052 people who remained after the data containing missing values were removed. In this study, Chi-squared Automatic Interaction Detection (CHAID) method, which is one of the data mining decision tree algorithms, was used for data analysis. As a result of the analysis, it was concluded that the variable that best explains the science literacy of Turkish students is "the number of books in the home". Other variables explaining the science literacy of Turkish students were also investigated in detail. In the research, gain values were examined to determine the most effective node in separating successful and unsuccessful students. As a result, it was seen that the most effective node consisted of students with science self-efficacy among students who had more than 200 books at home, and those who had more than 40 course hours in a week at school.

Keywords: PISA, science literacy, data mining, CHAID analysis

Introduction

In this period, in which we live in the information age, there are new developments in assessment every day. Miller et al. (2009, pp. 28) define assessment as "a general term that includes the full range of procedures used to gain information about student learning (observations, ratings of performances or projects, paper-and-pencil tests) and the formation of value judgments concerning learning progress". Russell and Airasian (2012, pp. 10) defined it as "a process of collecting, synthesizing and interpreting information in order to make a decision". As it can be understood from these definitions, the purpose of determining the situation is not to make a judgment about the student but to make an inference about his learning and performance. Due diligence studies are carried out at both national and international levels.

PISA is an international due diligence study. PISA is one of the largest international education studies organized by the Organization for Economic Cooperation and Development (OECD) that evaluates students' knowledge and skills in the fields of science, mathematics, and reading skills. This project covers the ability of students aged 15 who have reached the end of compulsory education to use what they have learned in school and in their out-of-school life, not how much they can remember what they have learned. Furthermore, it aims to determine the extent to which they can benefit from their knowledge and skills in order to understand new situations they will encounter, solve questions, make predictions about unfamiliar subjects, and make judgments. This purpose of PISA distinguishes it from other evaluation projects (Ministry of National Education [MoNE], 2010).

* This study is a part of master's thesis conducted under the supervision of Assoc. Prof. Dr. Hakan KOĞAR and prepared by Şerife ZEYBEKOĞLU.

** PhD. Student, Ankara University, Faculty of Education, Ankara-Turkey, serifezeybekoglu79@gmail.com, ORCID ID: 0000-0002-0378-263X
*** Assoc. Prof. Dr., Akdeniz University, Faculty of Education, Antalya-Turkey, hkogar@gmail.com, ORCID ID: 0000-0001-5749-9824

To cite this article:
Zeybekoğlu, Ş., & Koğar, H. (2022). Investigation of Variables Explaining Science Literacy in PISA 2015 Turkey Sample. *Journal of Measurement and Evaluation in Education and Psychology, 13*(2), 145-163. https://doi.org/10.21031/epod.798106

Received: 29.09.2020
Accepted: 23.01.2022

The experts in the countries participating in the PISA project form the evaluation framework and conceptual qualifications of the research, which are then approved by the joint decision of the participating countries as a result of the interviews. In this regard, a new understanding of the concept of "literacy" has been formed. The concept of literacy is defined as the capacity of students to make inferences from what they have learned in order to use their knowledge in daily life, to make logical inferences, to interpret and solve problems related to various situations. In this project, 15-year-old students are not asked to have learned everything they need like adults; they are expected to reach a solid foundation in science, mathematics and reading skills, to continue this learning throughout their lives and to be able to use the knowledge they have flexibly in their daily lives (MoNE, 2010).

The PISA project is carried out in three-year cycles. In each cycle, one of the areas of reading skills, mathematics and science literacy is examined in more detail, while the other areas are examined in less detail. PISA 2015 research focused on science literacy.

Science literacy is defined in PISA 2015 as "the ability to deal with science-related ideas and issues related to science as an active citizen". An individual who is scientifically literate wants to participate in speeches and conversations in the field of science-based on certain elements. For this, scientific explanation, scientific inquiry, and evaluation competencies are required. In science literacy, students' affective characteristics, such as their interests and attitudes towards science, can increase their motivation and affect their participation (MoNE, 2016).

Science literacy is important both at the national and international levels. Because humanity confronts great challenges in providing adequate water and food, controlling diseases, producing enough energy, and adapting to climate change. However, many problems also arise at the local level, where individuals may be faced with decisions regarding practices that affect their health and food resources, the appropriate use of materials, the use of new technologies, and energy use. In order to cope with all these difficulties, the contribution of science and technology is required. However, as discussed by the European Commission, "Unless young people have a certain scientific awareness, it is not a subject of conscious discussion." Also, that doesn't mean turning everyone into a scientific expert. But it is necessary to ensure that they take an enlightened role in making choices that affect their environment and to enable them to broadly understand the social implications of debates among experts. Given that the knowledge of science and science-based technology makes a significant contribution to the personal, social and professional lives of individuals, the understanding of science and technology is very important for the "preparation for life" of a teenager (OECD, 2017).

Everyone has individual differences, so it is not possible for all individuals to be scientifically literate at the same level. The important thing is to raise individuals who are interested in science, who can transfer and use what they have learned in their lives, and who have a level of knowledge to have a say in science-related issues in society. In this process, the desire for lifelong learning should be based. Some aim to pursue a career in science, while others seek leadership in science-related social issues. The important thing here is to give everyone the opportunity to learn the knowledge and skills that can meet the needs of society. The two main forces that shape our social life are science and technology. Therefore, in our age, societies want citizens to have the capacity to make decisions and make comparisons, when necessary, in the field of science and technology. The purpose of science education, which prepares children for the future, is to raise scientifically literate individuals because science and technology literacy is the basis of being effective in the world (Anagün, 2008).

The PISA application provides detailed information about the countries participating in the research, according to the established reference points, what their education levels are and what measures should be taken throughout the country. Thanks to this project, countries see the deficiencies in the education system, compare them with other countries and eventually find the opportunity to make new regulations regarding education.

In the literature review, many studies were found in which the variables affecting student success in PISA application were examined. In the study of Çeçen (2015), it was determined that the opportunities of the students at home, the cultural richness of the family, the education level of the parents and their positions at the workplace significantly predicted the PISA science literacy scores in 2003, 2006, 2009 and 2012 applications. In Karabay's (2013) study, it was found that the variables of the number of books in the home, having a room of their own, having a computer at home, and the education level of the parents were significant predictors for both PISA applications (2003, 2006, 2009) and application areas (science and mathematical literacy, reading skills). In addition to these, it was found that the variable of quality of educational resources at school was a significant predictor both during PISA applications (2003, 2006, 2009) and for application areas (science and mathematics literacy, reading skills) (Karabay, 2013). In the study of Özer and Anıl (2011), it was determined that the variable that most predicts students' science and mathematics achievement in the PISA 2006 application is the "time devoted to learning" variable. In the study of Anagün (2011), it was determined that the most important factor affecting science literacy in PISA 2006 application was "the time devoted to learning" and this was followed by the variable "inquiry-based learning activities".

In these studies, analysis was mostly made with parametric statistical methods, which require certain assumptions. In these studies, it was stated that assumptions such as missing data, extreme values, multicollinearity problem, normality, linearity, and homogeneity were met. Researchers generally used parametric methods such as factor analysis, analysis of variance, t-test, multiple linear regression, and structural equation modeling. In these studies, in which certain variables were selected from the student questionnaire applied in PISA, a limited number of variables were included in the analysis.

Despite the reforms in education in Turkey, it is quite remarkable that the average of success in the PISA project is below the OECD average. In this study, variables explaining science literacy in the PISA 2015 project were examined. The results of the research are important not only in science literacy but also in terms of taking necessary precautions and eliminating deficiencies in the education system. For this reason, the variables explaining the PISA 2015 science literacy of Turkish students were examined with CHAID analysis, one of the data mining methods. Thus, unlike other studies, the data did not need to provide various assumptions and many categorical and continuous predictor variables could be included in the analysis at the same time.

Purpose of the Study

The general purpose of the study is to examine the variables that explain science literacy with the answers given by Turkish students to the variables selected from the PISA 2015 student questionnaire. In line with the determined purposes, answers to the following questions are sought within the scope of this study:

1. Which predictor variable best explains the science literacy of Turkish students and divides the data set into homogeneous subgroups in the PISA 2015 Turkey sample?
2. Which variables respectively explain the science literacy of Turkish students in the PISA 2015 Turkey sample?
3. What is the order of importance of the predictor variables in classifying science literacy levels in the PISA 2015 Turkey sample?

Method

This research, conducted on Turkish students based on PISA 2015 data, is in a relational screening pattern of screening models.

Sample

In the PISA application, the sample setup is determined according to the stratified random sampling design through national centers. In this setup, the selection criteria (region, program type, school type) of the sample and schools are arranged. 187 schools and 5895 students from 61 provinces participated in the PISA 2015 application representing the 12 Statistical Regional Units Classification (İBBS) in Turkey (MoNE, 2016).

Since answering and not answering the questionnaire items in the student questionnaire used in the study may cause a bias, the missing values in the data set were analyzed first. Little (1988) states that if the missing values are randomly distributed, the list-based deletion method can be used. In this study, it was determined that the missing values were randomly distributed. Therefore, the missing values in the items in the student questionnaire were deleted. As a result, 3052 students who filled out the variables selected from the student questionnaire completely constitute the sample of the research.

Data Collection Tools

The measurement tools used in the research are the science literacy achievement test in the PISA 2015 application and the PISA 2015 student questionnaire applied to the students who answered this achievement test. In order to determine the variables that affect science literacy, student questionnaire items and indices created using these items were examined. The items and indexes selected from the student questionnaire within the scope of this research are given in Table 1.

Table 1

Variables Used in the Research on Science Literacy

	Variables	Description	Items	Number of Items
VARIABLES RELATED TO STUDENT, STUDENT'S FAMILY AND STUDENT'S HOME	ST001	Grade level	ST001Q01TA	1
	ST004	Gender	ST004Q01TA	1
	ST011	Home educational resources	ST011Q01TA, ST011Q02TA, ST011Q03TA, ST011Q04TA, ST011Q05TA, ST011Q06TA, ST011Q07TA, ST011Q08TA, ST011Q09TA, ST011Q10TA, ST011Q11TA, ST011Q12TA, ST011Q16TA	13
	ST012	Number of items at home	ST012Q01TA, ST012Q02TA, ST012Q03TA, ST012Q05NA, ST012Q06NA, ST012Q07NA, ST012Q08NA, ST012Q09NA	8
	ST013	Number of books in home	ST013Q01TA	1
	ST123	Parents emotional support	ST123Q01NA, ST123Q02NA, ST123Q03NA, ST123Q04NA	4
	ST125	Duration in early childhood care	ST125Q01NA	1
	ST126	Duration in early childhood education	ST126Q01TA	1
	MISCED	Mother's education	ST005, ST006	
	FISCED	Father's education	ST007, ST008	
	HISEI	Highest parental occupational status	ST014, ST015	
VARIABLES RELATED TO STUDENT'S OWN LIFE	ST118	Test anxiety	ST118Q01NA, ST118Q02NA, ST118Q03NA, ST118Q04NA, ST118Q05NA	5
	ST119	Achieving motivation	ST119Q01NA, ST119Q02NA, ST119Q03NA, ST119Q04NA, ST119Q05NA	5
	BSMJ	Student's expected occupational status	ST114	

Table 1 (continued)

	Variables	Description	Items	Number of Items
VARIABLES RELATED TO STUDENT'S SCHOOL	ST082	Collaborative problem solving	ST082Q01NA, ST082Q02NA, ST082Q03NA, ST082Q08NA, ST082Q09NA, ST082Q12NA, ST082Q13NA, ST082Q14NA	8
	ST034	Sense of belonging to school	ST034Q01TA, ST034Q02TA, ST034Q03TA, ST034Q04TA, ST034Q05TA, ST034Q06TA	6
	Unfairteacher	Teacher Fairness	ST039	
VARIABLES RELATED TO SCHOOL CALENDAR AND LEARNING TIME	ST060	The number of class periods per week attended in total	ST060Q01NA	1
	ST061	Average number of minutes in a class period	ST061Q01NA	1
	SMINS	Learning time per week in science (min)	ST059, ST061	
VARIABLES RELATED TO SCIENCE LEARNING IN SCHOOL	ST097	Disciplinary climate in science classes	ST097Q01TA, ST097Q02TA, ST097Q03TA, ST097Q04TA, ST097Q05TA	5
	ST098	Inquiry-based science teaching and learning practices	ST098Q01TA, ST098Q02TA, ST098Q03NA, ST098Q05TA, ST098Q06TA, ST098Q07TA, ST098Q08NA, ST098Q09TA, ST098Q10NA	9
VARIABLES RELATED TO STUDENT'S TENDENCY TO SCIENCE	ST092	Environmental awareness	ST092Q01TA, ST092Q02TA, ST092Q04TA, ST092Q05TA, ST092Q06NA, ST092Q08NA, ST092Q09NA	7
	ST093	Environmental optimism	ST093Q01TA, ST093Q03TA, ST093Q04TA, ST093Q05TA, ST093Q06TA, ST093Q07NA, ST093Q08NA	7
	ST094	Enjoyment of science	ST094Q01NA, ST094Q02NA, ST094Q03NA, ST094Q04NA, ST094Q05NA	5
	ST095	Interest in broad science topics	ST095Q04NA, ST095Q07NA, ST095Q08NA, ST095Q13NA, ST095Q15NA	5
	ST129	Science self-efficacy	ST129Q01TA, ST129Q02TA, ST129Q03TA, ST129Q04TA, ST129Q05TA, ST129Q06TA, ST129Q07TA, ST129Q08TA	8
	ST131	Epistemological beliefs	ST131Q01NA, ST131Q03NA, ST131Q04NA, ST131Q06NA, ST131Q08NA, ST131Q11NA	6

Data Analysis

In this study, the analysis of the data was carried out by using the SPSS Statistics 21 package program using the CHAID analysis, one of data mining decision trees algorithms. Decision trees are one of the most used classification methods. The creation and interpretation of decision trees is simpler than other methods. In addition, another advantage of decision trees is that the models they create are successful. In order to be able to classify decision trees in practice, a tree model is created in accordance with the available data, the data set is applied to this model and classification takes place in accordance with the result (Silahtaroğlu, 2013). The CHAID algorithm is one of the most widely used decision tree algorithms.

In this study, it was aimed to determine the variables that best explain science literacy from the variables related to the students themselves, their families, homes, and schools. In PISA 2015, 10 different PVSCIE coded science scores were calculated from the answers given to the achievement tests. Brown and Micklewright (2004) stated that it is difficult to easily combine success statistics into a single number in PISA. In order to combine these success statistics, the average score of each country in different tests can be taken into account (Brown & Micklewright, 2004). The average of these 10 different scores determined for science literacy was taken, and the average science literacy score was formed. Since the mean science literacy score has a correlation of around .95 with 10 different score types, it was chosen as the dependent variable in this study. Then, the mean of the dependent variable obtained was taken ($\bar{X}$=443.23), and this value was determined as the cut-off point. Students with a science literacy score below the determined value were categorized as “unsuccessful”, and students with a science literacy score equal to or higher than the average were categorized as “successful”. After this step, CHAID analysis from the data mining decision tree algorithm was performed to find answers to the research questions.

The performance of the tree model created as a result of the CHAID analysis was determined by the cross-validation method. Validation shows the generalizability of the established model to the universe. In the analysis phase, there are two types of validation: split-sample validation and cross validation (Aksu & Karaman, 2016). In split-sample validation, the data is separated into training and test data. The purpose of separating the data as training and testing is to determine the performance of the model on the data set it encounters for the first time. Some of the datasets are used to train the model, while the rest is used to test the model. In this way, models are produced by testing samples. In cross-validation, the sample is divided into k subsamples or multiples. A data set with a total of n samples is divided into k pieces, each of which contains n/k samples. Each time, a different dataset is reserved for testing, and the remaining k-1 dataset is used for training. This process is repeated k times, and at the end of each classification, the average of the performance values given to the tree is taken, and the performance of the model is determined in this way. In this study, the k number was determined as 10, and the data set was divided into 10 sub-samples.

The most important step in the creation of decision trees is to determine the criteria for branching in the tree or to create the tree structure according to which attribute values. There are various approaches developed for this in the literature. Han et al. (2012) explained these as gain ratio, gini index, χ^2 probability table statistics, and uncertainty coefficient, which takes into account the probability of each attribute value. In this study, χ^2 probability table statistics were used as the branching criterion in the tree.

Results

As a result of the CHAID analysis, a table is given about the correct classification of successful and unsuccessful students. Han et al. (2012) presented criteria to evaluate classification accuracy. These are accuracy (recognition rate), sensitivity (or recall), specificity, precision, F1, and F_β. When explaining these, four classifications are used: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

True positives (TP) refer to the positive tuples that were correctly labeled by the classifier. True negatives (TN) are the negative tuples that were correctly labeled by the classifier. False positives (FP) are the negative tuples that were incorrectly labeled as positive. False negatives (FN) are the positive tuples that were mislabeled as negative. The evaluation criteria starting from accuracy are introduced below:

- Accuracy: The percentage of test set tuples that are correctly classified by the classifier.
  [(TP+TN)/(P+N)]
- Sensitivity: The proportion of positive tuples that are correctly identified.
  [TP/P]
- Specificity: The proportion of negative tuples that are correctly identified.
  [TN/N]
- Precision: Measure of exactness (i.e., what percentage of tuples labeled as positive are actually such)
  [TP/(TP+FP)]

- Recall: A measure of completeness (what percentage of positive tuples are labeled as such). it is the same as sensitivity (or the true positive rate)
  [TP/(TP+FN)]=TP/P
- F_1 and F_β: An alternative way to use precision and recall is to combine them into a single measure.

The results related to the classification formed as a result of the analysis are given in Table 2.

Table 2
Classification Table for Success Status

	Predicted		
Observed	Successful	Unsuccessful	Success Percentage
Successful (P)	1207 (TP)	284 (FN)	81
Unsuccessful (N)	565 (FP)	996 (TN)	63.8
Total Percentage	58.1	41.9	72.2

Table 2 shows that 1207 of 1491 successful students were classified correctly in the model and the sensitivity (or recall) was 81%. Similarly, it was determined that 996 of 1561 unsuccessful students were classified correctly, and the specificity was 63.8%. It was seen that 1207 of 1772 students who were classified as successful were really successful, and the precision in classification was 68.1%. When the ratio of 1207 successful students and 996 unsuccessful students who were classified correctly to the total number of students was examined, it was determined that the classification accuracy was 72.2%. Apart from classification, the risk value table showing the margin of error of the model is given. Accordingly, the margin of error of the model was determined as approximately 27.8%.

As a result of the analysis, a decision tree with 52 nodes and 3 branches was formed. This decision tree is given at https://github.com/serifezeybekoglu/karar_agac/blob/main/karar_agac%C4%B1.pdf. Predictive variables explaining the students' science literacy, their order of importance on the dependent variable, and the frequency and percentage values related to the classification of successful and unsuccessful students are seen in the decision tree. When this tree is examined, the CHAID analysis, in which the average scores of science literacy are determined as the dependent variable, first of all, there is the variable that has the highest effect on the dependent variable.

There are 3052 students in this study. It is seen that 48.9% (1491) of the students were classified as successful and 51.1% (1561) as unsuccessful. According to the results of the CHAID analysis, among the items selected from the student questionnaire, the variable that best explains the science literacy of the students was the "number of books in home" variable (χ^2=326.14, p=.000). It was determined that five branches occurred at the starting node of this variable in question. It is seen that students who have books between "0-10" at home gather in Node 1. Students in this group constitute 21.8% of the entire data set, and it was determined that the majority of these students (73%) were unsuccessful. It is seen that students who have books between "11-25" at home gather in Node 2. It was determined that the majority of these students (61.3%) were unsuccessful. It is seen that the students who have "26-100" books at home gather in Node 3. It was determined that the majority of these students (59%) were successful. It is seen that students who have books between "101-200" in their home gather in Node 4. It was determined that the majority of these students (66.9%) were successful. It is seen that the students, who have more than 200 books in their homes, are gathered at Node 5. It was determined that the majority of these students (75.8%) were successful. It is seen that students are significantly more successful as the number of books in their homes increases.

It was determined that the variable that best explains the science literacy of the students who have books between "0-10" at home is the item "How informed are you about the use of genetically modified organisms?", which is related to environmental awareness (χ^2=62.86, p=.000). Two branches occurred at the first node to the variable in question. It is seen that the students who responded to this item as "I have never heard of this” and “I have heard about this but I would not be able to explain what it is really about” gathered in Node 6. It was determined that the majority of these students (90.7%) were unsuccessful. It is seen that the students who answered the above item as “I know something about this and could explain the general issue” and “I am familiar with this and I would be able to explain this well” gathered in Node 7. It was determined that the majority of these students (62.5%) were unsuccessful.

It was determined that the variable that best explains the science literacy of students who do not have knowledge about the use of genetically modified organisms is the item “Do you think problems associated with the use of genetically modified organisms will improve or get worse over the next 20 years?”, which is related to environmental optimism (χ^2=19.34, p=.000). According to the variable in question, two branches occurred in the sixth node. It is observed that the students who responded to this item in the form of “improve” and “stay about the same” were collected in Node 20. It was determined that almost all of these students (99.2%) were unsuccessful. It is seen that the students who answered “get worse” to the above item gathered at Node 21. It was determined that the majority of these students (82.9%) were unsuccessful. After these two branches, the branching is finished before a new node is formed.

It was determined that the variable that best explains the science literacy of the students who have knowledge about the use of genetically modified organisms is the item “This school year, approximately how many hours per week do you spend learning in addition to your required school schedule in school science?” (χ^2=40.72, p=.000). According to the variable in question, two branches occurred in the seventh node. The time that students who answered this item allocate to learning for school science varies between 0-800 minutes. It is seen that students with 200 minutes or less of time spent on learning congregate at Node 22. It was determined that the majority of these students (77.9%) were unsuccessful. Students with more than 200 minutes of learning each week for the school science seem to congregate at Node 23. It was determined that the majority of these students (52.4%) were successful. It is seen that students are significantly more successful when the time allocated to learning increases. After these two branches, the branching is finished before a new node is formed.

It was determined that the variable that best explains the science literacy of students who have “11-25” books at home is the item “Do you think problems associated with the nuclear waste problem will improve or get worse over the next 20 years?” (χ^2=71.36, p=.000). According to the variable in question, two branches occurred in the second node. It is seen that the students who responded to this item as “improve” and “stay about the same” gathered in Node 8. It was determined that the majority of these students (84%) were unsuccessful. It is seen that the students who answered “get worse” to the above item gathered at Node 9. It was determined that the majority of these students (52.4%) were unsuccessful.

It was determined that the variable that best explains the science literacy of students who approach the nuclear waste problem more optimistically is the item “How many of tablet computers are there at your home” (χ^2=14.91, p=.000). According to the variable in question, two branches occurred in the eighth node. It is seen that students who do not have any tablet computers at home gather in Node 24. It was determined that the majority of these students (94.3%) were unsuccessful. It is seen that students who have at least one tablet computer at home gather in Node 25. It was determined that the majority of these students (75.8%) were unsuccessful. After these two branches, the branching is finished before a new node is formed.

It was determined that the variable that best explains the science literacy of students who think that the nuclear waste problem will get worsen is the item “In a normal, full week at school, how many class periods are you required to attend in total?” (χ^2=52.35, p=.000). According to the variable in question, four branches occurred in the ninth node. The number of class periods attended by students who answered this item in a week varies between 10-60 hours. It is seen that students with 39 and less than 39 class periods in a week gather in Node 26. It was determined that the majority of these students (67.1%) were unsuccessful. It is seen that the students who have 40 class periods per week at the school gather at Node 27. It was determined that the majority of these students (60.3%) were successful. It is seen that the students whose number of class periods is between 40 and 45 per week are gathered in

Node 28. It was determined that the majority of these students (74.5%) were unsuccessful. It is seen that students with more than 45 class periods in a week gather in Node 29. It was determined that the majority of these students (55.7%) were successful. It is seen that the students who have 40 hours of class periods and over 45 hours in a week at school are more successful. However, it is also seen that the number of class periods more than 40 does not increase the success anymore. After these four branches, the branching is finished before a new node is formed.

It was determined that the variable that best explains the science literacy of students who have between "26-100" books at home is the item "In a normal, full week at school, how many class periods are you required to attend in total?" (χ^2=92.85, p=.000). According to the variable in question, four branches occurred in the third node. It is seen that students with 39 and less than 39 class periods in a week gather in Node 10. It was determined that the majority of these students (58.4%) were unsuccessful. It is seen that the students who have 40 class periods per week at school gather in Node 11. It was determined that the majority of these students (72.1%) were successful. It is seen that students whose number of class periods between 40 and 45 per week at school are gathered in Node 12. It was determined that the majority of these students (65%) were unsuccessful. It is seen that students with more than 45 class periods per week at school gather in Node 13. It was determined that the majority of these students (58.2%) were successful. It is seen that the students who have 40 hours of class periods and more than 45 hours in a week at school are more successful. However, it is also seen that the number of class periods more than 40 does not increase the success anymore.

It was determined that the variable that best explains the science literacy of students who attend school for 39 class periods or less in a week, is the item "How informed are you about the use of genetically modified organisms?" (χ^2=18.82, p=.000). According to the variable in question, two branches occurred in the tenth node. It is seen that the students who responded to this item as "I have never heard of this" and "I have heard about this but I would not be able to explain what it is really about" gathered at Node 30. It was determined that the majority of these students (80.3%) were unsuccessful. It is seen that the students who answered the above item as "I know something about this and could explain the general issue" and "I am familiar with this and I would be able to explain this well" gathered at Node 31. It was determined that the majority of these students (51.4%) were successful. It is seen that students with better environmental awareness are significantly more successful. After these two branches, the branching is finished before a new node is formed.

It was determined that the variable that best explains the science literacy of the students whose class periods in a week is 40 at school is the "Students' expected occupational status" scale, which is related to the student's own life (χ^2=56.02, p=.000). The scores students get from this scale range from 16 to 89. Higher scores on the scale indicate higher levels of expected occupational status. According to the variable in question, four branches occurred at the eleventh node. It is seen that students who score 61 or less from this scale gather at Node 32. It was determined that the majority of these students (52.3%) were successful. It is seen that students who score between 61 and 70 on the scale gather at Node 33. It was determined that the majority of these students (84.2%) were successful. It is seen that students who get 71 points from the scale in question gather at Node 34. It was determined that the majority of these students (61.8%) were successful. It is seen that students who score above 71 on this scale gather at Node 35. It was determined that the majority of these students (80.3%) were successful. After these four branches, the branching is finished before a new node is formed.

It was determined that the variable that best explains the science literacy of the students whose class periods in a week is between 40 and 45, is the item "learning time (minutes per week) in science" (χ^2=18.75, p=.000). According to the variable in question, two branches occurred in the twelfth node. It is seen that students who spend 200 minutes or less on learning time gather at Node 36. It was determined that the majority of these students (89.1%) were unsuccessful. It is seen that students with more than 200 minutes of learning time each week for science class gather at Node 37. It was determined that the majority of these students (50.6%) were unsuccessful. After these two branches, the branching is finished before a new node is formed.

It was determined that the variable that best explains the science literacy of students who have more than 45 class periods per week at school is the item "Is there a computer you can use for school work in your home?" (χ^2=13.08, p=.000). According to the variable in question, two branches occurred at the thirteenth node. It is seen that students who have a computer at home that they can use for school work gather at Node 38. It was determined that the majority of these students (69.1%) were successful. In this way, it is seen that students who do not have a computer congregate at Node 39. It was determined that the majority of these students (73.9%) were unsuccessful. It is seen that students who have a computer at home that they can use for school work are significantly more successful.

It was determined that the variable that best explains the science literacy of students who have books between "101-200" at home is the item "When learning school science at school, how often activity that students are allowed to design their own experiments occur?" (χ^2=32.94, p=.000). According to the variable in question, three branches occurred in the fourth node. It is seen that the students who answered this item as "in all lessons" gathered in Node 14. It was determined that the majority of these students (66%) were unsuccessful. It is seen that the students who answered the item as "in most lessons" and "in some lessons" gathered in Node 15. It was determined that the majority of these students (67.8%) were successful. It is seen that the students who answered this item as "never or hardly ever" gathered in Node 16. It was determined that the majority of these students (79.3%) were successful.

It was determined that the variable that best explains the science literacy of the students who stated that they were allowed to design their own experiments in "in all lessons" at school is the item "How much do you disagree or agree with the statement that a good way to know if something is true is to do an experiment" (χ^2=12.56, p=.001). According to the variable in question, two branches occurred at the fourteenth node. It is seen that students who answered "I agree, I disagree, and I strongly disagree" about the importance of experimenting gathered in Node 40. It was determined that the majority of these students (86.2%) were unsuccessful. It is seen that the students who answered "strongly agree" to the importance of experimenting gathered in Node 41. It was determined that the majority of these students (61.9%) were successful. It is seen that students who give importance to experimentation are significantly more successful. After these two branches, the branching is finished before a new node is formed.

It was determined that the variable that best explains the science literacy of the students who stated that they were allowed to design their own experiments as "in most lessons" and "in some lessons" at school is the item "Do you think problems associated with the nuclear waste problem will improve or get worse over the next 20 years" related to environmental optimism (χ^2=16.45, p=.000). According to the variable in question, two branches occurred at the fifteenth node. It is seen that the students who responded to this item as "improve" and "stay about the same" gathered in Node 42. It was determined that the majority of these students (60%) were unsuccessful. It is seen that the students who answered "get worse" to the above item gathered in Node 43. It was determined that the majority of these students (76.9%) were successful. It is seen that students who think that this environmental problem will worsen in the last 20 years are significantly more successful. After these two branches, the branching is finished before a new node is formed.

It was determined that the variable that best explains the science literacy of the students who stated that they were allowed to design their own experiments as "never or hardly ever" at school is the item "learning time (minutes per week) in science" (χ^2=19.69, p=.000). According to the variable in question, two branches occurred at the sixteenth node. It is seen that students who spend 150 minutes or less on learning gather in Node 44. It was determined that the majority of these students (62.5%) were unsuccessful. It is seen that students with more than 150 minutes of learning time each week for science gather at Node 45. It was determined that the majority of these students (85.7%) were successful. It is seen that students are significantly more successful when the time they spend on learning increases. After these two branches, the branching is finished before a new node is formed.

It was determined that the variable that best explains the science literacy of students who have books "more than 200" at home is the item "In a normal, full week at school, how many class periods are you required to attend in total?" (χ^2=35.22, p=.000). According to the variable in question, three branches occurred in the fifth node. It is seen that students with 39 and less than 39 class periods in a week gather in Node 17. It was determined that the majority of these students (51.2%) were unsuccessful. It is seen that students who have 40 class periods per week at school gather in Node 18. It was determined that the majority of these students (86.8%) were successful. It is seen that students who have more than 40 class periods in a week at school gather in Node 19. It was determined that the majority of these students (62.5%) were successful. It is seen that the number of class periods more than 40 does not increase the success more.

It was determined that the variable that best explains the science literacy of the students whose number of class periods in a week is 39 or less is the item "Are there technical reference books in your home" (χ^2=7.66, p=.006). According to the variable in question, two branches occurred at the seventeenth node. It is seen that students who have technical reference books at home gather in Node 46. It was determined that the majority of these students (64.3%) were successful. It is seen that students who do not have such a technical reference book gather in Node 47. It was determined that the majority of these students (80%) were unsuccessful. It is seen that students who have technical reference books at home are significantly more successful. After these two branches, the branching is finished before a new node is formed.

It was determined that the variable that best explains the science literacy of students whose number of class periods in a week is 40 at school is the item "My parents encourage me to be confident" (χ^2=15.50, p=.000). According to the variable in question, two branches occurred at the eighteenth node. It is seen that students who answered "strongly disagree" about receiving parental encourage gathered in Node 48. It was determined that the majority of these students (66.7%) were unsuccessful. It is seen that the students who answered the item as "disagree, agree, and strongly agree" gathered in Node 49. It was determined that the majority of these students (88.6%) were successful. After these two branches, the branching is finished before a new node is formed.

It was determined that the variable that best explains the science literacy of the students whose number of class periods in a week is more than 40 hours at the school is the item "Identify the better of two explanations for the formation of acid rain" related to science self-efficacy (χ^2=18.04, p=.000). According to the variable in question, two branches occurred at the nineteenth node. It is seen that the students who answered this qualification as "I could do this easily" gather in Node 50. It was determined that the majority of these students (92.6%) were unsuccessful. It is seen that the students who answered this qualification as "I could do this with a bit of effort", "I would struggle to do this on my own", and "I couldn't do this" gather in Node 51. It was determined that the majority of these students (59.5%) were unsuccessful. It is seen that students who have science self-efficacy and say "I could do this easily" are significantly more successful. After these two branches, the branching is finished before a new node is formed.

In addition, the gain values of the obtained nodes are given in Table 3 in order to determine which nodes (roots) are the best to classify successful students in the study and to reveal which of these nodes give more information.

Table 3
Gain Values for Success Status

Node	Node		Gain		Ratio of correct answer	Index
	n	%	n	%		
50. Node	27	0.9	25	1.7	92.6	189.5
49. Node	176	5.8	156	10.5	88.6	181.4
45. Node	105	3.4	90	6	85.7	175.5
35. Node	199	6.5	169	11.3	84.9	173.8
33. Node	95	3.1	80	5.4	84.2	172.4
43. Node	108	3.5	83	5.6	76.9	157.3
38. Node	68	2.2	47	3.2	69.1	141.5
46. Node	28	0.9	18	1.2	64.3	131.6
41. Node	21	0.7	13	0.9	61.9	126.7
34. Node	76	2.5	47	3.2	61.8	126.6
27. Node	290	9.5	175	11.7	60.3	123.5
29. Node	70	2.3	39	2.6	55.7	114.0
23. Node	212	6.9	111	7.4	52.4	107.2
32. Node	149	4.9	78	5.2	52.3	107.2
31. Node	148	4.8	76	5.1	51.4	105.1
37. Node	77	2.5	38	2.5	49.4	101.0
51. Node	37	1.2	15	1.0	40.5	83.0
42. Node	35	1.1	14	0.9	40.0	81.9
44. Node	16	0.5	6	0.4	37.5	76.8
48. Node	6	0.2	2	0.1	33.3	68.2
26. Node	143	4.7	47	3.2	32.9	67.3
39. Node	23	0.8	6	0.4	26.1	53.4
28. Node	98	3.2	25	1.7	25.5	52.2
25. Node	132	4.3	32	2.1	24.2	49.6
22. Node	204	6.7	45	3.0	22.1	45.2
47. Node	15	0.5	3	0.2	20.0	40.9
30. Node	66	2.2	13	0.9	19.7	40.3
21. Node	129	4.2	22	1.5	17.1	34.9
40. Node	29	1.0	4	0.3	13.8	28.2
36. Node	46	1.5	5	0.3	10.9	22.2
24. Node	105	3.4	6	0.4	57	11.7
20. Node	119	3.9	1	0.1	0.8	1.7

When Table 3 was examined, it was determined that the 50th node was the most effective node in separating successful and unsuccessful students (n=25, 1.7%). This node consists of 27 students who have science self-efficacy among the students who have more than 200 books in their home and who have more than 40 class periods in a week and who state "I could easily identify the better of two explanations for the formation of acid rain" and these students are 92.6% correctly classified. In the study, the gain values were examined to determine the second most effective node and it was seen that it was the 49th node (n=156, 10.5%). This node consists of 176 students who received encourage from their parents for self-confidence, among the students who have more than 200 books at home and 40 class periods per week at school, and these students were classified correctly at a rate of 88.6%. It was determined that the third most effective node in explaining successful and unsuccessful students was the 45th node (n=90, 6%). Among the students who have 101-200 books in home, those who were not allowed to design their own experiments in any school science at school, consisted of 105 students, who spent more than 150 minutes on learning each week for the school science, and these students were classified correctly at a rate of 85.7%. In addition, it was observed that the 20th node was the least informative node in distinguishing students' achievements (n=1, 0.1%). This node consists of 119 students who have less than 10 books in the home, who do not have environmental awareness about the use of genetically modified organisms, but who think that "the problem of using genetically modified organisms will improve or remain the same", and these students were classified correctly at a rate of 1.7%.

Discussion and Conclusion

In this study, the variables explaining the science literacy of Turkish students were examined. According to these variables, it was determined that the most important variable explaining the science literacy of 15-year-old students in Turkey is "the number of books in home". According to this research, the percentage of successful students increased significantly as the number of books in home increased. Similar to this result, Kaya and Doğan (2017) examined the variables affecting science literacy according to PISA 2012 data in their study and compared them with other countries. As a result of the research, when the findings were examined, a significant relationship was found between the number of books in the students' homes and science literacy. In addition, a significant difference was observed between the status of students having poetry books and world classics at home and their average science literacy in all four countries studied. Kahraman and Çelik (2017) aimed to determine the personal and environmental factors that affect the success of students according to PISA 2012 results, and it was concluded that the number of books at home was effective in the success of science and reading skills. Karabay (2013) examined family and school characteristics that affect students' success throughout PISA applications. It was concluded that the number of books in students' homes was a statistically significant predictor for both PISA applications (2003, 2006, 2009) and application areas (reading skills, mathematics, and science literacy). In the study of Karweit and Wasik (1992), it was determined that the number of books in the students' homes has a strong effect on science literacy. In a different study, Aslanoğlu (2007) examined the factors related to the reading comprehension skills of 4th-grade students in Turkey, using the PIRLS 2001 student, teacher, and school questionnaire. When the family characteristics of the student were examined, it was concluded that the number of books in the house was the most important family characteristic variable. It is very important for students to gain reading comprehension skills in order to be successful in national and international exams. It is important for students to be able to have books at an early age and to be able to choose reading materials from a wide range during their school years in order to become good readers.

Environmental awareness and environmental optimism of students are also variables that explain science literacy. As a result of the research, similar results were obtained from the literature. Acar and Öğretmen (2012) examined student and school characteristics that affect the science literacy of Turkish students based on PISA 2006 data, and a significant relationship was observed between students' environmental awareness and science performance. Öztürk (2018) examined the variables related to environmental literacy that affect the science literacy of students from different socioeconomic levels based on PISA 2015 data. As a result of the research, a significant relationship was found between students' environmental awareness, environmental optimism, and science literacy. In addition, environmental awareness and environmental optimism differed significantly according to the socioeconomic level of the student. A significant relationship was found between environmental awareness, environmental optimism, and science literacy at all socioeconomic levels. Çelebi (2010) examined the student and school characteristics affecting the science literacy of 15-year-old students in Turkey, Canada, and Sweden in PISA 2006. It was determined that students with environmental awareness and responsibility for sustainable development develop better science literacy skills. However, it was observed that the more awareness and responsibility students have, the less optimistic they are about the future benefits of scientific and technological developments in solving environmental problems. As a result of this study, it is seen that successful students in groups with environmental awareness are more successful than successful students in groups with low environmental awareness. Likewise, it is seen that successful students in groups who think that environmental problems will worsen gradually are more successful than successful students in groups that approach environmental problems more optimistically.

The time students spend on learning science and the number of class periods they attend school in a week are other variables that explain science literacy. As the time students spend on learning science increases, the percentage of successful students also increases significantly. When the class periods in a week at the school were examined, it was observed that the percentage of successful students increased up to 40 class periods, but the increase in the number of class periods was not effective in increasing the percentage of successful students. As a result of the research, similar studies were obtained. Duman (2014) examined the secondary school 6th-grade students' motivation to learn science in terms of various variables, and it was concluded that the motivations of the students did not show a significant difference in terms of the number of science lesson hours per week. Anagün (2011) examined the effects of variables related to learning-teaching processes on students' science literacy within the scope of PISA 2006. When the results of the research were examined, it was seen that the variable that most affected the students' science literacy in terms of learning-teaching processes was the "time devoted to learning". Özer and Anıl (2011) examined the variables affecting students' science and mathematics achievement based on PISA 2006 data. It was concluded that the most important variable affecting students' science achievement is the "time devoted to learning" for science lessons. When the components of this latent variable were examined, it was observed that "time students spare for studying science lessons at school and doing their homework about these lessons", "time allotted to private lessons for science lessons outside of school hours" and "duration of science lessons attended at school", respectively. The results of the research clearly reveal that the increase in the time allocated to education affects success.

Another variable that explains science literacy is having facilities and educational resources provided at home. It was observed that students who have one or more tablet computers at home, who have a personal computer that they can use for school work, and who have a technical reference book are significantly more successful. The results of the study were similar to the literature. In their study based on PISA 2012 results, Kahraman and Çelik (2017) found that the number of computers at home affected students' science and reading skills, and Kaya and Doğan (2017) found that there was a significant relationship between the number of computers and mobile phones at home and science literacy. In Karabay's (2013) study, it was concluded that the student's having a computer at home and having a room of his own were a significant predictor of both PISA applications (2003, 2006, 2009) and application areas (mathematics, science, reading skills). In another study by Karabay (2012), it was found that the facilities provided to the students at home significantly predicted the PISA science literacy scores during the application periods (2003-2006-2009). Likewise, Çeçen (2015) concluded that the facilities provided at home are a significant predictor of science literacy for the PISA application periods (2003, 2006, 2009, 2012). In the study of Chiu (2007), the relationship between the characteristics of students in 41 countries and science literacy was examined. It was concluded that students with educational resources were more successful. Considering that one of the most important variables affecting science literacy is the opportunities provided to students at home and the educational resources that students have, it is clearly seen that enriching the educational environment at home is effective on the success of students.

Students' expected occupational status is one of the variables that affect science literacy. High scores obtained from the scale created to determine the occupational expectations of students in PISA 2015 indicate better levels of expected occupational status. In general, it was observed that the percentage of successful students increased significantly as the occupational expectation increased. As a result of the research, similar studies were obtained. In the study conducted by Can and Taylı (2014) to examine the career development of secondary school students, it was found that students who think of choosing high schools that will take students with exams and high scores have a higher career development level than students who think of choosing high schools that will enter without exams. In many studies in the relevant literature, it has been concluded that the higher the perceived academic achievement, the higher the professional maturity level and career development of the students (Acısu, 2002; Bacanlı & Sürücü, 2011; Bal, 1998; Lawrence & Brown, 1976). However, in some studies, it has been determined that there is no relationship between perceived academic success and professional maturity level (Powell & Luzzo, 1998; Sahranç, 2000). It is very important for students to have a goal in line with their interests and abilities in the education process. Students who do not have any goals and continue to the educational ladder will not contribute enough to themselves or society in their future lives.

Another variable that explains science literacy is students' science self-efficacy. In this study, it was observed that students with high science self-efficacy were significantly more successful. As a result of the research, similar results were obtained with the literature. Usta (2009) examined the affective factors affecting students' science literacy based on PISA 2006 data. As a result of the research, it was determined that the students' self-sufficiency in science had a direct effect on science literacy. The fact that they consider themselves sufficient ensures that their science performance increases. Similarly, in the study conducted on PISA 2006 data, it was concluded that the science self-efficacy of the students was effective in science literacy (Acar & Öğretmen, 2012; Çalışkan, 2008). In a study conducted by Areepattamannil et al. (2011) on 15-year-old students in Canada, it was observed that students' motivational beliefs such as self-efficacy have a positive effect on science achievement. In a study conducted by Sun et al. (2012) based on the results of PISA 2006 in Hong Kong, it was observed that students with high motivation and self-efficacy tended to show a high level of science achievement.

Parental support is also one of the variables that explain science literacy. In this study, it was observed that the success percentage of students who stated that they did not encourage from their families in terms of self-confidence was lower. As a result of the literature review, many studies have been found showing that the support of the family affects student success. However, it has been observed that the studies conducted are mostly related to the socioeconomic status and demographic characteristics of families. In the study of Satır (1996), it was observed that the academic success of the children of the families who are interested in their child, who create a working environment for their child, who honor him when he succeeds, encourage him to work tirelessly when he fails, is higher. In Malkoç's (1993) research, it was concluded that a large part of school success was achieved with the contribution of the family. In the study of Diaz (1989), it was determined that students who are at risk of failing in the classroom and who have low academic success do not have family support and attention. In addition, it was concluded that problems in the family and parental inconsistencies were also negative factors in school success. It has been observed in many studies that findings related to family support affect academic achievement (Epstein, 1991; Fan & Chen, 2001; Yıldırım, 2000).

Epistemological beliefs are the beliefs held by individuals about the process of defining, creating, and evaluating information in mind. In this study, a significant relationship emerged between scientific epistemological belief dimensions, acquiring knowledge, verifying knowledge, and scientific literacy. It has been observed that the success percentage of students who attach great importance to the verification of knowledge is higher. Similar results have been obtained in the literature. In Özbay's (2016) study examining the relationship between middle school students' science achievement and their epistemological beliefs and mental risk-taking tendencies, it was concluded that there is a relationship between epistemological beliefs, mental risk-taking tendency, and science achievement. Evcim et al. (2011) examined the relationship between 8th-grade students' epistemological beliefs, their ability to solve problems they encounter in daily life, and their academic achievements. As a result of the research, a significant relationship was determined between the epistemological beliefs of the students and their academic achievement in general. Cano (2005) found a significant relationship between epistemological beliefs and academic achievement in his study on Hispanic students at the secondary school level. Muis and Franco (2009), in their study examining the relationship between epistemological beliefs, learning approaches and achievement, concluded that epistemological beliefs affect learning approaches and success.

The learning-teaching environment created for students is one of the variables that explain science literacy. As a result of the literature review, it was concluded that the creation of inquiry-based learning environments for students positively affects their science achievement (Akpullukçu, 2011; Anagün, 2011; Atun, 2016; Duban, 2008; Keçeci, 2014). However, in this study, it was determined that the success percentage of students who were prepared an inquiry-based learning environment and were allowed to design their own experiments was lower. The effect of the learning-teaching environment created in the course on student achievement surprisingly gave different results with the literature. It is difficult to explain the reason for this situation. The types of questioning that teachers use in the lesson and how students perceive it should be investigated. When the literature is examined, the types of inquiry are divided into confirmation inquiry, structured inquiry, guided inquiry, and open inquiry. Some researchers state that it is necessary to use open inquiry to develop students' research and higher-order thinking skills, and that structured and guided inquiry is not sufficient (Berg et al., 2003; Chinn & Malthora, 2002). However, some argue that structured and guided inquiry both prevents the student from wasting time and reduces their fear of failing and not reaching the right result (Trautmann et al., 2004). When the results obtained from the literature are examined in general, it can be said that structured inquiry activities are not sufficient to develop students' high-level thinking skills. In many studies, the questioning method used by the teacher in the learning environment was not mentioned. As can be seen, both structured and open inquiry methods have advantages and disadvantages. The important thing is to apply the method appropriate to the level of the students in the learning environment. The questioning methods applied by the teachers in the classroom environment need to be examined in more detail. In addition, based on the item "my teacher allows me to design my own experiments" in the scale, it should be questioned what kind of learning environment successful and unsuccessful students expect in this regard.

Some adjustments can be made in line with the results obtained as a result of this research. It is seen that the variable that best explains students' science literacy is the number of books in home. Therefore, it is recommended to ensure that students have access to various reading materials at home and to improve other facilities provided at home. Parents should be more careful about making their children feel their love, support, and trust. It is suggested that the curricula should be reviewed and the quality of the time allocated to science courses should be given due importance as much as the quantity. The learning and teaching environment should be rearranged to increase students' science self-efficacy and epistemological beliefs. Families and teachers should raise awareness of students and give the necessary support to raise their professional expectations.

This research has some limitations in terms of handling only the data in the PISA 2015 Turkey sample. Researchers can also examine the PISA applications made in different years and examine the changes in the variables that affect success over the years. In addition, the variables affecting success in other countries can be examined and compared with the results obtained in Turkey. Furthermore, it is recommended to re-examine the classification results obtained for science literacy with different analysis methods.

Declarations

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: Secondary data were used in this study. Therefore, ethical approval is not required.

References

Acar, T., & Öğretmen, T. (2012). Çok düzeyli istatistiksel yöntemler ile 2006 PISA fen bilimleri performanslarının incelenmesi [Analysis of 2006 PISA Science Performance via Multilevel Statistical Methods]. *Eğitim ve Bilim, 37*(163), 178-189. http://egitimvebilim.ted.org.tr/index.php/EB/article/view/1040/346

Acısu, S. (2002). *Koruma altında olan lise ve meslek lisesi öğrencilerinin mesleki olgunluk düzeylerinin bazı değişkenlere göre incelenmesi [Vocational maturity level of high school and vocational high school students who are under the protection of SHÇEK]* (Thesis No. 113167) [Unpublished master's thesis, Gazi University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Akpullukçu, S. (2011). *Fen ve teknoloji dersinde araştırmaya dayalı öğrenme ortamının öğrencilerin akademik başarı, hatırda tutma düzeyi ve tutumlarına etkisi [The effect of inquiry based learning environment in science and technology course on the students? academic achivements, attitudes and retention level]*

(Thesis No. 286479) [Master's thesis, Dokuz Eylül University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Aksu, G., & Karaman, H. (2016). CHAID Analizi (Otomatik Ki-Kare Etkileşim Belirleme Analizi) [CHAID Analysis (Automatic Chi-Square Interaction Determination Analysis)]. C. O. Güzeller (Ed.), *Herkes için çok değişkenli istatistik [Multivariate statistics for everyone]* (pp. 245-266). Maya Akademi.

Anagün, Ş. S. (2008). *İlköğretim beşinci sınıf öğrencilerinde yapılandırmacı öğrenme yoluyla fen okuryazarlığının geliştirilmesi: Bir eylem araştırması [Improving scientific literacy levels of the fifth grade students by constructivist learning: An action research]* (Thesis No. 229235) [Doctoral dissertation, Anadolu University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Anagün, Ş. S. (2011). PISA 2006 sonuçlarına göre öğrenme-öğretme süreci değişkenlerinin öğrencilerin fen okuryazarlığına etkisi [The impact of teaching-learning process variables to the students' scientific literacy levels based on PISA 2006 results]. *Eğitim ve Bilim*, *36*(162), 84-102. http://egitimvebilim.ted.org.tr/index.php/EB/article/view/503

Areepattamannil, S., Freeman, J. G., & Klinger, D. A. (2011). Influence of motivation, self-beliefs, and instructional practices on science achievement of adolescents in Canada. *Social Psychology of Education*, *14*(2), 233-259. https://doi.org/10.1007/s11218-010-9144-9

Aslanoğlu, A. E., (2007). *PIRLS 2001 Türkiye verilerine göre 4. Sınıf öğrencilerinin okuduğunu anlama becerileriyle ilişkili faktörler [Factors related to the reading comprehension skills of 4th grade students according to data of PIRLS 2001 Turkey]* (Thesis No. 234226) [Doctoral dissertation, Ankara University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Atun, T. (2016). *Sorgulamaya dayalı fen öğretiminin 5. Sınıf öğrencilerinde öğrenmeye yönelik öz düzenleme becerileri gelişimine etkisi [Effect of inquiry based science teaching on self-regulated learning skills development in 5th grade students]* (Thesis No. 435232) [Master's thesis, Hacettepe University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Bacanlı, F., & Sürücü, M. (2011). İlköğretim öğrencilerinin kariyer gelişimleri ile ebeveyne bağlanmaları arasındaki ilişkilerin incelenmesi [An examination of the relationships between attachment to parent and career development of elementary school students]. *Türk Eğitim Bilimleri Dergisi*, *9*(4), 679-700. https://dergipark.org.tr/tr/pub/tebd/issue/26098/274983

Bal, P. N. (1998). *Ergenlik döneminde mesleki karar verme olgunluğunun incelenmesi ve mesleki karar verme eğitim programının etkisinin araştırılması [Investigation of professional decision making maturity in adolescence and investigation of the effect of vocational decision making education program]* (Thesis No. 81253) [Unpublished doctoral dissertation, Marmara University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Berg, C. A. R., Bergendahl, C. V. B., Lundberg, B., & Tibell, L. (2003). Benefiting from an open-ended experiment? A comparison of attitudes to, and outcomes of, an expository versus an open-inquiry version of the same experiment. *International Journal of Science Education*, *25*(3), 351–372. https://doi.org/10.1080/09500690210145738

Brown, G., & Micklewright, J. (2004). *Using international surveys of achievement and literacy: A wiev from the outside*. Montreal, Rome: UNESCO Institute for Statistics.

Can, A., & Taylı, A. (2014). Ortaokul öğrencilerinin kariyer gelişimlerinin incelenmesi [An examinatıon of the career developments of primary school students in the second stage]. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, *14*(2), 321-346. https://dergipark.org.tr/tr/download/article-file/16810

Cano, F. (2005). Epistemological beliefs and approaches to learning: Their change through secondary school and their influence on academic performance. *British Journal of Educational Psychology*, *75*(2), 203-221. https://doi.org/10.1348/000709904X22683

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education*, *86*(2), 175–218. https://doi.org/10.1002/sce.10001

Chiu, M. M. (2007). Families, economies, cultures, and science achievement in 41 countries: Country-school and student-level analyses. *Journal of Family Psychology*, *21*(3), 510–519. https://doi.org/10.1037/0893-3200.21.3.510

Çalışkan, M. (2008). *The impact of school and student related factors on scientific literacy skills in the programme for international student assessment-PISA 2006* (Thesis No. 176784) [Unpublished doctoral dissertation, Middle East Technical University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Çeçen, Y. (2015). *Sosyokültürel ve sosyoekonomik değişkenlerin PISA fen okuryazarlığını yordama gücünün yıllara göre incelenmesi [Examination of the predictive powers of socio-cultural and socio-economic variables for PISA science literacy by years]* (Thesis No. 426119) [Unpublished master's thesis, İstanbul Aydın University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Çelebi, Ö. (2010). *A cross-cultural comparison of the effect of human and physical resources on students' scientific literacy skills in the Programme for International Student Assessment (PISA) 2006* (Thesis No. 285726)

[Unpublished doctoral dissertation, Middle East Technical University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Diaz, S. L. (1989, April 27-May 1). *The home environment and puerto rican children's achievement: A Researcher's Diary*. [Conference presentation]. In the National Association for Education Conference. https://eric.ed.gov/?id=ED299341

Duban, N. (2008). *İlköğretim fen ve teknoloji dersinin sorgulamaya dayalı öğrenme yaklaşımına göre işlenmesi: Bir eylem araştırması [Conducting science and technology course through inquiry-based learning approach in primary education: An action research]* (Thesis No. 229237) [Doctoral dissertation, Anadolu University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Duman, B. (2014). *Ortaokul 6. sınıf öğrencilerinin fen bilimleri dersindeki motivasyon durumları [Motivation statement of 6th grade students at secondary school at science lesson]* (Thesis No. 370247) [Unpublished master's thesis, Mehmet Akif Ersoy University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Epstein, J. L. (1991). Effects on student achievement of teachers' practices of parent involvement. S. B. Silvern (Ed.), *Advances in reading/language research: A research annual. Literacy through family, community, and school interaction* (pp. 261–276). Elsevier Science/JAI Press.

Evcim, İ., Turgut, H., & Şahin, F. (2011). İlköğretim öğrencilerinin epistemolojik inanışlarıyla, günlük yaşam problemlerini çözebilme ve akademik başarı düzeyleri arasındaki ilişki [The relationships between epistemological beliefs of elementary students and their problem solving abilities and academic achievement levels]. *Gaziantep Üniversitesi Sosyal Bilimler Dergisi*, *10*(3), 1199-1220. https://dergipark.org.tr/tr/download/article-file/223397

Fan, X., & Chen, M. (2001). Parental involvement and students' academic achievement: A meta analysis. *Educational Psychology Review*, *13*(1), 1-22. https://doi.org/10.1023/A:1009048817385

Han, J., Pei, J., & Kamber, M. (2012). *Data mining: concepts and techniques*. Elsevier.

Kahraman, Ü., & Çelik, K. (2017). Analysis of PISA 2012 results in terms of some variables. *Journal of Human Sciences*, *14*(4), 4797-4808. https://doi.org/10.14687/jhs.v14i4.5136

Karabay, E. (2012). *Sosyo-kültürel değişkenlerin PISA fen okuryazarlığını yordama gücünün yıllara göre incelenmesi [Examination of the predictive powers of socio-cultural variables for PISA science literacy by years]* (Thesis No. 311756) [Unpublished master's thesis, Ankara University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Karabay, E. (2013). *Aile ve okul özelliklerinin PISA okuma becerileri, matematik ve fen okuryazarlığını yordama gücünün yıllara göre incelenmesi [Investigation of the predictive power of family and school characteristics for PISA reading skills, mathematics and science literacy by years]* (Thesis No. 349068) [Unpublished master's thesis, Gazi University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Karweit, N. L., & Wasik, B. A. (1992). A review of the effects of extra-year kindergarten programs and transitional first grades. *Center for Research on Effective Schooling for Disadvantaged Students*. https://files.eric.ed.gov/fulltext/ED357894.pdf

Kaya, V. H., & Doğan, A. (2017). Determination and comparison of Turkish student characteristics affecting science literacy in Turkey according to PISA 2012. *Research Journal of Business and Management*, *4*(1), 34-51. https://doi.org/10.17261/Pressacademia.2017.369

Keçeci, G. (2014). *Araştırma ve sorgulamaya dayalı fen öğretiminin öğrencilerin bilimsel süreç becerilerine ve tutumlarına etkisi [The effects of inquiry-based science teaching on students' science process skills and attitudes]* (Thesis No. 363082) [Unpublished doctoral dissertation, Fırat University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Lawrence, W., & Brown, D. (1976). An investigation of intelligence, self-concept, socioeconomic status, race and sex as predictors of career maturity. *Journal of Vocational Behavior*, *9*(1), 43-52. https://doi.org/10.1016/0001-8791(76)90005-1

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*(404), 1198-1202. https://doi.org/10.2307/2290157

Malkoç, G. (1993). Aile eğitimi ve eğitimde nitelik geliştirme [Family education and quality development in education]. *Eğitimde Arayışlar*, *1*, 13-14.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in teaching* (10th ed.). Prentice-Hall Inc.

Ministry of National Education. (2010). *PISA 2009 Ulusal Ön Rapor*. Ankara: Eğitim Araştırma ve Geliştirme Dairesi Başkanlığı. http://pisa.meb.gov.tr/eski%20dosyalar/wp-content/uploads/2013/07/PISA-2009-Ulusal-On-Rapor.pdf

Ministry of National Education. (2016). *PISA 2015 Ulusal Rapor*. Ankara: Milli Eğitim Bakanlığı Ölçme Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü. http://pisa.meb.gov.tr/eski%20dosyalar/wp-content/uploads/2014/11/PISA2015_UlusalRapor.pdf

Muis, K. R., & Franco, G. M. (2009). Epistemic beliefs: Setting the standards for self-regulated learning. *Contemporary Educational Psychology, 34*(4), 306-318. https://doi.org/10.1016/j.cedpsych.2009.06.005

Organisation for Economic Co-operation and Development. (2017). *PISA 2015 assessment and analytical framework: Science, Reading, Mathematics, Financial Literacy and Collaborative Problem Solving*. OECD Publishing. https://www.oecd-ilibrary.org/docserver/9789264281820-en.pdf?expires=1654375516&id=id&accname=guest&checksum=1DF0FCC252345A0AEB86D757CEE42ACF

Özbay, H. E. (2016). *Ortaokul öğrencilerinin akademik başarılarının bilimsel epistemolojik inançlar ve zihinsel risk alma davranışları ile ilişkisinin incelenmesi [Investigation of association among scientific epistemological beliefs, intellectual risk taking and science achievement of middle school students]* (Thesis No. 436956) [Doctoral dissertation, İnönü University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Özer, Y., & Anıl, D. (2011). Öğrencilerin fen ve matematik başarılarını etkileyen faktörlerin yapısal eşitlik modeli ile incelenmesi [Examining the factors affecting students' science and mathemathics achievement with structural equation modeling]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *41*, 313-324. http://www.efdergi.hacettepe.edu.tr/yonetim/icerik/makaleler/702-published.pdf

Öztürk, Ö. (2018). *Using PISA 2015 data to analyze how to scientific literacy of students from different socioeconomic levels can be predicted by environmental awareness and by environmental optimism* (Thesis No. 505282) [Unpublished master's thesis, İhsan Doğramacı Bilkent University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Powell, D., & Luzzo, D. A. (1998). Evaluating factors associated with the career maturity of high school students. *The Career Development Quarterly*, *47*(2), 145–158. https://doi.org/10.1002/j.2161-0045.1998.tb00548.x

Russell, M. K., & Airasian, P. W. (2012). *Classroom Assessment: Concepts and applications* (7th ed.). McGraw-Hill

Sahranç, Ü. (2000). *Lise öğrencilerinin olgunluk düzeylerinin denetim odaklarına göre bazı değişkenler açısından incelenmesi [A Study on some variables affecting career maturity levels of high school students depending on their locus of control]* (Thesis No. 98528) [Unpublished doctoral dissertation, Hacettepe University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Silahtaroğlu, G. (2013). *Veri madenciliği kavram ve algoritmaları [Data mining concepts and algorithms]*. Papatya Yayınevi.

Sun, L., Bradley, K. D., & Akers, K. (2012). A multilevel modelling approach to investigating factors impacting science achievement for secondary school students: PISA Hong Kong sample. *International Journal of Science Education*, *34*(14), 2107-2125. https://doi.org/10.1080/09500693.2012.708063

Şatır, S. (1996). *Özel Tevfik Fikret Lisesi öğrencilerinin akademik başarılarıyla ilgili anne- baba davranışları ve akademik başarıyı artırmaya yönelik anne-baba eğitim gereksinmelerinin belirlenmesi [Views of parents related to the academic success of the students of private Tevfik Fikret High School and determination of education requirements of parents]* (Thesis No. 51615) [Unpublished Master's Thesis, Ankara University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Trautmann, N., MaKinster, J., & Avery, L. (2004, April 1-3). *What makes inquiry so hard? (And why is it worth it?)* [Conference presentation]. In annual meeting of the National Association for Research in Science Teaching, Vancouver, BC, Canada. http://www.ei.cornell.edu/pubs/NARST_04_CSIP.pdf

Usta, H. G. (2009). *PISA 2006 sınav sonuçlarına göre Türkiye'deki öğrencilerin fen okuryazarlığını etkileyen faktörler [The factors that effect students' scientific literacy according to PISA 2006 in Turkey]* (Thesis No. 293199) [Master's Thesis, Ankara University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Yıldırım, İ. (2000). Akademik başarının yordayıcısı olarak yalnızlık, sınav kaygısı ve sosyal destek [Loneliness, exam anxiety and social support as predictors of academic success]. *Hacettepe Eğitim Fakültesi Dergisi*, *18*, 167-176. http://efdergi.hacettepe.edu.tr/yonetim/icerik/makaleler/1093-published.pdf