Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN:1309-6575

Kış 2020
Winter 2020

Cilt: 11- Sayı: 4
Volume: 11- Issue: 4

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 sorumluğu ilgili yazarlara aittir.

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

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

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)

Hakem Kurulu / Referee Board

Hakem Kurulu / Referee Board

*Ada göre alfabetik sıralanmıştır. / Names listed in alphabetical order.

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (Aralık 2020, Sayı: 11-4)

Journal of Measurement and Evaluation in Education and Psychology (December 2020, Issue: 11-4)

İÇİNDEKİLER / CONTENTS

Comparison of Data Mining Classification Algorithms on Educational Data under Different Conditions *

İlhan KOYUNCU **   Selahattin GELBAL ***

Abstract

The purpose of this study was to examine the performance of Naive Bayes, *k*-nearest neighborhood, neural networks, and logistic regression analysis in terms of sample size and test data rate in classifying students according to their mathematics performance. The target population was 62728 students in the 15-year-old group who were participated in the Programme for International Student Assessment (PISA) in 2012 from The Organisation for Economic Co-operation and Development (OECD) countries. The performance of each algorithm was tested by using 11%, 22%, 33%, 44% and 55% of each dataset for small (500 students), medium (1000 students) and large (5000 students) sample sizes. 100 replications were performed for each analysis. As the evaluation criteria, accuracy rates, RMSE values, and total elapsed time were used. RMSE values for each algorithm were statistically compared by using Friedman and Wilcoxon tests. The results revealed that while the classification performance of the methods increased as the sample size increased, the increase of training data ratio had different effects on the performance of the algorithms. The Naive Bayes showed high performance even in small samples, performed the analyzes very quickly, and was not affected by the change in the training data ratio. Logistic regression analysis was the most effective method in large samples but had a poor performance in small samples. While neural networks showed a similar tendency, its overall performance was lower than Naive Bayes and logistic regression. The lowest performances in all conditions were obtained by the *k*-nearest neighborhood algorithm.

Key Words: Artificial neural networks, educational data mining, *k*-nearest neighborhood, logistic regression, naive Bayes

INTRODUCTION

Data mining is used to discover hidden patterns and relationships that help decision making by processing large amounts of data (Bhardwaj & Pal, 2011). A wide variety of methods based on mathematical and statistical algorithms are used to predict, cluster, and reveal relationship networks in many disciplines. Data mining has its roots in machine learning, artificial intelligence, computer science, and statistics (Dunham, 2003). Data mining methods, which are used in a wide range from marketing to engineering, from health sciences to business, have started to be used to examine large and complex educational datasets that have been increasing rapidly with technological developments. Although data mining is applied to a large number of industries and sectors, its applications in the context of education are limited (Ranjan & Malik, 2007).

Predicting student success is the focus of many kinds of research in education. In particular, today, while technology is developing rapidly and gaining more importance in education, there are databases that contain many factors that affect student success. In addition to the course management systems that include rich educational data sources such as Blackboard and Moodle, data is collected at the student, teacher, school, regional and country level in large scale assessments such as Trends In International

* This study is a part of the doctoral dissertation "Comparison of data mining methods in predicting PISA mathematical achievements of students".

** Assist. Prof., Adiyaman University, Faculty of Education, Adıyaman-Turkey, ilhankync@gmail.com, ORCID ID: https://orcid.org/0000-0002-0009-5279
*** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, sgelbal@gmail.com, ORCID ID: orcid.org/0000-0001-5181-7262

To cite this article:
Koyuncu, İ., & Gelbal, S. (2020). Comparison of data mining classification algorithms on educational data under different conditions . *Journal of Measurement and Evaluation in Education and Psychology, 11*(4), 325-345. doi: 10.21031/epod.696664.

Received: 01.03.2020
Accepted: 12.11.2020

Mathematics And Science Study (TIMMS), Programme for International Student Assessment (PISA), and Progress In International Reading Literacy Study (PIRLS). It is increasingly getting important in recent years to predict and compare students' performances by analyzing large educational datasets. For this purpose, educational data mining (EDM) has emerged as an independent research area in recent years (Baker, 2010).

EDM is a new discipline that emerged in order to apply data mining techniques to educational data (Baker & Yacef, 2009; Huebner, 2013). It can be used in various areas of education, from the effectiveness of teaching programs to predict student success, from educational institutions to the performance of teachers. There are different definitions of EDM in the related literature. According to Baker and Yacef (2009), EDM focuses on the development of new methods to make discoveries from characteristics data obtained from educational settings. EDM is a scientific research area that uses these methods to understand better students and learning environments (Baker & Yacef, 2009). However, Huebner (2013) considers that such definitions are limited, EDM covers an extensive educational area, and the scope and definitions of this area will change with future studies.

Romero and Ventura (2007) stated that data mining in education is an iterative cyclical process consisting of hypothesis creation, testing, and development. In this process, educators and academic specialists have the responsibility to design, plan, and develop educational systems. The outputs (demographic data, course information, academic data, etc.) obtained by the students' use and interaction with these systems can be used in data mining for various purposes (clustering, classification, association, etc.). The useful information discovered can be used by both educators and students (Romero & Ventura, 2007).

Baker (2010) stated that a wide variety of popular methods used in educational data mining are classified under five main categories: Prediction, clustering, discovering relationships, discovery with models, and distillation of data to evaluate individuals. Prediction makes inferences about a single piece of the data by using the other variables making up the majority of the data. An example of this is the use of features such as anxiety, attitude, self-efficacy, etc., in the rest of the data in order to make inferences about students' mathematics performance. Classification of individuals or observations according to a certain categorical variable is one of the most basic prediction techniques in data mining (Baker, 2010). Some popular prediction algorithms are decision trees, logistic regression, support vector machines, artificial neural networks, Bayes algorithms, k-nearest neighborhood, and density estimators based on various kernel functions. In order to evaluate the accuracy of an estimator, criteria such as converted performance metrics based on the error matrix (precision, recall, F criterion, etc.), root mean square error (RMSE), Kappa (Cohen, 1960) concordance coefficient, area under the ROC curve (Egan, 1975) and error rates are used.

In order to test the performance of algorithms in data mining, data is divided into two parts: training and test data. In this method, initial analyses are performed using a specific part of a data set (training data), and a predictive model is created. In the next step, by making use of this model, the prediction is made for individuals or objects in the rest of the data (test data). The reason for testing the performances of methods in data mining in this way is to avoid biased estimates of model error rates. The other methods used for similar purposes are bootstrapping (Efron, 1983) and cross-validation (Lachenbruch & Mickey, 1968) techniques (Michie, Spiegelhalter & Taylor, 1994). However, selecting one-third (33%) of all data as a test dataset and the rest of the data (67%) as training data is often preferred and used mostly for large samples to test the performance of the algorithms. In many studies in the field of data mining, the effect of the train/test ratio (e.g. Brain & Webb, 1999; Çölkesen, & Kavzoglu, 2010; Foody, Mathur, Sanchez-Hernandez, & Boyd, 2006; Heilman, & Madnani, 2015; Shao, Fan, Cheng, Wu & Cheng, 2013; Tadjudin & Landgrebe, 1998; Tayeh et al., 2015) and sample size (e.g. Beleites et al., 2013; Chu et al., 2012; Figueroa, Zeng-Treitler, Kandula, & Ngo, 2012; Heydari, SS, & Mountrakis, 2018; Raudys & Pikelis, 1980; Wharton, 1984) on the performances of the algorithms were assessed. For example, Brain and Webb (1999) showed that when the amount of test data was increased, the error variance decreased, but there was no significant change in bias. Tadjudin and Landgrebe (1998) developed a robust parameter estimation method that reduces the effect of varying test data rates by stating that the limited amount of test data causes errors in classification performance. Foody et al. (2006) stated that even a

90% reduction in the rate of test data did not cause a decrease in some algorithms' performance. Heilman and Madnani (2015) found that increasing test data increased performance, but increasing sample size did not have the same effect. Shao et al. (2013) showed that the minimum rate of test data can be found for some methods. Çölkesen and Kayzoğlu (2010) found in their study that some methods show higher performance in small training sets than others. In the present study, the ideal amount of test and training data are examined for educational data.

In classification studies in the field of education, the performance of methods such as decision trees, support vector machines, logistic regression, neural networks, Bayes algorithms, *k*-nearest neighborhood are examined and compared (e.g., Bahadır, 2013; Barker, Trafalis & Rhoads, 2004; Berens, Schneider, Gortz, Oster, & Burghoff, 2019; Çırak, 2012; Dekker, Pechenizkiy & Vleeshouwers, 2009; Göker, 2012; Hamalainen & Vinni, 2006; Hamalainen & Vinni, 2011; Minaei-Bidgoli, Kashy, Kortemeyer & Punch, 2003; Osmanbegović & Suljić, 2012; Romero, Espejo, Zafra, Romero & Ventura, 2013; Romero, Ventura, Espejo & Hervas, 2008; Shahiri, Husain & Rashid, 2015; Sweeney, Lester, Rangwala, & Johri 2016; Şengür, 2013; Tepehan, 2011; Tezbaşaran, 2016; Tosun, 2007; Yurdakul & Topal, 2015). In addition, methods were compared according to the different number of categories of the dependent variable (Minaei-Bidgoli, Kashy, Kortemeyer & Punch, 2003; Nghe, Janecek & Haddawy, 2007), the data structure (Romero et al., 2008; Romero et al., 2013), amount of missing and noisy data (Hamalainen & Vinni, 2011) and sample sizes (Hamalainen & Vinni, 2006; 2011).

In the literature, in general, it can be seen that different results are obtained for different data structures. For example, in their study, Kotsiantis et al. (2003) compared some data mining methods; the Naive Bayes algorithm generally yielded better results than any other method. In the study conducted by Tosun (2007), artificial neural networks showed about 92% correct classification performance, while decision trees showed 86% accuracy. In the research conducted by Tepehan (2011) with PISA data, neural networks were as successful as logistic regression. Çırak (2012) found that the correct classification performance (66.1%) of logistic regression analysis was lower than the performance of artificial neural networks (70.16%). Similarly, Bahadır (2013) showed that the prediction performed with artificial neural networks was better with the logistic regression method. Göker (2012) compared many methods to develop a program for predicting students' success before taking an exam and used the Naive Bayes method, which has the highest correct classification rate (87.27%).

Minaei-Bidgoli et al. (2003) have shown that increasing the number of categories of the dependent variable causes significant performance differences in all mining methods, especially in Naive Bayes and k-nearest neighborhood methods. In their study, Nghe et al. (2007) showed that decision trees produce better results than Bayes networks for the different number of categories of the dependent variable. In the study conducted by Barker et al. (2004), when different training and test datasets of different years were combined for the same data structure, different techniques produced the same results, and neural network methods showed good performance when the data of previous years were used as a training set.

In their study, Hamalainen and Vinni (2006) showed that when more variables are added to the model, the support vector machines perform better in small samples; while the number of variables is less, Bayes algorithms show higher performance. Hamalainen and Vinni (2011), while Naive Bayes classifiers are effective for their accuracy in small samples, neural networks and nearest neighborhood classifiers require much larger samples. In another study, Osmanbegović and Suljić (2012) compared Naive Bayes (76.65%), decision trees (73.93%), and artificial neural network (71.20%) methods, and they found that neural networks method took a little time for training the algorithm while other methods did not.

In their study, Sweeney et al. (2016) analyzed students' versatile data with many data mining methods in order to estimate the attendance status of students and found the least erroneous results with Factorization Machines (FM), Random Forests (RF), the Personalized Linear Multiple Regression, and hybrid FM-RF methods. Tezbaşaran (2016) compared the generalized Hebb algorithm and principal component analysis results to confirm the data structure of a scale. She found that the two structures were very similar, and the error and fit indexes were very close to each other. Berens et al. (2019) showed that the AdaBoost algorithm, which combines regression analysis, neural networks, and decision trees, is effective instead of using a single algorithm in predicting school attendance status through longitudinal data of students attending at two German universities. In the present study, unlike the previous studies, we aimed to compare the performances of Naive Bayes, k-nearest neighborhood, logistic regression, and neural networks classifiers in terms of sample size and test data rate. Therefore, the general structure of these methods will be briefly explained.

One of the most used classification algorithms in data mining is the Naive Bayes method based on Bayes' theorem. This classifier performs comparable performance with decision trees and neural networks classifiers in predicting probabilities of class memberships. The classifier calls “naive” because of the assumption that any value of a property belongs to a class is independent of the probability that other properties' values belong to the same class (Han, Kamber, & Pei, 2011). While this classifier has advantages such as being simple, useful, easy to interpret, and resistant to complexity, it can be used in small data sets and applied to categorical and continuous data (from Gauss distribution). There are disadvantages, such as the fact that the assumption of conditional independence is difficult to provide, and in the categorical data, when the limits of classes are complex, it is difficult to estimate its power (Hamalainen & Vinni, 2011).

Another method most commonly used is the *k*-nearest neighborhood algorithm. This algorithm is mostly used for classification purposes besides estimation and prediction. The method is based on the principle of classifying a new sample according to its similarity with the samples in training data (Larose, 2004). The class in which the sample will be assigned can be the most common class among neighboring samples or a neighboring class distribution. The most important problems to be encountered in calculations are what will be the value of *k* and how to calculate the distance (*d*). Another question that may come to mind is how to weight the sample cases in the training set. This algorithm's advantages are that there are only two parameters (*k* and *d*) in training the model and classification. The classification performance is very well in some problems, and the classification is robust to the complexity and missing data. The most important disadvantage is that there are difficulties in choosing the distance function (*d*) and *k* value (Hamalainen & Vinni, 2011).

Artificial neural networks (ANN) are used to discover relationships and patterns in a data set using certain mathematical and statistical algorithms. As a result of training neural networks, guiding information is obtained in making certain decisions (Sivanandam, Sumathi, & Deepa, 2006). ANN is used effectively in almost every field, especially in computer sciences, engineering, cognitive sciences, neurophysiology, physics, biology, environmental science, and marketing. When applied in educational technologies, it can be problematic if there is not enough numerical data, and it is exactly not known how to train the model (Hamalainen & Vinni, 2011). ANN was developed using the structure of biological cell networks. Neural networks, a subject that has been studied since the 1940s, have been reported in the form of many network architectures in the literature because of the complexity of the structures of real nerve cells and inadequate understanding of their working principles (Sivanandam et al., 2006). Some of the advantages of ANN are that they can easily learn nonlinear boundaries, represent basically different types of classifiers, fully convert variables when they are not discriminatory, robust to complexity (noise), and update themselves with new data. Some disadvantages are that ANNs require more data than typical data sets in education. They are very sensitive to overfitting. They require numerical data, and categorical data should be quantitated (Hamalainen & Vinni, 2011).

Logistic regression analysis is one of the prediction and classification algorithms that are used more than many other data mining methods. This analysis method effectively predicts group memberships when the predicted variable is categorical, and the predictors are categorical, continuous, or a mixture of the two. Discriminant analysis and multiple regression methods seek answers to similar research problems in logistic regression. However, the logistic regression has no strict assumptions such as normality, linearity, homogeneity of variances, etc. (Cox & Snell, 1989; Tabachnick & Fidell, 2013). This analysis method proposed in the early 1960s (Cabrera, 1994) began to take place as a routine package in statistical software since the early 1980s (Peng, Lee, & Ingersoll, 2002). It has become a frequently used method in social sciences and education until today (Cabrera, 1994; Peng & So, 2002). The logistic regression analysis has become popular by means of its advantages, such as being effective in a wide variety of complex data sets and a lack of assumptions about the distribution of predictive variables. However, in order for the analysis to be effective, it is required that the predictors are well-chosen and have a theoretical basis, there are sufficient samples in variables and category distributions, there is a linear relationship between continuous predictors and logit of the predicted variable, there is no multicollinearity and extreme values, errors and observations are independent of each other. (Tabachnick & Fidell, 2013).

It is possible to come across many studies on the applicability and effectiveness of data mining methods on educational data in the last decade. These researches aim to predict and evaluate student performance in general and to determine the factors affecting performance. However, only a few of these studies addressed the impact of sample size and training data size on the performance of these algorithms, as well as the comparison of data mining algorithms. In addition, studies on EDM and related to Naive Bayes and *k*-nearest neighbor techniques (e.g., Göker, 2012; Yurdakul & Topal, 2015) are limited in Turkey. In the present study, it was aimed to make a comprehensive application by using the data received in PISA (2012) assessment for these deficiencies in the literature.

In addition, data mining techniques have been used in order to predict and classify students' PISA performance in recent studies (e.g., Aksu, & Guzeller, 2016; Bulut, & Yavuz, 2019; Gorostiaga, & Rojo-Álvarez, 2016; Güre, Kayri, & Erdoğan, 2020; Kiray, Gok, & Bozkir, 2015; Martínez-Abad, Gamazo, & Rodríguez-Conde, 2020; Tepehan, 2011). For example, Kiray, Gok, & Bozkir (2015) examined the factors influencing Turkish students' performances in TIMMS 1999 and PISA 2003/2006 studies. Similarly, Aksu and Güzeller (2016) found that CHAID analysis and J.48 decision tree methods in data mining effectively classify Turkish students participating in PISA 2012 study. Moreover, Gorostiaga, and Rojo-Álvarez (2016), proposed a feature selection method in predicting Spanish students' PISA 2009 performance by using data mining techniques in addition to logistic regression. Besides, Bulut and Yavuz (2019) developed "Rattle" which is a R package used to apply data mining with graphical representations by using PISA 2015 data. Martínez-Abad et al. (2020) found that as a data mining technique, decision trees were more effective in explaining inter-school variance when compared to hierarchical linear modeling for PISA 2015 Spanish data. Güre et al. (2020) the performances of multilayer perceptron and random forest methods of data mining in determining factors affecting students' PISA 2015 mathematics literacy. In the literature related to PISA and data mining, the efficiency of different methods in predicting or classifying students' success and development of new techniques or tools were investigated.

As education systems are evaluated worldwide by PISA studies, a careful and systematic way is followed at every stage of the data collection process. Therefore, at the end of each application, a large data pool with high reliability and validity is obtained in terms of measurement and evaluation processes. Since the data of PISA (2012) assessment is used in the present study, the results obtained for the methods are considered to be important for the theory and real-life practice. In addition, in order to increase the reliability of the results obtained from different performance criteria, different data sets were selected by putting with replacement method, and the analyzes were replicated 100 times. Thus, we aimed to obtain results with high precision on real data regarding the methods used in the area of educational data mining.

Purpose of the Study

The aim of this study is to examine the performance of Naive Bayes, k-nearest neighborhood, neural networks, and logistic regression analysis in terms of sample size and training data ratio in classifying students according to their PISA mathematics performance. In accordance with this purpose, the sub-goals are to test whether;

- The performances of algorithms vary for small, medium, and large sample sizes,
- The performances of algorithms vary for different test data ratios,
- There is also a common effect of different sample sizes and test data ratios,

- Some of these algorithms perform better/worse under different conditions or not.

For this purposes, it is sought to find answers to the following research problem: For sample sizes of 500, 1000, and 5000 students, do the performances of Naive Bayes, k-nearest neighborhood, multilayer perceptron methods of artificial neural networks, and logistic regression methods differ for the ratio of test data 11%, 22%, 33%, 44%, and 55% in predicting students' PISA mathematics achievement?

METHOD

Since it is aimed to determine and explain the performances of Naive Bayes, k-nearest neighborhood, artificial neural networks, and logistic regression algorithms under different conditions, the present study is fundamental research. In this type of studies, it is aimed to produce knowledge by conducting studies based on methodological analysis (Büyüköztürk, Çakmak-Kılıç, Akgün, Karadeniz & Demirel, 2015). Fundamental research aims to add new information to existing knowledge (Karasar, 2005). Research is also quantitative relational research in terms of examining the relationships between methods. Relational studies aim to seek, explain, and discover the relationships between quantitative variables (Fraenkel & Wallen, 2006).

Sample

The research population of the study is 15 years-old students from OECD countries.The samples representing the population for each country were selected by PISA practitioners through stratified random sampling. The total number of people participating in the PISA (2012) assessment from OECD countries is 295416 students. In this study, after the missing data, residual and extreme values were examined and extracted, the target population of 62728 students was obtained. Table 1 shows the distribution of students in the target population by OECD countries.

Table 1. Distribution of The Target Population by OECD Countries

Country	*f*	*%*	*Country*	*f*	*%*	*Country*	*f*	*%*
Australia	2982	4.75	Finland	2001	3.19	Mexico	6062	9.66
Austria	976	1.56	France	993	1.58	Holland	1054	1.68
Belgium	1754	2.80	UK	2647	4.22	Norway	1032	1.65
Canada	4910	7.83	Greece	1190	1.90	New Zeland	852	1.36
Switzerland	2558	4.08	Hungary	1088	1.73	Poland	1010	1.61
Chile	1480	2.36	Ireland	1237	1.97	Portugal	1210	1.93
Czech Republic	1339	2.13	Iceland	780	1.24	Slovakia	1072	1.71
Germany	833	1.33	Italy	7479	11.92	Slovenia	1269	2.02
Denmark	1614	2.57	Japan	1512	2.41	Sweden	977	1.56
Spain	5502	8.77	Korea	1242	1.98	Turkey	834	1.33
Estonia	1140	1.82	Luxemburg	1017	1.62	USA	1082	1.72
Total	62728	100.0						

In data mining, the sample to be used in analysis is expressed as 'medium' when it consists of 1000 subjects, 'small' when it has less than this value, and 'large' when it has more than this value (Michie, Spiegelhalter & Taylor, 1994). In the sample selection, the bootstrapping method recommended by Efron (1983) was used. Accordingly, the samples of the research are 500 (small), 1000 (medium), and 5000 (large) students selected randomly by putting with replacement from the target population. In this sample selection method, the probability of each individual being selected is equal. In order to obtain results with high precision regarding the performance of the methods studied, a total of 180 datafiles consisting of 100 datafiles each including a sample of 500 students, 50 datafiles each including a sample of 1000 students, and 30 datafiles each including a sample of 5000 students were created. As the sample size decreases, the reason why more data files were drawn from all the data is to avoid biased or erroneous generalizations and increase the representativeness of small samples. To prevent the fact that different researchers can obtain different results with the same datasets, the weighted average of analysis results obtained with these datasets were evaluated by considering standard deviations of 100 replications.

Data Collection Instruments

The data collection tools of this study are mathematics cognitive test developed to measure students' academic performance in PISA (2012) assessment and a student questionnaire prepared to evaluate the students with all their existing characteristics. PISA study is an assessment that examines 15-year-old students' knowledge and skills in mathematics, science, and reading in order to evaluate and compare education systems worldwide in three-year periods (OECD, 2014b). Mathematics cognitive test consists of change and relationships, quantities, distances and shapes, uncertainty and data, tasks, formulation, and interpretation subfields. The test items consist of a mixture of multiple-choice items and items that students create their own answers. In the student questionnaire, students were expected to fill in forms containing various information about themselves, their homes, schools, and learning experiences. Besides the student questionnaire, one of the questionnaires that some countries chose for their students is related to the students' familiarity with the information and communication technologies, and the other is related to students’ education processes that question whether they are in preparation for a career for their future or a break during their education process. The student questionnaire consisting of three forms has 53 items in two forms and 54 items in the other. While each of these forms used in the PISA assessment is answered by one-third of the students, there are also students who answer the two forms in addition to the common items in the forms (OECD, 2014a).

Data Collection Procedure

In this study, open-access data obtained by PISA practitioners (OECD, 2014a) were taken from the OECD's public database. Detailed information about the data collection process in PISA assessment can be found in PISA documents (see OECD, 2014a; 2014b).

Data Analysis

In this study, a systematic process was followed in preparing data for the analysis. Firstly, data from OECD countries was drawn from PISA student questionnaire data. The demographic variables and all variables related to mathematics were taken from this existing data file. Then, considering the PISA 2012 technical report published by OECD (2014b), variables consisting of the combination of other variables were taken, and the remaining variables were removed from the file. In the data obtained, all individuals containing missing data related to basic affective variables such as math anxiety and math self-efficacy were excluded from the data. Thus, out-of-school mathematics lessons, class size, basic and applied mathematics experience in school, familiarity with mathematical concepts, time devoted to mathematics lessons, and out-of-school working time consisted of completely missing data. The stratum variable was not interpreted similarly in every country, and in some countries, school type was added as a layer. In this case, when a particular sample is selected from all data, some cells of this independent variable remain empty, and this is especially problematic for logistic regression analysis. A similar situation is valid for the test language variable. For these reasons, when all the mentioned variables above are removed from the analysis and all missing data, and extreme values in the file are deleted, the target population consisting of 35 variables and 62728 students was obtained.

Although data mining algorithms work with a lot of variables, keeping the variables that do not contribute to the classification causes the analysis to take a lot of time and decrease the classification performance. For this purpose, variable (feature) selection, which is a data preprocessing process, is one of the important techniques frequently used in data mining (Blum & Langley, 1997; Liu & Motoda, 2001). Variable selection methods designed according to different evaluation criteria are generally divided into three categories as filtering, winding, and hybrid models (Liu & Yu, 2005). Models other than filtering models require an analysis method to define the significance of variables in classification.

In this study, since different analysis methods were compared, the filtering method, which allows sorting the variables according to gaining the information, was used without requiring an additional analysis method. The filtering method aims to select and evaluate the subset of variables based on the general characteristics of the data, without including any data mining method (Liu and Yu, 2005).

In this study, Information Gain Ranking Filter, Chi-Squared Ranking Filter, Gain Ratio Feature Evaluator, and Symmetrical Uncertainty Ranking Filter in WEKA Version 3.9.0 software (Hall et al., 2009) methods are used to select variables for the analysis. Information Gain Ranking Filter measures the information obtained by classes; Chi-Squared Ranking Filter calculates the Chi-square value according to the class; Gain Ratio Feature Evaluator measures the ratio obtained according to the class; Symmetrical Uncertainty Ranking Filter measures symmetric uncertainty by class and evaluates the importance order of a variable (Frank, Hall & Witten, 2016).

In the present study, the variable selection process was performed on the data belonging to the target population (N = 62728). As the dependent variable, the first of 5 plausible values (PV1MATH) corresponding to students' mathematics performance was used. Plausible values correspond to the ability distribution a student may have, based on the students' responses to the items, and are obtained by subtracting random values from the posterior probability distribution for the Ɵ ability values in the Item Response Theory (IRT) (Wu, 2005). In the simulation study conducted by Wu (2005), it was found that using any of the plausible values alone is sufficient to estimate the population parameters with high accuracy. Therefore, the first plausible value 'PV1MATH' variable was converted to a new variable with two categories that represent the students below and above the medium level (482) according to proficiency levels determined by PISA practitioners (OECD, 2014a). Then, by doing feature selection analyses, the top 10 variables that have the greatest contribution to the classification of students according to their mathematics performance were selected. Then, the first 10 variables that have the greatest contribution to the classification of students according to all filtering methods in terms of mathematics performance were selected. These variables are mathematics self-efficacy, mathematics self-concept, mathematics anxiety, economic, social and cultural status index, openness to the problem solving, country, father's education level (ISCED), mothers' education level (ISCED), teacher behavior: directing students, and calculator use. In this study, all analyzes were performed by using these variables.

After selecting the variables to be used in the analysis, the assumptions and prerequisites of the algorithms were checked. Although logistic regression (LR) analysis does not require any assumptions regarding the distribution of independent variables, the ratio of the number of individuals to the number of variables, the suitability of the expected frequencies, the moderate linear relationship between the continuous variables, the absence of missing and extreme values, and the sufficient model fit values are some preconditions for the analysis (Tabachnick and Fidell , 2013). Although the Naive Bayes (NB) algorithm is based on the conditional independence of all independent variables, this assumption is rarely provided, but this algorithm still yields good results (Hamalainen & Vinni, 2011).

In the *k*-nearest neighborhood algorithm (KNN), choosing the appropriate *k* value and *d* distance criteria are important requirements (Larose, 2004). One of the most used methods for selecting the most appropriate *k* value are taking the square root of the sample size of training data (Dunham, 2003). Some researchers suggest that it is difficult to make a definitive judgment, but they recommend to try values close to this value, to use odd and prime numbers, to use Bayes methods, and *k*-layered cross-validation (Aha, Kibler & Albert, 1991; Ghosh, 2006; Hall, Park & Samworth, 2008). In this study, the square root of the number of students in the training data is taken (Dunham, 2003), and the most appropriate *k* number is selected for each analysis with the *k*-fold cross-validation method (Frank, Hall, and Witten, 2016).

Although the selection of the number of layers is an important issue in the multilayer perceptron (MLP) algorithm of artificial neural networks, reasonable results are obtained in educational data when there is enough numerical data, and the model is well trained (Hamalainen & Vinni, 2011). In this study, in order to select an appropriate number of layers, the values 1 to 5 were tested as the number of layers for the rate of test data of 33%. More than 5 values are not tried because when the number of layers increases, the model becomes complicated, and the analyzes take a lot of time. The experimental design for the number of layers revealed that 3 gives the ideal results. In addition, Akpınar (2014) states that it will be sufficient to select 3 layers in the solution of many classification problems. Still, it will be useful to examine additional layers if necessary to save time. For these reasons, the number of layers was taken 3 for the multilayer perceptron as the artificial neural network algorithm used in this study.

In the present study, in order to determine the standard conditions that the analysis was performed, some important assumptions and prerequisites were checked, and the following results have been obtained.

- The sample size is sufficient.
- There are no missing and extreme values.
- Continuous variables do not show a significant deviation from the standard normal distribution.
- Variance and covariance matrices are not homogenous.
- Linear relationships between variables are at a low or medium level.
- There is no multicollinearity or singularity problem.
- Conditional independence assumption could not be achieved for the Naive Bayes algorithm.

After the data were prepared for analysis, for each algorithm and datafiles (180 files), the analyses were performed for critical test data ratios 11%, 22%, 33%, 44%, and 55%. Although selecting one-third (33%) of all data as test dataset and the rest of data (67%) as training data is often used in the related literature, we aimed to test the effect of different amounts of test and training data on the performance of algorithms for educational data. For this purpose, one-third of the ideal test data ratio (33%) used in the related literature was drawn from all data. This value was then added and subtracted from 33%, and test values 11%, 22%, 33%, 44%, and 55% were obtained. After the data were prepared for analysis, for each algorithm and datafiles (180 files), 100 replications were performed for a different rate of test data (11%, 22%, 33%, 44%, and 55%) in which training data were randomly selected in the 'Experiment' section of the WEKA Version 3.9.0 software. Therefore, the test data for every replication of each algorithm was selected randomly. A total of 10000 analyzes were carried out for the sample of 500 students (100 datafiles), 5000 for 1000 students (50 datafiles), and 3000 for 5000 students (30 datafiles), and the average of the accuracy rates and RMSE values were reported and interpreted together with the total elapsed times for each algorithm. Selecting different datafiles from whole data and making and averaging 100 replications is to reduce the possible biased and erroneous results that could stem from getting different results for different algorithms. In the analysis, IBM SPSS Statistics 23, Microsoft Office Excel 2016 and WEKA Version 3.9.0 software were used.

In this study, since individuals showed a balanced distribution to the categories of the dependent variable, the accuracy rate, root mean square error (RMSE) values, and total elapsed time of the models were used in the evaluation of the performances of algorithms. The accuracy rate gives the correct classification percentage of a classifier. RMSE is a standard measure of the difference between values estimated by predicted and actual values. It is also a standard measure of accuracy rate that takes into account errors and allows to compare models.

In data mining, hypothesis testing is used to compare different methods and select the method with the least errors. For this purpose, when the assumptions of parametric analyzes are satisfied, the most preferred method is to use the t or F test. In this study, since the RMSE values used in the statistical comparison of methods did not meet the assumptions of parametric methods, the Friedman test was used to compare these values. Binary comparisons of the methods were made with the Wilcoxon Signed Ranks test.

RESULTS

The findings obtained for the data mining algorithms under different conditions with respect to different evaluation criteria are given in Table 2.

Table 2. Performances of Data Mining Techniques under Different Conditions.

Sample size	Percent of test data	NB			LR			MLP			KNN		
		Accuracy (%)	*RMSE*	*Time (sec)*	*Accuracy (%)*	*RMSE*	*Time (sec)*	*Accuracy (%)*	*RMSE*	*Time (sec)*	*Accuracy (%)*	*RMSE*	*Time (sec)*
500	11	75.93	0.42	0.00	75.30	0.42	0.04	72.80	0.48	0.72	71.81	0.48	0.03
	22	75.87	0.42	0.00	74.69	0.43	0.03	72.62	0.48	0.59	71.44	0.48	0.02
	33	75.78	0.42	0.00	73.94	0.44	0.03	72.27	0.48	0.50	71.05	0.48	0.02
	44	75.66	0.42	0.00	72.97	0.45	0.03	71.97	0.49	0.42	70.55	0.48	0.02
	55	75.41	0.43	0.00	71.62	0.47	0.02	71.56	0.49	0.35	70.06	0.48	0.01
1000	11	76.35	0.42	0.00	76.93	0.40	0.07	73.92	0.45	1.32	72.11	0.47	0.10
	22	76.27	0.42	0.00	76.70	0.40	0.06	73.65	0.46	1.16	71.97	0.47	0.09
	33	76.16	0.42	0.00	76.37	0.41	0.05	73.42	0.47	0.98	71.78	0.48	0.07
	44	76.04	0.42	0.00	75.88	0.41	0.04	73.07	0.47	0.82	71.57	0.48	0.06
	55	75.94	0.42	0.00	75.15	0.42	0.04	72.66	0.48	0.66	71.18	0.48	0.05
5000	11	76.60	0.42	0.00	78.30	0.39	0.36	76.36	0.41	6.54	74.52	0.46	2.05
	22	76.60	0.42	0.00	78.25	0.39	0.30	76.20	0.41	7.44	74.34	0.46	1.71
	33	76.58	0.42	0.00	78.19	0.39	0.25	76.04	0.41	4.93	74.14	0.47	1.50
	44	76.53	0.42	0.01	78.10	0.39	0.24	75.77	0.42	4.13	73.83	0.47	1.17
	55	76.48	0.42	0.01	77.96	0.39	0.25	75.53	0.42	0.35	73.50	0.47	1.03

Note: NB: Naïve Bayes, LR: Logistic Regression, MLP: Multilayer Perceptron, KNN: *k*-Nearest Neighborhood, RMSE: Root mean squared error.

According to Table 2, it has been observed that the methods chosen for classifying students according to their PISA mathematics achievement generally show above average or high performance under different conditions. The accuracy rates for all methods range from 70.06 to 78.30. While the NB method showed the highest performance in the sample of 500 students, the LR method showed the highest performance in the samples of 1000 and 5000 students. The MLP method performed less than the NB and LR method in all conditions but higher than the KNN method. The results for comparing the performances of the methods were examined separately according to different evaluation criteria. In Figure 1, the change of the accuracy rates of the methods for different conditions is given.

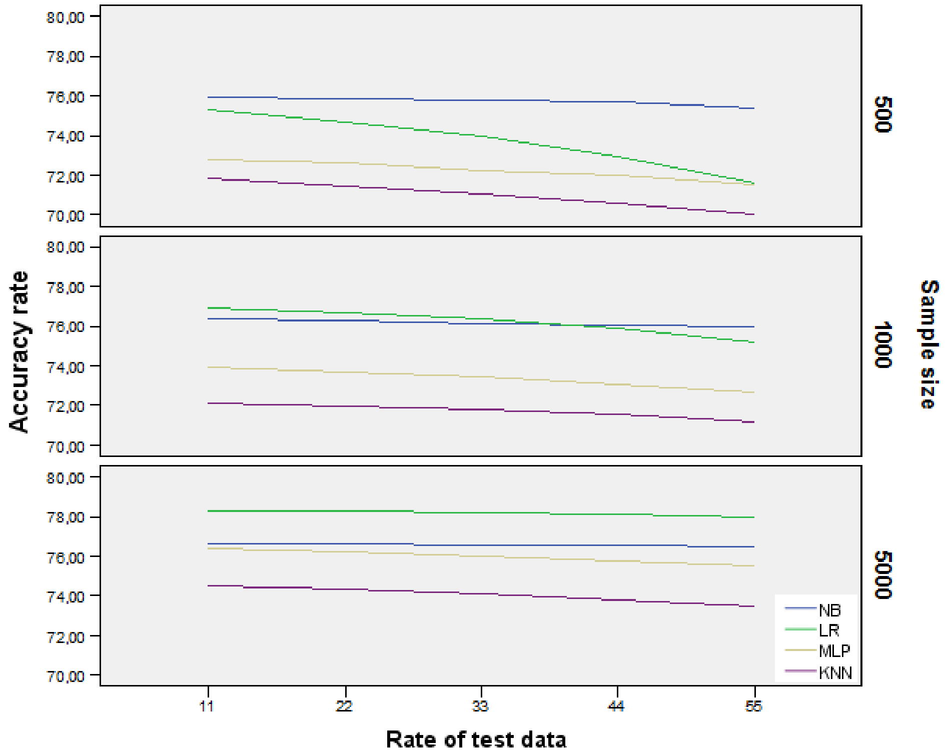


Figure 1. Change of Accuracy Rates of The Algorithms

When Figure 1 is examined, increasing the sample size leads to an increase in the classification performance of all methods, although much less in the NB method. In the samples of 500 and 1000 students, increasing the test data rate causes a significant decrease in the performance of the LR method. While the NB method is not affected by this, other methods are much less affected than the LR method. In the sample of 5000 students, the NB and LR methods are not affected by the increase in the rate of test data, while the MLP and KNN methods decrease slightly, as in other sample sizes. As a result, when the sample size is increased, the LR method is less affected by the change of the test data rate, while the NB method is not. MLP and KNN methods, on the other hand, show lower performance even if the sample size is increased, similarly being affected by increasing the test data rate. In Figure 2, the change of RMSE values of the methods for different conditions is given.

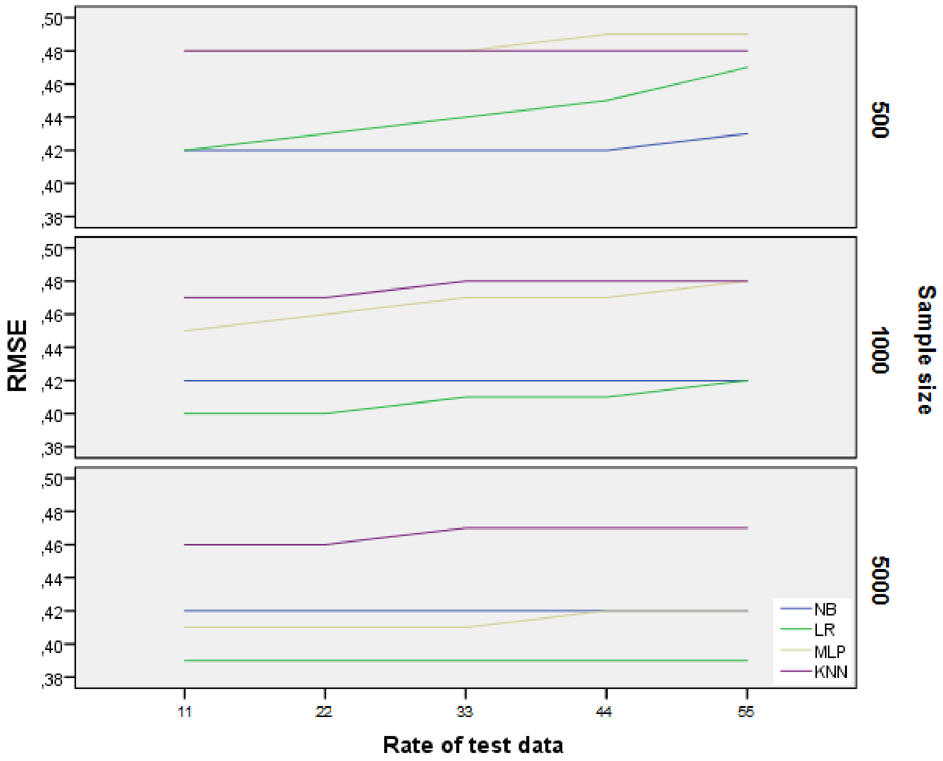


Figure 2. Change of RMSE Values of The Algorithms

For all methods, RMSE values range from approximately 0.39 and to 0.48. In small samples, the least erroneous estimations were made with NB and LR methods. According to Figure 2, the amount of error of the LR method increased significantly when the training data was reduced in small samples. In contrast, other methods were not significantly affected by this situation. In medium-sized samples, the error amount of the methods decreased compared to the small samples except the NB method. The least erroneous results in this sample size were obtained with the LR method. The decrease of the training data rate increased the error amount of the other methods except the NB method. In large samples, the estimation error amounts of the other methods have decreased except for the NB method. The NB method has approximately the same amount of error in all conditions. While increasing the test data rate does not affect the error amount of LR method in large samples, the error amount of MLP and KNN methods increased. The differentiation in RMSE values, which allow comparison of methods under different conditions across different methods, was analyzed by the Friedman test and binary comparisons of the methods were performed with the Wilcoxon test. The results are given in Table 3.

Table 3. Statistical comparison of data mining techniques under different conditions.

Sample size	Percent of test data	Test statistics (Friedman)			Multiple comparisons (Wilcoxon)**
		Chi-Square	*df*	*p*	
500	11	257.251*	3	0.000	1<3, 1<4, 2<3, 2<4
	22	265.013*	3	0.000	1<2, 1<3, 1<4, 2<3, 2<4
	33	275.340*	3	0.000	1<2, 1<3, 1<4, 2<3, 2<4, 3<4
	44	284.642*	3	0.000	1<2, 1<3, 1<4, 2<3, 2<4, 4<3
	55	271.014*	3	0.000	1<2, 1<3, 1<4, 2<3, 2<4, 4<3
1000	11	149.705*	3	0.000	2<1, 1<3, 1<4, 2<3, 2<4, 3<4
	22	149.149*	3	0.000	2<1, 1<3, 1<4, 2<3, 2<4, 3<4
	33	146.351*	3	0.000	2<1, 1<3, 1<4, 2<3, 2<4, 3<4
	44	144.013*	3	0.000	2<1, 1<3, 1<4, 2<3, 2<4, 3<4
	55	137.068*	3	0.000	1<3, 1<4, 2<3, 2<4
5000	11	88.729*	3	0.000	2<1, 3<1, 1<4, 2<3, 2<4, 3<4
	22	88.052*	3	0.000	2<1, 3<1, 1<4, 2<3, 2<4, 3<4
	33	87.632*	3	0.000	2<1, 3<1, 1<4, 2<3, 2<4, 3<4
	44	89.022*	3	0.000	2<1, 1<4, 2<3, 2<4, 3<4
	55	87.769*	3	0.000	2<1, 1<3, 1<4, 2<3, 2<4, 3<4

Note: 1: Naïve Bayes, 2: Logistic Regression, 3: Multilayer Perceptron, 4: *k*-Nearest Neighborhood, RMSE: Root mean squared error, df: Degree of freedom

*p<0.001

**p<0.0166 (Calculated based on Bonferroni correction)

According to Table 3, when the sample size increases, the LR method performs analysis with significantly less error than all methods. The NB method, on the other hand, provides significantly less erroneous estimations when the sample size decreases. The KNN method has more errors in medium and large samples compared to other methods at statistically significant level. When the test data rate is increased in small samples, the error amount of MLP method is significantly higher than other methods. In Figure 3, the change of the total elapsed time of the methods for the analysis under different conditions is given.

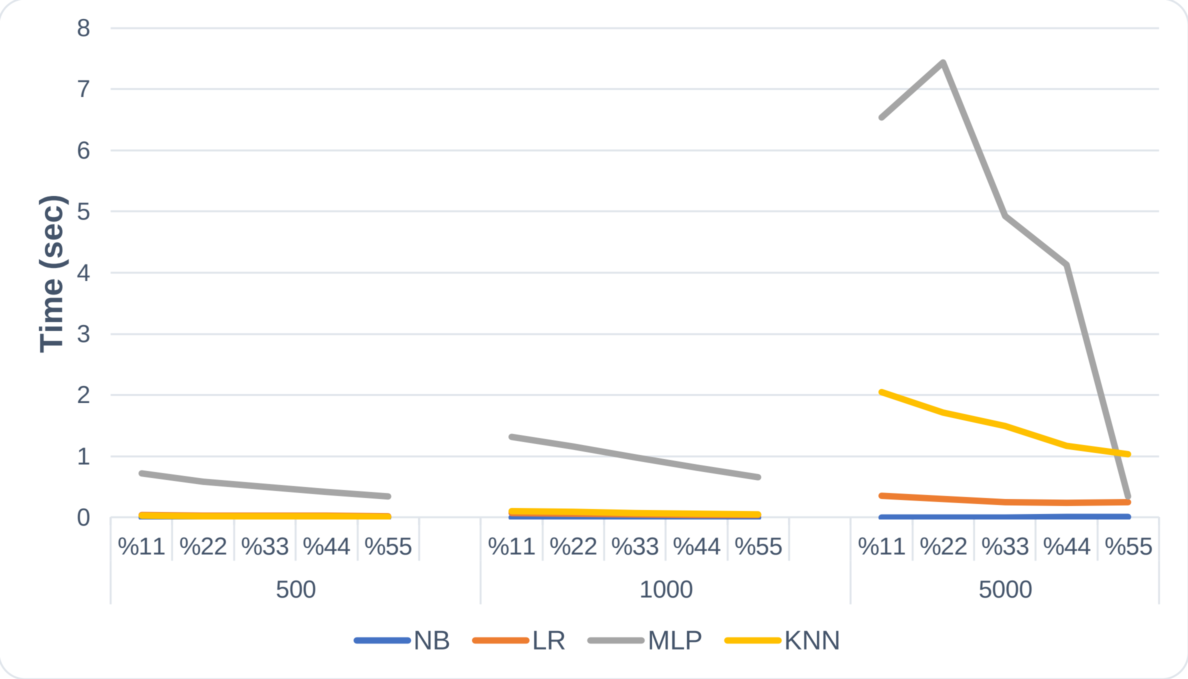


Figure 3. Total Elapsed Time for Each Analysis Under Different Conditions.

According to Figure 3, the NB method performs analyzes without taking any time in almost all conditions. In samples of 500 and 1000 students, LR and KNN methods operate in a much shorter time than MLP method. In small and medium sample sizes, LR and KNN methods carried out analysis in a much shorter time than MLP method. In large samples, KNN method takes more time than other methods for test data rate of 55%. The MLP method takes a lot of time when the test data rate is low, as the training data is high. Due to the k-fold cross-validation method used in the selection of the k value, in larger samples, the KNN method performed analyzes in much longer time than the LR method. However, since the total elapsed times are obtained under standard conditions on a computer with certain features, analysis can be completed in a shorter time on computers with more advanced features.

DISCUSSION and CONCLUSION

In this study, the performance of different data mining methods for different sample sizes and test data rates were compared on educational data in terms of accuracy rate, RMSE value, and total elapsed time for the analysis. It has been observed that the accuracy rates of the methods vary slightly for different conditions. This situation stems from the data selection and analysis procedure used in the present study. We selected 180 datasets from a huge data dataset of 62728 students by random selection with replacement and replicated each analysis 100 times. Therefore, the average of 10000 analyses for small samples, 5000 analyses for medium samples, and 3000 analyses for large sample sizes were evaluated. The results obtained seem to be close to each other due to these numerous amounts of the analyses. However, statistical hypothesis tests have shown that these seemingly small differences differ significantly.

In small sample sizes, high accuracy rates were obtained, less erroneous estimates were made, and the analyzes were completed in a very short time with the NB method compared to other methods. In addition, the NB method gives acceptable results even with a small amount of training data. In some studies, NB method has been shown to give better results than other methods in small samples (Göker, 2012; Hamalainen & Vinni, 2006; Hamalainen & Vinni, 2011, Kotsiantis et al., 2003; Osmanbegović & Suljić, 2012). However, Nghe et al. (2007) showed that decision trees produce better results than Bayes networks. Data structure might be a preliminary reason for this situation. Hence, it is very important to know which method is the best for a certain data type.

In the study, LR method showed higher performance in all conditions than MLP method. Although this result is different from some research results (Bahadır, 2013; Çırak, 2012; Tepehan, 2011), the most important reason for this situation is that the data structure is suitable for LR analysis. LR method produces less erroneous and higher accuracy estimates than other methods in medium and large samples. In the study conducted by Dekker et al. (2009), the LR method performed better in samples with similar size than the Bayes method.

After NB and LR methods, the highest accuracy rates were obtained by MLP and KNN methods, respectively. In the study of Romero et al. (2013), KNN method performed lower for numerical and categorical data compared to other classifiers. Similarly, in this study, the MLP method gave less erroneous results than the KNN method in medium and large samples. However, the opposite is true in small samples. This was due to the fact that the KNN method has a simpler statistical structure than the MLP method and that the selected k value was more stable in small samples in determining the closest neighborhood. In the MLP method, selecting the number of layers as three was effective in training the network, but in small samples, it yielded a high amount of error.

In this study, KNN method showed lower correct classification performance in all conditions than other methods. However, some studies have shown that the KNN method performs as well as ANN and LR methods (Minaei-Bidgoli et al., 2003; Yurdakul & Topal, 2015). Similarly, Shahiri et al. (2015) compared the studies published in international databases between 2002 and 2015 and found that NB method showed lower performance than KNN and ANN methods in terms of average performance. However, in this study, NB method showed higher classification performance, especially in small and medium-size samples. Some researchers have stated that it is not true to say that a classification method is best for different conditions and data structures (Romero et al., 2013; Shahiri et al., 2015). Barker et al. (2004), for example, made the classification of students who graduated in different years according to their graduation status and showed that different methods could be effective according to the structure of the data in different years.Barker et al. (2004), on the other hand, made the classification of students who graduated in different years according to their graduation status and showed that different methods could be effective according to the structure of the data in different years. For this reason, the results obtained from the present study have been interpreted within the framework of the structure of the data used and the analysis conditions. Since it is possible to obtain a different result with different data types (Romero et al., 2013), it is important to determine the structure of the data and choose the most appropriate method before the analyses.

Although the rate of test data is generally taken as one-third of all data in the related literature, it has been found that using a general valid rate may not be a proper approach. The results showed that the test data rate is closely related to the number of variables used, sample size, structure of the data and the structure of the method. However, except for the NB method, in general, increasing the rate of test data decreased the performance of the methods and increased the error of the results obtained. Therefore, as increasing the sample size increases classification performance and reduces the amount of error, it will be appropriate to use as much larger sample sizes as possible to achieve high performance from all methods. In many studies, it was found that different train/test ratios (e.g., Brain & Webb, 1999; Çölkesen, & Kavzoglu, 2010; Tadjudin & Landgrebe, 1998; Foody et al., 2006; Heilman, & Madnani, 2015; Shao et al., 2013; Tayeh et al., 2015) have different effects on the performance of the methods. For example, Brain and Webb (1999) showed that error variance decreases when the amount of test data is increased, but there is no significant change in the amount of bias. Similarly, Tadjudin and Landgrebe (1998) stated that the lack of test data caused errors in classification performance. However, Foody et al. (2006) stated in their study that even a 90% reduction in the rate of test data did not cause a decrease in the performance of some algorithms. Heilman and Madnani (2015) found that increasing test data increased performance, but increasing sample size did not have the same effect. Çölkesen and Kayzoğlu (2010) found in their study that some methods show higher performance in small training sets than others.

Limitations and Suggestions

In this study, although the analyses were performed with data for which the conditional independence assumption of the Naive Bayes method was not satisfied, acceptable results were obtained. This result has shown that, as stated by Hamalainen and Vinni (2011), Naive Bayes can perform well even if the conditional independence assumption is not met. In future studies, the acceptability of the results obtained under satisfying this assumption can be examined and compared with the performance of other methods. In the present study, it was seen that the *k* value to be selected for the *k*-nearest neighborhood method affects the classification performance. Accordingly, in other studies, different methods can be used to select the *k* value, or new methods can be developed. In the artificial neural networks method, since many parameters such as the number of layers, the number of nodes in layers, weightings affect the classification performance of the models, the effects of changes in these parameters on the performance of the method can be examined. The results obtained for logistic regression analysis and artificial neural network methods were obtained under the condition that homogeneity of variance-covariance matrices is not satisfied. Although these methods give effective results even when this assumption is violated, the classification performances of the methods can be evaluated and compared under the conditions in which the variance-covariance matrices are homogeneous. The results of the present study are also limited to the PISA 2012 data. For different data types, the performance of the algorithms can be compared in future studies. Besides, a simulation study under similar conditions could be done and compared with the results obtained with student data.

Similar to the results of the present study, it was found that different data types may yield different results (Romero et al., 2013). Therefore, identifying the structure of data and choosing the best analysis might be a solution to this issue. In addition, as a better solution to this problem, the procedure followed by (Göker, 2012; Yurdakul & Topal, 2015) can be used. As a two-step method, this procedure consists of selecting the method with the lowest error and then reporting the results obtained or performing further analysis with this method.

Using the Naive Bayes method in applications to be carried out under similar conditions will provide better results in a shorter time. Other methods may be preferred to the *k*-nearest neighborhood method to obtain higher classification performance under similar conditions. When the sample size is large, preferring Naive Bayes and logistic regression methods to multilayer perceptron will provide higher classification performance and time-saving.

REFERENCES

Aha, D. W., Kibler, D. & Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning 6*, 37-66.

Aksu, G., & Guzeller, C. O. (2016). Classification of PISA 2012 mathematical literacy scores using decision-tree method: Turkey sampling. *Education and Science, 41*(185), 101-122.

Akpınar, H. (2014). *Veri madenciliği veri analizi.* Papatya Yayınları, İstanbul.

Baker, R. S. J. (2010). Data mining for education. *International Encyclopedia of Education, 7*(3), 112-118.

Baker, R.S.J. & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3-17.

Bahadır, E. (2013). *Yapay sinir ağları ve lojistik regresyon analizi yaklaşımları ile öğretmen adaylarının akademik başarılarının tahmini* (Doktora tezi, Marmara Üniversitesi, İstanbul). Retrieved from http://tez2.yok.gov.tr/

Barker, K., Trafalis, T. & Rhoads, T. R. (2004). Learning from student data. In *Proceedings of the 2004 Systems and Information Engineering Design Symposium* (pp. 79-86). IEEE.

Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C., & Popp, J. (2013). Sample size planning for classification models. *Analytica Chimica Acta, 760*, 25-33.

Berens, J., Schneider, K., Gortz, S., Oster, S., & Burghoff, J. (2019). Early detection of students at risk - predicting student dropouts using administrative student data from German universities and machine learning methods. *Journal of Educational Data Mining, 11*(3), 1-41. https://doi.org/10.5281/zenodo.3594771

Bhardwaj, B. K. & Pal, S. (2011). Data mining: A prediction for performance improvement using classification. *(IJCSIS) International Journal of Computer Science and Information Security*, *9*(4), 136-140.

Blum, A. L. & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence, 97*(1), 245–271.

Brain, D., & Webb, G. (1999). On the effect of data set size on bias and variance in classification learning. In *Proceedings of the Fourth Australian Knowledge Acquisition Workshop, University of New South Wales* (pp. 117-128), December 5-6, Sydney, Australia.

Bulut, O., & Yavuz, H. C. (2019). Educational data mining: A tutorial for the" Rattle" package in R. *International Journal of Assessment Tools in Education*, *6*(5), 20-36.

Büyüköztürk, Ş., Çakmak-Kılıç, E., Akgün, Ö., Karadeniz, Ş. & Demirel, F. (2015). *Bilimsel araştırma yöntemleri.* Ankara: Pegem.

Cabrera, A. F. (1994). Logistic regression analysis in higher education: An applied perspective. *Higher Education: Handbook of Theory and Research, 10,* 225–256.

Chu, C., Hsu, A. L., Chou, K. H., Bandettini, P., Lin, C., & Alzheimer's Disease Neuroimaging Initiative. (2012). Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *Neuroimage, 60*(1), 59-70.

Cox, D. R. & Snell, E. J. (1989). *The analysis of binary data* (2nd ed.). London: Chapman and Hall.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20*(1), 37-46.

Çırak, G. (2012). *Yükseköğretimde öğrenci başarılarının sınıflandırılmasında yapay sinir ağları ve lojistik regresyon yöntemlerinin kullanılması* (Yüksek lisans tezi, Ankara Üniversitesi, Ankara). Retrieved from http://tez2.yok.gov.tr/

Çölkesen, I., & Kavzoglu, T. (2010). Farklı boyutta eğitim örnekleri için destek vektör makinelerinin sınıflandırma performansının analizi. In *Proceedings of III. Uzaktan Algılama ve Coğrafi Bilgi Sistemleri Sempozyumu* (pp. 161-170), 11 – 13 Ekim, Gebze, Kocaeli, Türkiye.

Dekker, G. W., Pechenizkiy, M. ve Vleeshouwers, J. M. (2009). Predicting students drop out: A case study. In *Proceedings of 2nd International Conference on Educational Data Mining* (pp. 41-50). Spain, Cordoba.

Dunham, M.H. (2003). *Data mining introductory and advanced topics*. Upper Saddle River, NJ: Pearson Education, Inc.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvements on crossvalidation. *J. Amer. Stat. Ass.*, *78*(382), 316–331.

Egan, J. P. (1975). *Signal detection theory and ROC analysis*. New York: Academic Press.

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, *12*(1), 8.

Foody, G. M., Mathur, A., Sanchez-Hernandez, C., & Boyd, D. S. (2006). Training set size requirements for the classification of a specific class. *Remote Sensing of Environment*, *104*(1), 1-14.

Fraenkel, J. R. & Wallen, N. E. (2011). *How to design and evaluate research in education* (6th ed.). New York: McGraw-Hill, Inc.

Frank, E., Hall M. A. & Witten, I. H. (2016). *The WEKA workbench: Online appendix for "Data mining: Practical machine learning tools and techniques"* (4th ed.). Morgan Kaufmann.

Ghosh, A. K. (2006). On optimum choice of k in nearest neighbor classification. *Computational Statistics and Data Analysis*, *50*(11), 3113-3123.

Gorostiaga, A., & Rojo-Álvarez, J. L. (2016). On the use of conventional and statistical-learning techniques for the analysis of PISA results in Spain. *Neurocomputing, 171*, 625-637.

Göker, H. (2012). *Üniversite giriş sınavında öğrencilerin başarılarının veri madenciliği yöntemleri ile tahmin edilmesi* (Yüksek lisans tezi, Gazi Üniversitesi, Ankara). Retrieved from http://tez2.yok.gov.tr/

Güre, Ö. B., Kayri, M., & Erdoğan, F. (2020). Analysis of factors effecting PISA 2015 mathematics literacy via educational data mining. *Education and Science*, *45*(202), 393-415.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10-18.

Hall, P., Park, B. U. & Samworth, R. J. (2008). Choice of neighbor order in nearest-neighbor classification. *The Annals of Statistics*, *36*(5), 2135-2152.

Han, J., Kamber, M. & Pei, J. (2011). *Data mining: concepts and techniques* (3rd ed.). MA, USA: Elsevier.

Hamalainen, W. & Vinni, M. (2006). Comparison of machine learning methods for intelligent tutoring systems. In *Proceedings of International Conference on Intelligent Tutoring Systems* (pp. 525-534). Springer Berlin/Heidelberg.

Hamalainen, W. & Vinni, M. (2011). *Classifiers for educational technology*. In C. Romero, S. Ventura, M. Pechenizkiy, R.S.J.d. Baker (Eds.), Handbook of educational data mining (pp. 54-74). CRC Press.

Heilman, M., & Madnani, N. (2015). The impact of training data on automated short answer scoring performance. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 81-85), June 4, Association for Computational Linguistics, Denver, Colorado.

Heydari, S. S., & Mountrakis, G. (2018). Effect of classifier selection, reference sample size, reference class distribution and scene heterogeneity in per-pixel classification accuracy using 26 Landsat sites. *Remote Sensing of Environment*, *204*, 648-658.

Huebner, R. A. (2013). A survey of educational data-mining research. *Research in Higher Education Journal*, *19*, 1-13.

Karasar, N. (2005). *Bilimsel araştırma yöntemi*. Ankara: Nobel Yayın Dağıtım.

Kiray, S. A., Gok, B., & Bozkir, A. S. (2015). Identifying the factors affecting science and mathematics achievement using data mining methods. *Journal of Education in Science, Environment and Health*, *1*(1), 28-48.

Kotsiantis, S. B., Pierrakeas, C. J. & Pintelas, P. E. (2003). Preventing student dropout in distance learning using machine learning techniques. In *Knowledge-Based Intelligent Information and Engineering Systems* (pp. 267-274). Springer Berlin/Heidelberg.

Lachenbruch, P. A. & Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, *10*(1), 1-11.

Larose, D. T. (2004). *K-nearest neighbor algorithm*. In Larose, D.T. and Larose, C.D. (Eds.), *Discovering knowledge in data: An introduction to data mining* (pp. 90-106). Hoboken, NJ, USA John Wiley and Sons, Inc.. https://doi.org/10.1002/0471687545.ch5.

Liu, H. & Motoda, H. (2001). *Feature extraction, construction and selection: A data mining perspective*. Boston: Kluwer Academic Publishers.

Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering, 17*(4), 491-502.

Martínez-Abad, F., Gamazo, A., & Rodríguez-Conde, M. J. (2020). Educational Data Mining: Identification of factors associated with school effectiveness in PISA assessment. *Studies in Educational Evaluation*, *66*, 100875. https://doi.org/10.1016/j.stueduc.2020.100875

Michie, D., Spiegelhalter, D.J. & Taylor, C.C. (1994*). Machine learning, neural and statistical classification*. Ellis Horwood Limited.

Minaei-Bidgoli, B., D.A. Kashy, G. Kortemeyer, & W. Punch. Predicting student performance: An application of data mining methods with an educational web-based system. In *Proceedings of 33rd Frontiers in Education Conference*, (pp. 13-18). Westminster, CO..

Nghe, N. T., Janecek, P. & Haddawy, P. (2007). A comparative analysis of techniques for predicting academic performance. In *Frontiers in Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports*, (pp. T2G-7). IEEE.

Organisation for Economic Co-operation and Development (2014a). *PISA 2012 results: What students know and can do - student performance in mathematics, reading and science* (Volume I, Revised edition). PISA, OECD Publishing.

Organisation for Economic Co-operation and Development (2014b). *PISA 2012 technical report.* PISA, OECD Publishing.

Osmanbegović, E. & Suljić, M. (2012). Data mining approach for predicting student performance. *Economic Review*, *10*(1), 3-12.

Peng, C.Y.J., Lee, K. L. & Ingersoll, G. M. (2002) An introduction to logistic regression analysis and reporting. *The Journal of Educational Research, 96*(1), 3-14. doi:10.1080/00220670209598786.

Peng, C. Y. J. & So, T. S. H. (2002). Logistic regression analysis and reporting: A primer. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, *1*(1), 31-70.

Ranjan, J. & Malik, K. (2007). Effective educational process: A data mining approach. *VINE, 37*(4), 502-515.

Raudys, S., & Pikelis, V. (1980). On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (3), 242-252.

Romero, C., Espejo, P. G., Zafra, A., Romero, J. R. & Ventura, S. (2013). Web usage mining for predicting final marks of students that use Moodle courses. *Computer Applications in Engineering Education*, *21*(1), 135-146.

Romero, C., Ventura, S., Espejo, P. G. & Hervás, C. (2008). Data mining algorithms to classify students. In *Proceedings of the 1st International Conference on Educational Data Mining* (pp. 8-17). Montréal, Québec, Canada.

Romero, C. & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135-146.

Romero, C. & Ventura, S. (2013). Data mining in education. *WIREs Data Mining Knowledge Discovery 3*(1), 12-27.

Shahiri, A. M., Husain, W. & Rashid, N. A. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, *72*, 414-422.

Shao, L., Fan, X., Cheng, N., Wu, L., & Cheng, Y. (2013). Determination of minimum training sample size for microarray-based cancer outcome prediction–an empirical assessment. *PloS one*, *8*(7), e68579. https://doi.org/10.1371/journal.pone.0068579

Sivanandam, S., Sumathi, S., & Deepa, S. (2006). *Introduction to neural networks using Matlab 6.0.* New Delhi: Tata McGraw-Hill Publishing Company.

Şengür, D. (2013). *Öğrencilerin akademik başarılarının veri madenciliği metotları ile tahmini* (Yüksek lisans tezi, Fırat Üniversitesi, Elazığ). Erişim adresi: http://tez2.yok.gov.tr/

Sweeney, M., Lester, J., Rangwala, H., & Johri, A. (2016). Next-term student performance prediction: A recommender systems approach. *JEDM | Journal of Educational Data Mining*, *8*(1), 22-51. https://doi.org/10.5281/zenodo.3554603

Tabachnick, B. G. & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson.

Tadjudin, S., & Landgrebe, D. (1998). *Classification of high dimensional data with limited training samples* (Report No. 56). West Lafayette, Indiana: Purdue University, School of Electrical and Computer Engineering. http://docs.lib.purdue.edu/ecetr/56

Tayeh, N., Klein, A., Le Paslier, M. C., Jacquin, F., Houtin, H., Rond, C., ... & Burstin, J. (2015). Genomic prediction in pea: effect of marker density and training population size and composition on prediction accuracy. *Frontiers in Plant Science*, *6*(941), 941. https://doi.org/10.3389/fpls.2015.00941

Tepehan, T. (2011). *Türk öğrencilerinin PISA başarılarının yordanmasında yapay sinir ağı ve lojistik regresyon modeli performanslarının karşılaştırılması* (Doktora tezi, Hacettepe Üniversitesi, Ankara). Retrieved from http://tez2.yok.gov.tr/

Tezbaşaran, E. (2016). *Temel bileşenler analizi ve yapay sinir ağı modellerinin ölçek geliştirme sürecinde kullanılabilirliğinin incelenmesi* (Doktora tezi, Mersin Üniversitesi, Mersin). Retrieved from http://tez2.yok.gov.tr/

Tosun, S. (2007). *Sınıflandırmada yapay sinir ağları ve karar ağaçları karşılaştırması: Öğrenci başarıları üzerine bir uygulama* (Yüksek lisans tezi, İstanbul Teknik Üniversitesi, İstanbul). Retrieved from http://tez2.yok.gov.tr/

Wharton, S. W. (1984). An analysis of the effects of sample size on classification performance of a histogram based cluster analysis procedure. *Pattern Recognition, 17*(2), 239-244.

Yurdakul, S. & Topal, T. (2015). Veri madenciliği ile lise öğrenci performanslarının değerlendirilmesi. *XVII. Akademik Bilişim Konferansında* sunulan bildiri. Anadolu Üniversitesi, Eskişehir.

Veri Madenciliği Sınıflandırma Algoritmalarının Farklı Koşullar için Eğitsel Bir Veride Karşılaştırılması

Giriş

Öğrenci başarısının yordanması eğitimde yapılan birçok araştırmanın odak noktasını oluşturur. Özellikle, teknolojinin hızla geliştiği ve eğitimde daha fazla önem kazandığı günümüzde öğrenci başarısını etkileyen birçok faktörü içinde barındıran veri tabanları bulunmaktadır. Blackboard ve Moodle gibi zengin eğitimsel veri kaynaklarını içeren ders yönetim sistemlerinin yanında, uluslararası düzeyde yapılan TIMMS (Uluslararası Matematik ve Fen Eğilimleri Araştırması), PISA (Uluslararası Öğrenci Değerlendirme Programı) ve PIRLS (Uluslararası Okuma Becerilerinde Gelişim Projesi) gibi çalışmalarda öğrenci, öğretmen, okul, bölge ve ülke düzeyinde bilgiler toplanmaktadır. Elde edilen eğitimsel içerikli veri yığınlarını analiz etmek ve öğrencileri karşılaştırarak başarılarını yordamak son yıllarda gittikçe önem kazanmaktadır. Bu amaçla, eğitsel veri madenciliği (EVM) son yıllarda bağımsız bir araştırma alanı olarak ortaya çıkmıştır (Baker, 2010).

EVM, veri madenciliği tekniklerini eğitim içerikli verilere uygulamak amacıyla ortaya çıkan yeni bir disiplindir (Baker ve Yacef, 2009; Huebner, 2013). Öğretim programlarının etkililiğinden öğrenci başarısının yordanmasına, eğitim kurumlarından öğretmenlerin performansına kadar eğitimin her alanında kullanılabilmektedir. İlgili alan yazında EVM ile ilgili farklı tanımlamalar mevcuttur. Baker ve Yacef (2009), EVM'yi, eğitim ortamlarından elde edilen kendine özgü verilerden keşifler yapmak amacıyla yeni metotların geliştirilmesini merkez alan, öğrencileri ve öğrenme ortamlarını daha iyi anlamak için bu metotları kullanan bilimsel araştırma alanı olarak tanımlamaktadır. Ancak, Huebner (2013) bu şekilde tanımlamaların sınırlı olduğunu, EVM'nin çok geniş bir alanı kapsadığını ve ileride yapılacak çalışmalarla birlikte bu alanın kapsamının ve tanımlarının değişeceğinin belirtmiştir.

Veri madenciliğinde bireylerin ya da gözlemlerin belirli bir kategorik değişkene göre sınıflandırılması en temel yordama tekniklerinden biridir (Baker, 2010). Bazı popüler yordama algoritmaları, karar ağaçları, lojistik regresyon, destek vektör makineleri, sinir ağları, Bayes algoritmaları, k-en yakın komşuluk ve çeşitli kernel fonsiyonlarına dayanan yoğunluk kestiricileridir. Bir kestiricinin doğruluğunu değerlendirmek amacıyla hata matrisine dayanan dönüştürülmüş performans değerlendirme ölçütleri (kesinlik, çağrışım, F ölçütü, vb.), Root mean square error (RMSE), Kappa (Cohen, 1960), ROC eğrisinin altında kalan alan (Egan, 1975) ve yordama hata oranları gibi ölçütler kullanılmaktadır.

Veri madenciliğinde algoritmaların performansını artırmak amacıyla veri öğrenme ve test verisi olmak üzere iki parçaya ayrılır. Bu metotta, bir veri setinin belirli bir bölümü kullanılarak ilk analizler gerçekleştirilir ve bir yordama modeli oluşturulur. Sonraki aşamada, elde edilen bu modelden yararlanılarak verinin kalan kısmındaki bireyler ya da nesneler için yordama işlemi gerçekleştirilir. Yöntemin etkililiğinin test edildiği verinin bu parçasına test verisi denir. Bu veri, tüm verinin belirli bir oranından edildiğinden dolayı test verisi oranı olarak ifade edilir. Veri madenciliğinde yöntemlerin etkililiğinin bu şekilde test edilmesinin nedeni model hata oranlarının yanlı kestirimlerinin önüne geçmektir. Benzer amaçlar için kullanılan diğer yöntemler, önyükleme (Efron, 1983) ve çapraz geçerleme (Lachenbruch ve Mickey, 1968) teknikleridir (Michie, Spiegelhalter ve Taylor, 1994). Ancak, tüm veriden belirli oranda (genellikle 1/3 oranında - %33) test verisi seçilerek bu veri ile yordama işleminin gerçekleştirilmesi sıklıkla tercih edilen ve büyük örneklemler için de çoğunlukla kullanılan etkili bir yöntemdir.

Veri madenciliği yöntemlerinin eğitim verileri üzerinde uygulanabilirliği ve etkililiği üzerine son on yıllık süreçte birçok araştırmaya rastlamak mümkündür (Barker, Trafalis ve Rhoads, 2004; Dekker, Pechenizkiy ve Vleeshouwers, 2009; Kotsiantis, Pierrakeas ve Pintelas, 2003; Hamalainen ve Vinni, 2006; Hamalainen ve Vinni, 2011; Minaei-Bidgoli, Kashy, Kortemeyer ve Punch, 2003; Nghe, Janecek ve Haddawy, 2007; Osmanbegović ve Suljić, 2012; Romero, Espejo, Zafra, Romero ve Ventura, 2013; Romero, Ventura, Espejo ve Hervas, 2008; Shahiri, Husain ve Rashid, 2015). Bu araştırmalar, genel olarak öğrenci performansının yordanması, değerlendirilmesi ve performansı etkileyen faktörlerin belirlenmesi amacı taşımaktadır. Romero ve Venturo (2007) 1995 ve 2005 yılları arasında eğitim alanında yapılan veri madenciliği çalışmalarını derleyerek çeşitli özelliklerine göre sınıflandırmışlardır. Ancak, bu araştırmalardan çok az bir kısmı veri madenciliği algoritmalarının karşılaştırılmasının yanında örneklem büyüklüğü ve eğitim setinin büyüklüğü bu algoritmaların performansına etkisine değinmiştir. Hâlbuki istatistik, mühendislik, sağlık ve sosyal bilimler gibi birçok alanda farklı veri yapısının veri madenciliği algoritmaları üzerindeki etkileri önemli bir araştırma konusu haline gelmiştir. Ayrıca, Türkiye'de EVM ile ilgili uygulamalara ve yukarıda anlatılan yöntemlerden Naive Bayes ve k-en yakın komşuluk tekniklerine yönelik çalışmalar sınırlı düzeydedir. Bu çalışmada, alan yazında görülen bu eksikliklere yönelik PISA (2012) uygulamasında alınan bir veri kullanılarak kapsamlı bir uygulama yapılması hedeflenmiştir.

Bu çalışmanın amacı, öğrencilerin, çeşitli özellikleri bakımından PISA (2012) matematik başarılarını yordamada Naive Bayes, k-en yakın komşuluk, lojistik regresyon ve yapay sinir ağları çok katmanlı algılayıcı yöntemlerinin performanslarının farklı örneklem büyüklükleri (küçük, orta, büyük) ve test verisi oranlarına (%11, %22, %33, %44 ve %55) göre nasıl değiştiğini gözlemlemektir.

Yöntem

Çalışmanın yöntem kısmı burada özetlenmelidir. Naive Bayes, *k*-en yakın komşuluk, yapay sinir ağları ve lojistik regresyon algoritmalarının farklı koşullar altında performanslarının belirlenmesi ve açıklanması hedeflendiğinden, bu çalışma temel bir araştırmadır. Bu tür araştırmalarda metodolojik analize dayalı çalışmalar yaparak bilgi üretilmesi amaçlanmaktadır (Büyüköztürk, Çakmak-Kılıç, Akgün, Karadeniz ve Demirel, 2015). Temel araştırmalar mevcut bilgiye yeni bilgiler eklemeyi amaçlamaktadır (Karasar, 2005). Araştırma aynı zamanda yöntemler arasındaki ilişkileri incelemek açısından nicel ilişkisel araştırmadır. Bu tür çalışmalar nicel değişkenler arasındaki ilişkileri araştırmayı, açıklamayı ve keşfetmeyi amaçlamaktadır (Fraenkel ve Wallen, 2006).

Araştırmanın evreni, PISA uygulamasına katılan OECD ülkelerindeki 15 yaş grubundaki öğrencilerdir. Her bir ülke için evreni temsil eden örneklemler PISA uygulayıcıları tarafından tabakalı tesadüfi örnekleme yoluyla seçilmiştir. OECD ülkelerinden PISA uygulamasına katılan toplam kişi sayısı 295416 kişidir. Bu çalışmada, kayıp veriler, artık ve uç değerler incelenip çıkartıldıktan sonra 62728 kişilik hedef evrene ulaşılmıştır. Araştırmada, incelenen yöntemlerin performanslarına yönelik yüksek kesinlikte sonuçlar elde etmek amacıyla 500 kişilik örneklem (küçük) için 100 veri dosyası, 1000 kişilik örneklem (orta) için 50 veri dosyası, 5000 kişilik örneklem (büyük) için 30 veri dosyası olmak üzere toplam 180 veri dosyası oluşturulmuştur.

Araştırmanın veri toplama araçları, PISA (2012) uygulamasında öğrencilerin matematik alanındaki akademik performanslarını ölçmek amacıyla geliştirilen matematik bilişsel testi ve öğrenciyi var olan tüm özellikleri ile değerlendirmeyi amacıyla hazırlanan öğrenci anketidir. Öğrenci anketinde ise öğrencilerin evleri, okulları, kendileri ve öğrenme deneyimleri hakkında çeşitli bilgileri içeren formları doldurmaları beklenmiştir (OECD, 2014a). Bu çalışmada, PISA uygulayıcıları tarafından takip edilen süreçler (OECD, 2014a) sonucunda elde edilen veri OECD'nin herkese açık veri tabanından alınarak kullanılmıştır.

Verilerin analizinde, öncellikle, PISA (2012) öğrenci anketinden elde edilen veriden öğrencilerin demografik bilgileri ve matematiğe ilişkin tüm değişkenleri alınmıştır. Daha sonra OECD (2014b) tarafından yayınlanan PISA 2012 teknik raporu göz önünde bulundurularak diğer değişkenlerin bileşiminden oluşan değişkenler alınmış ve kalan değişkenler dosyadan çıkartılmıştır. Daha sonra ise kayıp verilerden oluşan değişkenler, kalan değişkenlere ait tüm kayıp veriler ve uç değerler silindiğinde matematik performansı ile birlikte 35 değişken ve 62728 kişiden oluşan hedef evren elde edilmiştir.

Veri madenciliği yöntemleri çok fazla değişkenle çalışmakla birlikte, sınıflandırmaya katkısı olmayan değişkenlerin analizde bulundurulması yapılacak analizlerin çok zaman almasına ve sınıflandırma performansının düşmesine neden olmaktadır. Bu amaçla, bir veri ön işleme süreci olan değişken seçme veri madenciliğinde sıkça kullanılan önemli tekniklerden biridir (Blum ve Langley, 1997; Liu ve Motoda, 2001). Bu çalışmada, WEKA Version 3.9.0 yazılımında (Hall ve diğerleri, 2009) yer alan Information Gain Ranking Filter, Chi-squared Ranking Filter, Gain Ratio Feature Evaluator ve Symmetrical Uncertainty Ranking Filter metotları kullanılmıştır. Araştırmanın bağımlı değişkeni ve birinci makul değer olan PV1MATH (Plausible Value 1) değişkeni, PISA uygulayıcıları tarafından belirlenen ve öğrencilerin matematikte yeterliğini temsil eden altı düzeyden (OECD, 2014a) orta düzeyin (482) altında ve üstünde yer alan öğrenciler şeklinde iki kategorili bir değişkene dönüştürülmüştür. Daha sonra öğrencilerin matematik performanslarına göre sınıflandırmaya en çok katkı sağlayan ilk 10 değişken çalışmaya dâhil edilmiştir. Bu değişkenler, matematik öz-yeterliği, matematiksel benlik algısı, matematik kaygısı, ekonomik, sosyal ve kültürel statü indeksi, problem çözmeye açık olma, ülke, babanın eğitim düzeyi (ISCED), anne eğitim düzeyi (ISCED), öğretmen davranışı: öğrenciyi yönlendirme ve hesap makinesi kullanımıdır. Bu çalışmada, tüm analizler bu değişkenler kullanılarak gerçekleştirilmiştir.

Araştırmada kullanılacak değişkenlere karar verildikten sonra analizlerin varsayımları kontrol edilmiştir. Bu çalışmada yapılacak analizlere yönelik olarak yapılan varsayım kontrollerinde örneklem büyüklüğünün yeterli olduğu, kayıp ve uç değer olmadığı, normalliğin sağlandığı, varyans-kovaryansların homojen olmadığı, doğrusallığın kısmen sağlandığı, çoklu bağlantı ve tekilliğin olmadığı görülmüştür. Ayrıca, Naive Bayes yöntemi için koşullu bağımsızlık varsayımı sağlanamamıştır. Analizlerde, IBM SPSS Statistics 23, Microsoft Office Excel 2016 ve WEKA Version 3.9.0 yazılımlarından yararlanılmıştır. Model değerlendirilmesinde doğruluk oranı, RMSE değerleri ve modellerin işlem hızları kullanılmıştır. Yöntem karşılaştırma ölçütü olarak kullanılan RMSE değerleri, parametrik yöntemlerin varsayımlarını karşılamadığından, bu değerlerin karşılaştırılmasında Friedman testi kullanılmıştır. Yöntemlerin ikili karşılaştırmaları ise Wilcoxon İşaretli Sıralar testi ile yapılmıştır.

Sonuç ve Tartışma

Bu araştırmada, farklı örneklem büyüklüklerinin ve test verisi oranlarının yöntemlerin performansları üzerinde yarattığı etkiler şu şekildedir:

1. Örneklem büyüdükçe, tüm yöntemlerin doğru sınıflandırma performansları artmış geçerliği ve güvenirliği yüksek sonuçlar elde edilmiştir.

2. Örneklem büyüdükçe, Naive Bayes yönteminin analiz süresi değişmemekle birlikte diğer yöntemlerin analiz işlem süreleri uzamıştır.

3. Test verisi oranı örneklem büyüklüğüne göre yöntemlerin sınıflandırma performanslarında farklı etkiler yaratmıştır.

4. Örneklem büyüdükçe test verisi oranının arttırılmasının yöntemlerin performansları üzerindeki etkisi azalmıştır.

5. Test verisi oranı tüm verinin üçte birinden az olduğunda da yüksek doğru sınıflandırma performansları elde edilmiştir.

6. Örneklem büyüdükçe test verisi oranı tüm verinin üçte birinden fazla olduğunda bile güvenilir sınıflandırma performansları elde edilebilmiştir.

7. Tüm örneklem büyüklükleri için test verisi oranının değişimden en az etkilenen yöntem Naive Bayes yöntemidir.

8. Örneklem büyüklüğünün artmasından en fazla etkilenen yöntem lojistik regresyon analizidir.

9. Tüm koşullarda en düşük doğruluk oranları en yakın komşuluk yöntemi ile elde edilmiştir.

Küçük örneklemlerde, NB yöntemi ile diğer yöntemlere göre, yüksek doğruluk oranları, daha az hatalı kestirimler yapılmış ve analizler çok kısa sürede tamamlanmıştır. Yapılan bazı araştırmalarda da küçük örneklemlerde NB yönteminin diğer yöntemlere göre daha iyi sonuçlar verdiği görülmüştür (Göker, 2012; Hamalainen ve Vinni, 2006; Hamalainen ve Vinni, 2011, Kotsiantis ve diğerleri, 2003; Osmanbegović ve Suljić, 2012). Araştırmada LR yöntemi, tüm koşullarda YSA yöntemine göre daha yüksek performans göstermiştir. Bu bulgu yapılan bazı araştırma sonuçlarından farklı olmakla birlikte (Bahadır, 2013; Çırak, 2012; Tepehan, 2011) bu durumun oluşmasının en önemli nedeni veri yapısının LR analizi için uygun olmasıdır. LR yöntemi orta ve büyük örneklemlerde, daha az hatalı ve daha yüksek doğrulukta kestirimler yapmaktadır. Dekker ve diğerleri (2009) tarafından yapılan çalışmada, benzer büyüklükte örneklemde LR yöntemi Bayes yöntemine göre daha iyi performans göstermiştir.

NB ve LR yöntemlerinden sonra en yüksek doğruluk oranları sırasıyla MLP ve KNN yöntemleri ile elde edilmiştir. Romero ve diğerlerinin (2013) yaptıkları çalışmada, numerik ve kategorik veri için KNN yönteminin diğer sınıflandırıcılara göre daha düşük performans göstermiştir. Benzer şekilde, bu çalışmada, orta ve büyük örneklemlerde, MLP yöntemi KNN yönteminden daha az hatalı sonuçlar vermiştir. Ancak, küçük örneklemlerde tersi bir durum söz konusudur. Bu durum, KNN yönteminin MLP yöntemine göre daha basit bir istatistiksel yapıya sahip olması ve seçilen k değerinin en yakın komşuluğu belirlemede küçük örneklemlerde daha kararlı davranmasından kaynaklanmıştır. MLP yönteminde ise katman sayısının 3 seçilmesi ağın eğitilmesinde etkili olmasına rağmen küçük örneklemlerde hata miktarının fazla olmasına neden olmuştur.

Bu çalışmada, KNN yöntemi diğer yöntemlere göre tüm koşullarda daha düşük doğru sınıflandırma performansı göstermiştir. Ancak, yapılan bazı araştırmalarda KNN yönteminin de en az YSA ve LR yöntemleri kadar performans gösterdiği görülmüştür (Minaei-Bidgoli, Kashy, Kortemeyer ve Punch, 2003; Yurdakul ve Topal, 2015). Benzer şekilde, Shahiri ve diğerleri (2015), 2002 ile 2015 yılları arasında uluslararası veri tabanlarında yayınlanan çalışmaları karşılaştırmış ve ortalama performans açısından NB yönteminin KNN ve YSA yöntemlerine göre daha düşük performans gösterdiği görülmüştür. Ancak, bu çalışmada, NB yöntemi özellikle küçük ve orta büyüklükteki örneklemlerde daha yüksek sınıflandırma performansı göstermiştir. Bazı araştırmacılar, farklı koşullar ve veriler için bir sınıflandırma yönteminin en iyi olduğunu söylemek doğru olmadığını ifade etmişlerdir (Romero ve diğerleri, 2013; Shahiri ve diğerleri, 2015). Barker ve diğerleri (2004) ise farklı yıllarda mezun olan öğrencilerin mezun olma durumlarına göre yaptıkları sınıflandırmada farklı yıllarda verinin yapısına göre farklı yöntemlerin etkili olabileceğini göstermişlerdir. Bu nedenle, bu araştırmadan elde edilen bulgular, kullanılan verinin yapısı ve analiz koşulları çerçevesinde yorumlanmıştır.

Bu araştırmada, Naive Bayes yöntemi için koşullu bağımsızlık varsayımının sağlanmadığı bir veri ile analizler gerçekleştirilmiştir. Bu sonuç, Hamalainen ve Vinni (2011) tarafından belirtildiği gibi, Naive Bayes'in koşullu bağımsızlık varsayımı karşılanmamış olsa bile iyi performans gösterebileceğini göstermiştir. Yapılacak araştırmalarda bu varsayımın sağlandığı, k-en yakın komşuluk yöntemi için seçilecek k değerinin farklı şekillerde belirlendiği, yapay sinir ağları yönteminde, katman sayısinin seçildiği, lojistik regresyon analizi ve yapay sinir ağları yöntemleri için varyans-kovaryans matrislerinin homojenliğinin sağlandığı koşullarda yöntemlerin sınıflandırma performansları değerlendirilip karşılaştırılabilir.

Benzer koşullarda yapılacak uygulamalarda Naive Bayes yönteminin kullanılması, zaman kaybı yaşanmadan geçerliği ve güvenirliği yüksek sonuçların elde edilmesini sağlayacaktır. benzer koşullar için yapılacak uygulamalarda daha yüksek sınıflandırma performansı sağlayabilmek için diğer yöntemler, k-en yakın komşuluk yöntemine tercih edilebilir. Örneklem büyüklüğü fazla olduğunda Naive Bayes ve lojistik regresyon yöntemlerinin YSA'ya tercih edilmesi daha yüksek performans ve zaman tasarrufu sağlayacaktır.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(4); 346-361

Examination of Meeting the Needs of University Students from Social Support Systems

Betül DÜŞÜNCELİ * T. Seda ÇOLAK ** Süleyman DEMİR *** Mustafa KOÇ ****

Abstract

In this study, it is aimed to determine which social support systems respectively preferred by students to meet their basic needs. The research was conducted with 347 university students from Sakarya University Faculty of Education, 243 of whom were female and 104 of whom were male. A ranking chart was used to determine the rank of fulfilment of the five basic needs of the students, as in the Maslow's hierarchy of needs, (physiological, safety, love and belonging, esteem, and self-actualization), by social support systems (family, relatives, friends, teacher-school, and society). The data was analyzed by rank order judgment scaling. As a result of the research, it was found that university students regard family as the primary social support system in meeting all their needs (physiological, safety, love and belonging, esteem, and self-actualization). The ranking does not change in meeting the needs of safety, love and belonging, and esteem; in meeting the physiological needs, it was observed that relatives are preferred more than friends. Another finding of the research is that in meeting the need for self-actualization, relatives are preferred the least.

Key Words: Maslow, hierarchy of needs, social support, rank order judgment scaling.

INTRODUCTION

Research into the underlying factors of human behavior has been the subject of psychology science for many years. Maslow’s approach to psychological needs is one of the most popular theories in this field (Roediger, Capaldi, Paris, Polivy, & Herman, 1996). Maslow first introduced a hierarchical structure of needs in 1943 and introduced his theory of needs in his book “Motivation and Personality” in 1954 (Maslow, 1970). He argued that there are differences between human motives and animal motives and expressed human motives in the form of a pyramid. At the base of this pyramid are biological motives, and at the top are psychological motives (Cüceloğlu, 2003). It is possible to say that Maslow’s theory is still valid today. The hierarchy of needs can be examined under five headings: physiological needs, need for safety, need for love and belonging, need for self-esteem and need for self-actualization (Plotnik, trans. 2009).

Physiological Needs

Maslow claims that human comes to earth from the lowest level (McConnell & Philipchalk, 1992). Needs such as food, water, sexuality, breathing, and sleeping are discussed in the category of physiological needs (Burger, 2006; Plotnik, trans. 2009; Roediger et al. 1996). One has to satisfy their physical needs before meeting psychological or social needs (McConnell & Philipchalk, 1992). Once this need is adequately met, it will be possible for the individual to be motivated to meet other needs.

* Assist. Prof. Dr., Sakarya University, Education Faculty, Sakarya-Turkey, bbayraktar@sakarya.edu.tr, ORCID ID: 0000-0002-6794-8811
** Assist. Prof. Dr., Düzce University, Education Faculty, Düzce-Turkey, tugbacolak@duzce.edu.tr , ORCID ID: 0000-0002-7219-1999
*** Assist. Prof. Dr., Sakarya University, Education Faculty, Sakarya-Turkey, suleymand@sakarya.edu.tr , ORCID ID: 0000-0003-3136-0423
**** Prof. Dr., Düzce University, Education Faculty, Düzce-Turkey, mustafa.koc@duzce.edu.tr, ORCID ID: 0000-0002-8644-4109

To cite this article:
Düşünceli, B., Çolak, T. S., Demir, S., & Koç, M. (2020). Examination of meeting the needs of university students from social support systems. *Journal of Measurement and Evaluation in Education and Psychology, 11*(4), 346-361. doi: 10.21031/epod.706475.

Received: 19.03.2020
Accepted: 26.10.2020

Need for Safety

Babies become ready to explore the physical environment once their basic needs are met. However, for that, they need to feel safe first (McConnell & Philipchalk, 1992). The need for safety can be met through protection from crime, fire, extreme heat or cold, wild animals, or dangers such as economic disaster (Plotnik, trans. 2009; Roediger et al. 1996). For example, a student who hears about negative things happen in a school on the news and social media may not feel safe with the possibility that these may also happen at his/her own school (Shaughnessy, Moffitt, & Cordova, 2018). This prevents the student from being driven to the upper categories in the hierarchy, for example, to be accepted by their peers and to be successful at school.

Need for Love and Belonging

The need for love and belonging can be met by connecting with others, by being accepted by others (Plotnik, trans. 2009), and by acquiring a place in a group (Roediger et al. 1996). According to Maslow, this need can only be met by other people (McConnell & Philipchalk, 1992). Once the needs at the lower level are satisfied, the human looks for a resource to connect with, so it is considered as the need that drives people to be social.

Need for Esteem

Success can be met by gaining competence, approval, and attestation (Plotnik, trans. 2009). One reason people connect with others is that they help themselves to set their life goals. Through feedback from people, individuals are able to gain insight into how far they have achieved their life goals (McConnell & Philipchalk, 1992). Having reached this level, individuals can move on to the final stage: self-actualization.

Need for Self-Actualization

It can be defined as experiencing one’s own potential (Plotnik, trans. 2009). Maslow emphasized the potential of the individual by saying, “what a man is and what he could be” (p. 272) while explaining the need for self-actualization (Maslow, 1970). This can include being a parent, being an athlete, a musician, or whatever appropriate (Roediger et al. 1996). While in the way of performing themselves, individuals’ characteristics to be creative, to love, and to be healthy and strong, in line with their goals, come to the fore (Gençtanırım, 2013). Maslow states that unless one is self-confident, they will not dare to express themselves in their own way, cannot contribute to society, and thus cannot unleash this innate potential (McConnell & Philipchalk, 1992).

Maslow assumes that all people go through these five levels in some way, and all people deal with problems with the lower level before moving to an upper level (McConnell & Philipchalk, 1992). The hierarchy of needs has been increased to seven levels by adding two more needs as “*need for knowing/understanding*” and “*need for being aesthetic*” (İnceoğlu, 2004). When lower-level motives reach satisfaction, the individual becomes ready for higher-level motives (Cüceloğlu, 2003; Maslow, 1970; Seker, 2014). If the basic needs are not met at a higher level, one can go backwards within the hierarchy (Plotnik, trans. 2009). Maslow defines physiological, safety, love-belonging, and esteem needs as the deficiency needs and states that it is compulsory that these needs are met. However, according to Maslow, once these needs are met, they will decrease. Needs such as knowing, understanding, and appreciating beauty, which he describes as growth needs, can never be fully met (Slavin, trans. 2013). In a sense, as the growth needs are satisfied, it can be said that there are dynamic structures that can replace them with new developmental needs.

When we look at the hierarchy of needs today, it can be said that people need others to meet their needs from the first step to the last step. Examples include that a baby needs its mother to feed it, a person feels more confident with his/her family, loved by her friends at school, valued by her boss at work as an adult, or needs an environment where s/he can reveal his/her potential. Actually, that an individual needs the presence of others even when meeting a very basic need brings to mind the role of environmental factors in meeting these needs. At this point, the concept of social support comes into play. According to Yıldırım (1997), factors such as family, environment of family, friends, relations with the opposite sex, teachers, colleagues, neighbors, ideological, religious or ethnic groups, and the society in which the individual lives can be said to constitute the sources of social support for the individual. Social support allows the person to cope with the difficulties in life and acts as a protective buffer (Arslantaş & Ergin, 2011; Lin, Thompson & Kaslow, 2009; Terzi, 2008).

It is possible to come across many studies reporting that social support plays an important role in the school adaptation processes of university students (Mallinckrodt, 1988; Rahat & İlhan, 2016; Tinajero, Martínez-López, Rodríguez, Guisande, & Páramo, 2015). Changes in the individual's self or source of support can cause the individual's level of social support to change (Yıldırım, 1997), but it is a fact that social support systems such as family, friends, and teachers have an important place in the lives of individuals. Khallad and Jabr's (2016) study on the mental health of university students found that for Jordanian university students, social support of family is essential, and for Turkish university students, social support of friends is at an important point. Similarly, another study in Turkey shows that somatization, anger/aggression, depression, and anxiety symptoms decrease as family support increases in university students (Doğan, 2016). Haskan-Avcı and Yıldırım (2014) found that adolescents with a high propensity for violence have low levels of support from family, friends, and teachers. Indeed, the literature shows the importance of social support not only in the early years of life but also in the later years. Studies conducted with individuals over 60 years showed that individuals with higher levels of social support are less depressed (Aksüllü & Doğan, 2004; Bozo, Toksabay, & Kürüm, 2009).

The universities that include the sample group of this study have the duty to be one of the important institutions that enable students to step into adulthood and professional life. Especially in Turkey, universities have an important role in presenting the experiences that lead many students to start organizing their lives independently from their parents. Sarı, Yenigün, Altıncı, and Öztürk (2011) state that the basic psychological needs of university students must be satisfied in order to increase their self-sufficiency perceptions and decrease their trait anxiety. This study aims to emphasize the importance of social support systems that exist in the lives of university students in the process of meeting these needs. Identifying one's social support systems has an important place in preventive mental health services (Terzi, 2008). In this context, the aim of this study is to determine which social support systems (family, relatives, friends, teacher-school, and society) respectively preferred by students to meet their basic needs (physiological, safety, love and belonging, esteem, and self-actualization). It is thought that determining which social support systems are functional in meeting which needs to raise healthier individuals, and setting out what is needed for non-functional social support systems to become functional will contribute to raising individuals who are more mentally healthy. It is thought that this study will shed light on this issue.

METHOD

In this study, survey model was used to determine the level of meeting the needs of university students from social support systems by rank order judgment scaling. The purpose of survey model is to reveal the current situation. In order to achieve this, the attitudes, interests, and abilities of a group are measured by quantitative data collection methods (Creswell 2009; Karasar, 1998).

Participants

The research was conducted with 347 university students studying at Sakarya University, Faculty of Education. The gender of the participants is 70% female and 30% male. Considering that the university period is the transition period to adulthood when students are separated from their families and started to take responsibility for their own lives, data were collected from university students.

Data Collection Instruments

In order to obtain personal information about the participants, their gender was asked. Also, a ranking chart was prepared by the researchers to determine the rank of fulfilment of the five basic needs of the students, as in the Maslow's hierarchy of needs (physiological, safety, love and belonging, esteem, and self-actualization), by social support systems (family, relatives, friends, teacher-school, and society). Each participant was asked to rank separately for five needs and each need from 1 to 5 in the order of priority according to whichever social support system they met those needs with. During the ranking procedure, 1 was used for *the highest level of social support system meeting the need*, and 5 was used for *the lowest level of social support system*. For example; a participant who ranks as family (1), relatives (2), friends (3), teacher-school (4), and society (5) in meeting her/his physiological needs, s/he was able to make another ranking as family (1), friend (2), relatives (3), society (4), and teacher / school (5) in meeting their security needs. In order for participants' perceptions of the concepts in the study to be similar, the ranking chart provides information on the five basic needs in Maslow's hierarchy of needs and social support concepts.

Data Analysis

Data collection by sorting objects, individuals, situations or methods (stimulus) by scorers according to a specific rule is a method often used in the social sciences. However, in analyzing such data, generally the stimulus written mostly in the first rank is taken into account. In case of this study, rank order judgment method was used. The rank order judgment scaling is a method that can be used to analyze data by considering all the rankings made, not just the stimulus in the first place (Baykul & Turgut, 1992; Guilford 1954). The rank order judgment scaling begins with the creation of the Frequency Matrix of Rank Ordering regarding the order in which each stimulus is preferred. For each stimulus, the probability of being preferred in binary comparison with other stimuli is calculated using the $P_{j>k} = \sum_{i=1}^{n} \left[f_{ji}(f_{k<i} + f_{ki} / 2 \right]$ formula[1] , and the Probability Matrix of Rank Ordering showing the probability of preference of the stimuli in binary comparisons is created. From this stage to the end, calculation stages are as follows:

1. For each unit in the probability matrix, the Z values for the corresponding unit normal distribution are calculated, and the Unit Normal Deviate Matrix is created.
2. Mean Z values are obtained by taking the average of each column.
3. The smallest mean Z value is shifted so that it equals to zero.
4. The obtained mean Z values constitute the scale values of each stimulus (Anıl & Güler, 2006; Baykul & Turgut, 1992; Guilford, 1954).

In this study, the rank order judgment scaling was used to analyze the data. The analysis of the data was made via Excel with the using formulas in the literature (Anıl & Güler, 2006; Anıl & İnal, 2019; Baykul & Turgut, 1992).

RESULTS

In the analysis of the data, the levels of meeting the needs of the students by their social support systems were calculated using the scaling method based on rank frequency tables and rank orderings. Calculation of scale values obtained by rank order judgment scaling, were reported only for

[1] j and k: stimulus; i: rank value

$P_{j>k}$: The probability that stimulus (j) is preferred over stimulus (k)

f_{ji} : Frequency of which the value (i) is given to stimulus (j)

f_{ki}: Frequency of which the value (i) is given to stimulus (k) physiological needs, and graphs obtained according to scale values for all needs were provided. In Table 1, the frequency matrix meeting the physiological needs of university students from social support systems was given.

Table 1. The Frequency of Meeting the Physiological Needs of University Students from Social Support Systems

	Family	Relative	Friend	Teacher-School	Society
1st rank	325	2	6	2	12
2nd rank	10	150	163	12	11
3rd rank	1	116	135	55	41
4th rank	2	40	36	182	87
5th rank	9	39	7	96	196
Total	347	347	347	347	347

According to Table 1, it can be said that in meeting their physiological needs, university students prefer family mostly in the first rank among the social support systems, relative and friend in the second and third rank, teacher/school in the fourth rank, and society in the last rank.

In order to analyze data by rank order judgment scaling, a probability matrix of rank ordering has been created primarily using Table 1. The obtained probability matrix of rank ordering is given in Table 2.

Table 2. Probability Matrix of Rank Ordering in Meeting the Physiological Needs of University Students by Social Support Systems

	Family	Relative	Friend	Teacher-School	Society
Family		0.96	0.95	0.97	0.96
Relative	0.04		0.44	0.80	0.82
Friend	0.05	0.56		0.88	0.88
Teacher-School	0.03	0.20	0.12		0.62
Society	0.04	0.18	0.12	0.38	

The probability values in Table 2 represent the probability that the stimulus in the row is preferred instead of the stimulus in the column. For example, the probability of choosing the family instead of the relative from social support systems is 0.96, while the probability of choosing the relative instead of the family is 0.04. The Z values for the normal distribution were calculated using the probability values in Table 2, and the unit normal deviate matrix in Table 3 was formed.

Table 3. Unit Normal Deviate Matrix for Social Support Systems in Meeting the Physiological Needs of University Students

	Family	Relative	Friend	Teacher-School	Society
Family		1.75	1.66	1.90	1.77
Relative	-1.75		-0.14	0.84	0.92
Friend	-1.66	0.14		1.16	1.18
Teacher-School	-1.90	-0.84	-1.16		0.30
Society	-1.77	-0.92	-1.18	-0.30	
Mean Z	-1.42	0.03	-0.16	0.72	0.83
Scale Values	0.00	1.44	1.25	2.14	2.25

In Table 3, a row with the mean Z values is created by averaging the Z values in each row, and scale values are obtained by shifting the smallest mean Z value to zero. With these scale values, the graph in Figure 1 is obtained.

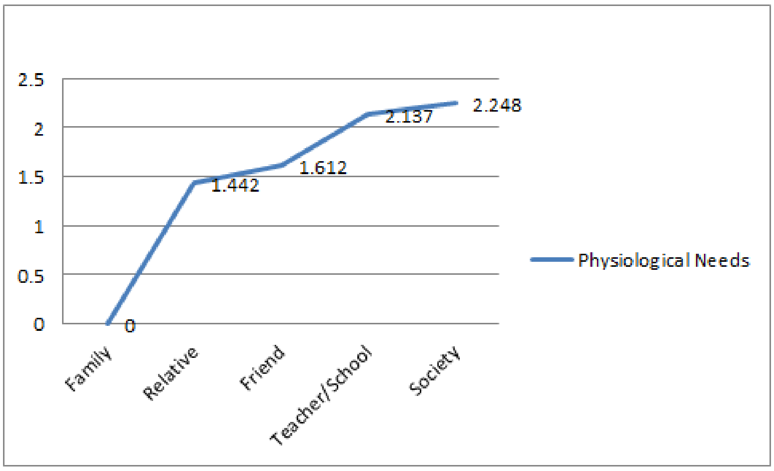


Figure 1. The Level of Meeting the Physiological Needs of University Students from Social Support Systems

It is seen in Figure 1 that university students prefer family first among the social support systems to meet their physiological needs and then relative, friend, teacher/school, and society. It can be said that the preference of friend and relative among the social support systems in meeting physiological needs is close to each other. Similarly, it can be said that the choices of teacher/school and society from social support systems in meeting physiological needs are close to each other.

Table 4. The Frequency of Meeting the Need for Safety of University Students from Social Support Systems

	Family	Relative	Friend	Teacher-School	Society
1st rank	297	5	18	4	24
2nd rank	28	137	123	30	28
3rd rank	8	97	136	75	32
4th rank	7	62	54	165	59
5th rank	7	46	16	73	204
Total	347	347	347	347	347

According to Table 4, it can be said that in meeting their need for safety, university students prefer family mostly in the first rank among the social support systems, relative and friend in the second and third rank, teacher/school in the fourth rank, and society in the last rank.

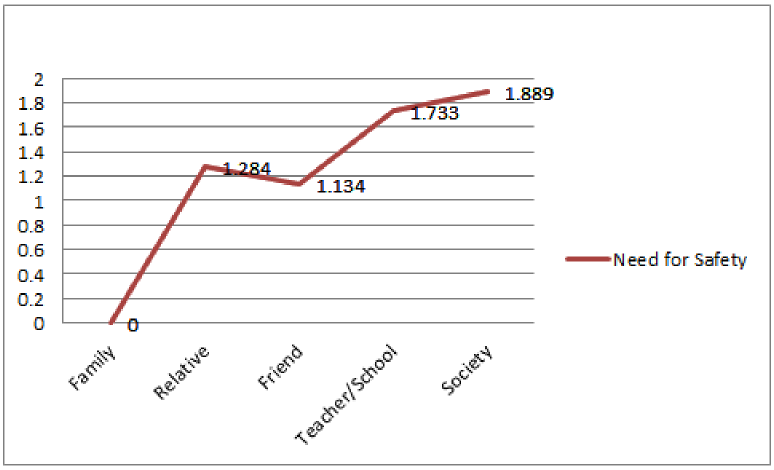


Figure 2. The Level of Meeting the Need for Safety of University Students from Social Support Systems

It is seen in Figure 2 that university students prefer family first among the social support systems to meet their need for safety, then prefer friend, relative, teacher/school, and society. It can be said that the preference of friend and relative among the social support systems in meeting safety needs is close to each other. Similarly, it can be said that the choices of teacher/school and society among the social support systems in meeting need for safety are close to each other.

Table 5. The Frequency of Meeting the Need for Love and Belonging of University Students' from Social Support Systems

	Family	Relative	Friend	Teacher-School	Society
1st rank	311	1	22	3	10
2nd rank	20	108	204	7	8
3rd rank	4	146	102	64	32
4th rank	3	50	15	207	71
5th rank	9	42	4	66	226
Total	347	347	347	347	347

According to Table 5, it can be said that in meeting their need for love and belonging, university students mostly prefer family in the first rank among the social support systems, friend in the second rank, relative in the third rank, teacher/school in the fourth rank, and society in the last rank.

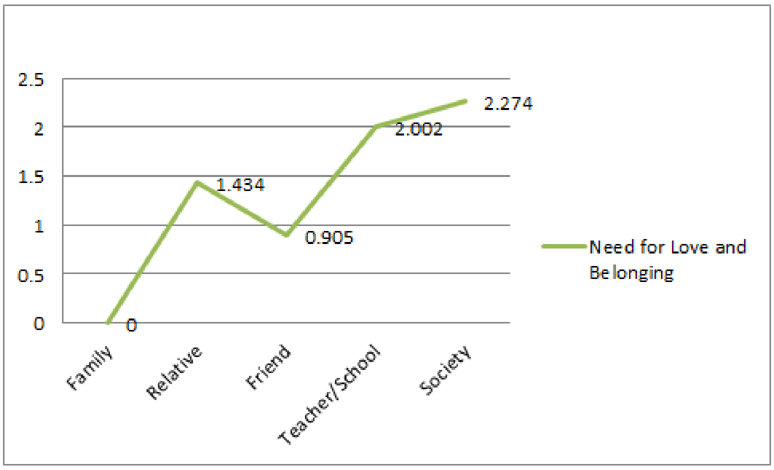


Figure 3. The Level of Meeting the Need for Love and Belonging of University Students from Social Support Systems

It is seen in Figure 3 that university students prefer family first among the social support systems to meet their need for love and belonging, then prefer friend, relative, teacher/school, and society. Similarly, it can be said that the choices of teacher/school and society among the social support systems in meeting their need for love and belonging are close to each other.

Table 6. The Frequency of Meeting the Need for Esteem of University Students from Social Support Systems

	Family	Relative	Friend	Teacher-School	Society
1st rank	229	5	52	17	43
2nd rank	60	96	122	47	21
3rd rank	27	90	115	77	38
4th rank	16	75	42	140	75
5th rank	15	81	16	66	170
Total	347	347	347	347	347

According to Table 6, it can be said that in meeting their need for esteem, university students prefer family mostly in the first rank among the social support systems, relative and friend in the second and third rank, teacher/school in the fourth rank, and society in the last rank.

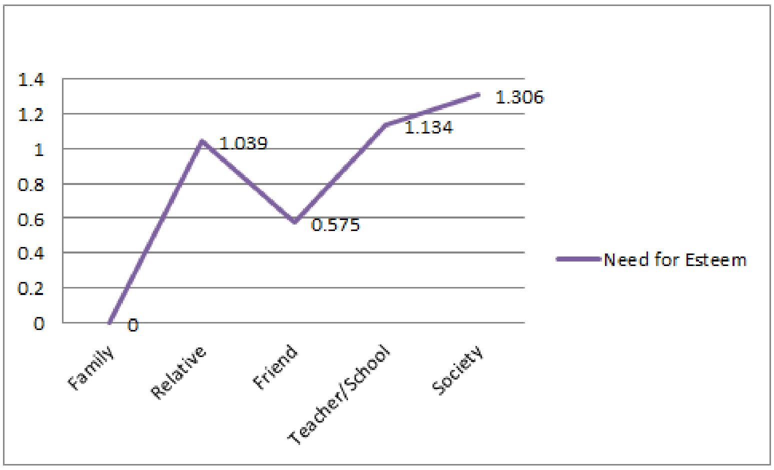


Figure 4. The Level of Meeting the Need for Esteem of University Students from Social Support Systems

It is seen in Figure 4 that university students prefer family first among social support systems to meet their need for esteem, then prefer friend, relative, teacher/school, and society. It can be said that the choices of relative, teacher/school, and society among the social support systems in meeting their need for esteem are close to one another.

Table 7. The Frequency of Meeting the Need for Self-Actualization of University Students from Social Support Systems

	Family	Relative	Friend	Teacher-School	Society
1st rank	197	5	53	49	45
2nd rank	70	54	134	56	32
3rd rank	40	78	100	93	35
4th rank	21	87	49	111	80
5th rank	19	123	11	38	155
Total	347	347	347	347	347

According to Table 7, it can be said that in meeting their need for self-actualization, university students prefer family mostly in the first rank among the social support systems, friend in the second rank, teacher/school in the third and fourth rank, and relative and society in the last rank.

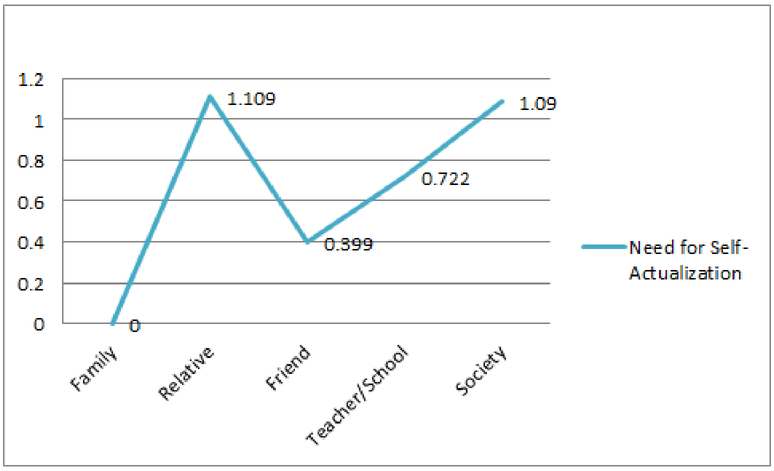


Figure 5. The Level of Meeting the Need for Self-actualization of University Students from Social Support Systems

It is seen in Figure 5 that university students prefer family first among the social support systems to meet their need for self-actualization, then prefer relative, friend, teacher/school, and society. In order to meet the need for respect, it can be said that the preferences of relative and society among the social support systems are close to each other.

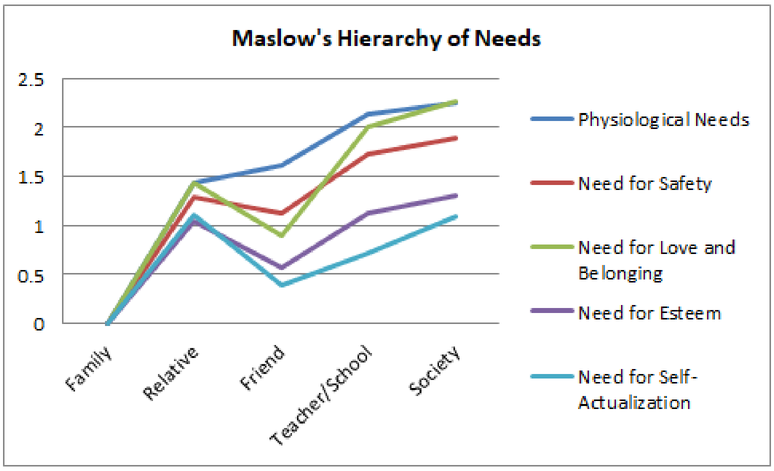


Figure 6. Level of Social Support Systems to Meet the Needs of University Students

As a result of the rank order judgment scaling analysis, the social support systems that the university students prefer to meet their needs are seen holistically in Figure 6. It can be said that university students regard family as the primary social support system in meeting all their needs (physiological, safety, social, esteem, and self-actualization). It is observed that the rank orders in meeting the needs of safety, love-belonging, and esteem are the same (first: family, second: friend, third: relative, fourth: teacher/school, fifth: society). In meeting physiological needs, it is seen that the relative is more preferred than the friend. In meeting the need for self-actualization, it is seen that the teacher and the society are preferred over the relative in the last rank.

DISCUSSION and CONCLUSION

As a result of research, it is found that in meeting their needs, all university students see family as the primary social support system; and the order of the ranking does not change in meeting the needs for safety, love-belonging, and esteem [family (1), friend (2), relative (3), teacher/school (4), society (5)]; and in the fulfillment of physiological needs, relative is preferable than friend. Yılmaz, Yılmaz, and Karaca's (2008) study also reveals that university students see family as their social support rather than friends and people special for them. As for the importance of the family, Engin, Özen, and Bayoğlu, (2009) state that 91.3% of the basic needs required in education and training activities are always met by students' families. Dwyer and Cummings (2001), in their study with students at the University of Canada, also found that family and friends have an important place in providing social support. As Türkdoğan and Duru (2012) stated, the permanent presence of family and friends in lives of university students may also provide an explanatory perspective on why they rank family first among the social support systems in meeting needs in the present study. This finding may also be related to the fact that social support systems of family and friends have less changing and more stable characteristics in human life. In addition, the family's role as a support provider in the development of the individual from the moment s/he was born may have led the family to be the primary choice.

Actually, that a friend ranks before a relative in meeting needs for safety, love-belonging, and respect may be due to the fact that relatives in Turkish society play a controlling role in the lives of individuals (Aksoy, 2011). When the conditions of the study group are evaluated, the point that they have more contact with their friends than their relatives as a result of the university environment can be considered one of the factors of that finding. In addition, due to the age of university students, according to Erikson's psychosocial development theory, it can be said that they are in a period of being alone rather than gaining camaraderie. Young people who have difficulty associating with others are likely to fall into unhealthy psychological loneliness (Senemoğlu, 2013). That university students who have difficulty finding friends become more depressed (Özdel, Bostancı, Özdel, Oguzhanoğlu, 2002) shows that friendships have an important place in lives of university students. Friends can be said to contribute to meeting the need for intimacy of university students in a healthy way. As a matter of fact that teacher/school and society are at the bottom of the rankings in meeting the needs of safety, love-belonging, and esteem may be due to the fact that these social support systems are the ones in which the individual is more distant. However, it is an important finding of the research that the relative ranks before the teacher/school and society, especially when it comes to meeting the need for esteem. This can be interpreted as an indication of the importance attributed to relatives in Turkish culture. Actually, that relative ranks after a friend may be related to the fact that university students spend more time with their friends in accordance with the period of life they are in.

That the relative ranks after the family in meeting the physiological needs may be due to the fact that relatives may be seen by the students as an important mechanism that can provide financial support after the family. Because of the university students and their friends are agemates and they get financial support from their parents may be among the reasons why friends ranking in meeting physiological needs is low among university students.

Another finding of the research is that, in meeting the need for self-actualization, teacher and society are preferred over the relative in the last rank. The teacher is expected to see the student as a whole, not from a narrow perspective (Farmer, 1984). In this way, the teacher can take a supporting role in exposing the student's potential. According to Ercoşkun and Nalçacı (2005), teachers contribute to the student's self-actualization process by creating appropriate learning environments. Students who can communicate effectively with their teacher are expected to increase their positive behavior (Hoşgörür, 2006). In the research, that the teacher ranks higher than relatives in meeting the need for self-actualization when compared to other needs can be considered related to the importance attributed to the teacher within the education system (Sünbül, 1996). In addition, the fact that teachers are a source of identification for students suggests that the teacher is an important factor for the student. In fact, that relatives rank last in the process of self-actualization can be considered a factor that comes from living as a nuclear family. This result may be due to limited contact with relatives in the nuclear family, while contact with relatives was greater in living as a wider family.

One of the first concepts that come to mind when it comes to teachers and schools is academic achievement. As a matter of fact, Parickova (1982) considers the increase in academic achievement as a factor that increases the level of self-actualization of the individual (as cited in Akbaş, 1989). Even a small success can be encouraging for the student who wants to continue learning (Crump, 1995). In their study, Yıldırım and Ergene (2003) suggest that family and teacher significantly predict academic success, and this supports that family and teachers have an important role in revealing the potential of the individual in the current study. Furthermore, the role of school in developing social relations becomes important when it is thought that developing social relations will help them use social support resources effectively (Terzi, 2008). Shaughnessy et al. (2018) have stated that family, teachers, and psychological counselors should be sensitive about the fulfilment of the basic needs of students; otherwise the situation may have negative repercussions on students' school life.

This study is limited to 347 university students studying at Sakarya University, Faculty of Education. Application to individuals with different demographic characteristics in future researches may change the results of the research. Especially, it may be suggested to carry out a comparative study on which social support systems meet the needs of the elderly and the young. In the survey, the needs of individuals are the five basic needs in Maslow's hierarchy of needs (physiological, safety, love-belonging, esteem, and self-actualization), and the social support systems are family, relative, friend, teacher/school, and society. The work can be expanded by incorporating knowing-understanding and aesthetic needs added to Maslow's hierarchy of needs later (İnceoğlu, 2004), or by using different social support classification systems. In addition, the factors that make the family the first to satisfy the needs can be determined by identifying the socio-economic conditions of the families in the following studies. It is thought that the correct determination of the current situation of individuals in the process of meeting their needs will contribute significantly to the elimination of deficiencies in meeting these needs.

REFERENCES

Akbaş, A. (1989). Ergenlerin kendini gerçekleştirme düzeylerini etkileyen bazı faktörler. *Ondokuz Mayıs Üniversitesi Eğitim Fakültesi Dergisi, 8*(1), 1-12. Retrieved from https://dergipark.org.tr/en/download/article-file/188143

Aksoy, İ. (2011). Türklerde aile ve çocuk eğitimi. *Uluslararası Sosyal Araştırmalar Dergisi, 4*(16), 11-19. Retrieved from https://www.sosyalarastirmalar.com/cilt4/sayi16_pdf/aksoy_ilhan.pdf

Aksüllü, N., & Doğan, S. (2004). Relationship of social support and depression in institutionalized and non-institutionalized elderly. *Anatolian Journal of Psychiatry*, *5*, 76-84. Retrieved from https://search.proquest.com/openview/777659b5add12fe4592a68b9685eef8a/1?cbl=136214&pq-origsite=gscholar

Anıl, D., & Güler, N. (2006). İkili karşılaştırma yöntemi ile ölçekleme çalışmasına bir örnek. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *30*, 30-36. Retrieved from https://dergipark.org.tr/en/download/article-file/87655

Anıl, D., & İnal, H. (2019). *Psikofizikte ölçekleme uygulamaları*. Ankara: Pegem Akademi Yayıncılık.

Arslantaş, H., & Ergin, F. (2011). 50-65 yaş arasındaki bireylerde yalnızlık, depresyon, sosyal destek ve etki eden faktörler. *Turkish Journal of Geriatrics, 14*(2), 135-144

Baykul, F., & Turgut, Y. (1992). *Ölçekleme teknikleri*. Ankara: ÖSYM Yayınları.

Bozo, Ö., Toksabay, N. E., & Kürüm, O. (2009). Activities of daily living, depression, and social support among elderly Turkish people. *The Journal of Psychology, 143*(2), 193-206. doi: 10.3200/JRLP.143.2.193-206

Burger, J. (2006). *Kişilik: Psikoloji biliminin insan doğasına dair söyledikleri*. İstanbul: Kaknüs Psikoloji.

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, CA: Sage.

Crump, C. A. (1995). *Motivating students: A teacher's challenge*. Paper presented at the Sooner Communication Conference, Norman, Oklahoma.

Cüceloğlu, D. (2003). *İnsan ve davranışı*. İstanbul: Remzi.

Doğan, T. (2016). Psikolojik belirtilerin yordayıcısı olarak sosyal destek ve iyilik hali. *Türk Psikolojik Danışma ve Rehberlik Dergisi, 3*(30), 30-44. Retrieved from https://dergipark.org.tr/tr/download/article-file/200121

Dwyer, A. L., & Cummings A. L. (2001). Stress, self-efficacy, social support, and coping strategies in university students. *Canadian Journal of Counselling, 35*(3), 208-220. Retrieved from https://cjc-rcc.ucalgary.ca/article/view/58672/44160

Engin, A. O., Özen, Ş., & Bayoğlu, V. (2009). Öğrencilerin okul öğrenme başarılarını etkileyen bazı temel değişkenler. *Sosyal Bilimler Enstitüsü Dergisi*, (3), 125-156. Retrieved from https://www.kafkas.edu.tr/dosyalar/sobedergi/file/003/03%20(9).pdf

Ercoşkun, M. H., & Nalçacı, A. (2005). Öğretimde psikolojik ihtiyaçların yeri ve önemi. *Kazım Karabekir Eğitim Fakültesi Dergisi*, (11), 353-370. Retrieved from https://dergipark.org.tr/en/download/article-file/31453

Farmer, R. (1984). Humanistic education and self-actualization theory. *Education, 105*(2), 162-172.

Gençtanırım, D. (2013). Bireysel farklılıklar. In Ş. I. Terzi (Ed.), *Eğitim psikolojisi* (pp. 251-286). Ankara: Pegem Akademi.

Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.

Haskan-Avcı, Ö., & Yıldırım, İ. (2014). Ergenlerde şiddet eğilimi, yalnızlık ve sosyal destek. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 29*(1), 157-168. Retrieved from https://dergipark.org.tr/tr/download/article-file/87085

Hoşgörür, V. (2006). İletişim. In Z. Kaya (Ed.), *Sınıf yönetimi* (pp. 151-179). Ankara: Pegem A Yayınları.

İnceoğlu, M. (2004). *Tutum, algı, iletişim*. Ankara: Elips Kitap.

Karasar, N. (1998). *Bilimsel araştırma yöntemi*. Ankara: Nobel Yayıncılık.

Khallad, Y., & Jabr, F. (2016). Effects of perceived social support and family demands on college students' mental well-being: A cross-cultural investigation. *International Journal of Psychology, 51*(5), 348-355. doi: 10.1002/ijop.12177

Lin, J., Thompson, M. P., & Kaslow, N. J. (2009). The mediating role of social support in the community environment psychological distress link among low-income African American women. *Journal Of Communıty Psychology, 37*(4), 459-470. doi: 10.1002/jcop.20307

Mallinckrodt, B. (1988). Students retention, social support, and dropout intention: Comparison of black and white students. *Journal of Counseling Psychology*, *129*(1), 60-64.

Maslow, A. H. (1970). *Motivation and Personality* (2nd ed.). NewYork: Harper and Row.

McConnell, J. V., & Philipchalk, R. P. (1992). *Understanding human behaviour* (7th ed.). Florida: International Edition.

Özdel, L., Bostancı, M., Özdel, O., & Oğuzhanoğlu, N. K. (2002). Üniversite öğrencilerinde depresif belirtiler ve sosyodemografik özelliklerle ilişkisi. *Anadolu Psikiyatri Dergisi, 3*, 155-161. Retrieved from https://www.researchgate.net/profile/Nalan_Oguzhanoglu/publication/265922906_Universite_ogrencilerinde_depresif_belirtiler_ve_sosyodemografik_ozelliklerle_iliskisi/links/551514bb0cf260a7cb2e7cc8.pdf

Plotnik, R. (2009). *Psikolojiye giriş* (Trans. T. Geniş). İstanbul: Kaknüs Yayınları.

Rahat, E., & İlhan, T. (2016). Coping styles, social support, relational self-construal, and resilience in predicting students' adjustment to university life. *Educational Sciences: Theory & Practice*, *16*(1), 187-208. doi: 10.12738/estp.2016.1.0058

Roediger, H. L., Capaldi, E. D., Paris, S. G., Polivy, J., & Herman, C. P. (1996). *Psychology* (4th ed.). Minnesota: Best West Publishing Company.

Sarı, İ., Yenigün, O., Altıncı, E., & Öztürk, A. (2011). Temel psikolojik ihtiyaçların tatmininin genel öz yeterlik ve sürekli kaygı üzerine etkisi (Sakarya üniversitesi spor yöneticiliği bölümü örneği). *Spormetre Beden Eğitimi ve Spor Bilimleri Dergisi, IX*(4). 149-156. Retrieved from https://dergipark.org.tr/tr/download/article-file/602163

Seker, S. E. (2014). Maslow'un ihtiyaçlar piramiti. *YBS Ansiklopedisi, 1*(1), 43-45. Retrieved from http://ybsansiklopedi.com/wp-content/uploads/2014/10/15.Maslow%E2%80%99un-%C4%B0htiya%C3%A7lar-Piramiti.pdf

Senemoğlu, N. (2013). *Gelişim, öğrenme ve öğretim, kuramdan uygulamaya*. Ankara: Yargı yayınevi.

Shaughnessy, M. Moffitt, B., & Cordova, M. (2018). Maslow, basic needs and contemporary teacher training ıssues. *Archives of Current Research International, 14*(4), 1-7. doi: 10.9734/ACRI/2018/42858

Slavin, R. E. (2013). *Educational psychology theory and practice* (10th ed.), (Trans. G. Yüksel). Ankara: Nobel Yayıncılık.

Sünbül, A. M. (1996). Öğretmen niteliği ve öğretimdeki rolleri. *Kuram ve Uygulamada Eğitim Yönetimi, 8*(8), 597-608. Retrieved from https://www.pegem.net/dosyalar/dokuman/1240-20120208173433-sunbul.pdf

Terzi, Ş. (2008). Üniversite öğrencilerinin psikolojik dayanıklılıkları ve algıladıkları sosyal destek arasındaki ilişki. *Türk Psikolojik Danışma ve Rehberlik Dergisi*, *3*(29), 1-11. Retrieved from http://turkpdrdergisi.com/index.php/pdr/article/view/239/169

Tinajero, C., Martínez-López, Z., Rodríguez, M. S., Guisande M. A., & Páramo, M. F. (2015), Gender and socioeconomic status differences in university students' perception of social support. *European Journal of Psychology of Education*, *30*, 227-244. doi: 10.1007/s10212-014-0234-5

Türkdoğan, T., & Duru, E (2012). Üniversite öğrencilerinde öznel iyi oluşun yordanmasında temel ihtiyaçların karşılanmasının rolü. *Kuram ve Uygulamada Eğitim Bilimleri*, *12*(4), 2429-2446.

Yıldırım, İ., & Ergene, T. (2003). Lise son sınıf öğrencilerinin akademik başarılarının yordayıcısı olarak sınav kaygısı, boyun eğici davranışlar ve sosyal destek. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *25*, 224-234. Retrieved from https://dergipark.org.tr/tr/download/article-file/87880

Yıldırım, İ. (1997). Algılanan sosyal destek ölçeğinin geliştirilmesi güvenirliği ve geçerliği. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *13,* 81-87. Retrieved from https://dergipark.org.tr/en/download/article-file/88127

Yılmaz, E., Yılmaz, E., & Karaca, F. (2008). Üniversite öğrencilerinin sosyal destek ve yalnızlık düzeylerinin incelenmesi. *Genel Tıp Dergisi, 18(*2), 71-79. Retrieved from http://geneltip.org/upload/sayi/56/GTD-00447.pdf

Üniversite Öğrencilerinin İhtiyaçlarının Sosyal Destek Sistemlerinden Karşılanmasının İncelenmesi

Giriş

İnsan davranışlarının altında yatan etmenleri araştırmak psikoloji biliminin uzun yıllardır konusu olmuştur. Maslow’un psikolojik ihtiyaçlara ilişkin yaklaşımı bu alanda en bilindik teorilerdendir (Roediger, Capaldi, Paris, Polivy & Herman, 1996). Maslow, ihtiyaçlara ilişkin hiyerarşik bir yapıyı ilk olarak 1943 yılında ortaya koymuş, 1954 yılında yayınladığı “Motivasyon ve Kişilik” kitabında ihtiyaçlara ilişkin teorisini tanıtmıştır (Maslow,1970). İnsan güdülerinin hayvan güdülerinden ayrılan noktalarının olduğunu savunmuş ve insan güdülerini bir piramit şeklinde ifade etmiştir. Bu piramidin temelinde biyolojik güdüler, en üstünde ise psikolojik güdüler yer almaktadır (Cüceloğlu, 2003). Maslow’un bu teorisinin günümüzde halen geçerliliğini koruduğunu söylemek mümkündür. İhtiyaçlar hiyerarşisi fizyolojik ihtiyaçlar, güvenlik ihtiyacı, sevgi ve ait olma ihtiyacı, saygı ihtiyacı ve kendini gerçekleştirme ihtiyacı olmak üzere beş başlık altında incelenmektedir (Plotnik, çev. 2009).

Günümüzde ihtiyaçlar hiyerarşisine bakıldığında ilk basamaktan son basamağa kadar ihtiyaçların karşılanmasında bir başkasına ihtiyaç duyulduğu söylenebilir. Bu noktada da sosyal destek kavramı işin içine girmektedir. Sosyal desteğin tanımına bakıldığında ise Yıldırım’a (1997) göre aile, aile çevresi, arkadaşlar, karşı cins ile ilişkiler, öğretmenler, iş arkadaşları, komşular, ideolojik, dinsel veya etnik gruplar ile bireyin içinde yaşadığı toplum gibi faktörlerin o bireyin sosyal destek kaynaklarını oluşturduğu söylenebilir. Sosyal destek kişinin yaşamda karşılaştığı güçlüklerle başa çıkabilmesine olanak sağlamakta, koruyucu bir tampon görevi görmektedir (Arslantaş & Ergin, 2011; Lin, Thompson, & Kaslow, 2009; Terzi, 2008).

Bu çalışmanın çalışma grubunu da içine alan üniversiteler, öğrencilerin yetişkinliğe ve meslek hayatına adım atmasını sağlayan önemli kurumlardan biri olma görevini üstlenmektedir. Özellikle Türkiye’de üniversiteler birçok öğrencinin ailesinden bağımsız olarak hayatlarındaki süreci kendilerinin organize etmeye başlamasına neden olan yaşantıları sunması gibi önemli bir role de sahiptir. Daha sağlıklı bireyler yetiştirmek için hangi sosyal destek sistemlerinin hangi ihtiyaçları karşılamada işlevsel olduğunu belirlemek, işlevsel olamayan sosyal destek sistemlerinin işlevsel hale gelmesi için gerekenleri ortaya koymanın ruh sağlığı daha yerinde bireyler yetiştirme konusuna fayda sağlayacağı düşünülmektedir. Bu bağlamda yapılan çalışmada üniversite öğrencilerinin ihtiyaçlarını sosyal destek sistemlerinden karşılama düzeylerinin sıralama yargılarına dayalı olarak ölçekleme yöntemi ile belirlenmesi amaçlanmıştır. Bu amaçla öğrencilerin beş temel ihtiyacını (fizyolojik, güvenlik, ait olma-sevme, saygı ve kendini gerçekleştirme), sosyal destek sistemleri (aile, akraba, arkadaş, öğretmen-okul, toplum) tarafından kaçıncı sırada karşılandığı belirlenmeye çalışılmıştır.

Yöntem

Üniversite öğrencilerinin ihtiyaçlarını sosyal destek sistemlerinden karşılama düzeylerinin sıralama yargılarına dayalı olarak ölçekleme yöntemi ile belirlenmesi amaçlanan bu çalışmada tarama modeli kullanılmıştır. Araştırma Sakarya Üniversitesi Eğitim Fakültesi’nde öğrenim gören toplam 347 üniversite öğrencisi ile gerçekleştirilmiştir. Araştırmada kullanılan veri toplama araçları araştırmacılar tarafından oluşturulan kişisel bilgi formu ve Maslow'un ihtiyaçlar hiyerarşisinde yer alan beş temel ihtiyacın (fizyolojik, güvenlik, ait olma-sevme, saygı ve kendini gerçekleştirme), sosyal destek sistemleri (aile, akraba, arkadaş, öğretmen-okul, toplum) tarafından kaçıncı sırada karşılandığını belirlemeye yönelik bir sıralama çizelgesinden oluşmaktadır. Verilerin analizinde ise Sıralama yargılarına dayalı ölçekleme yöntemi kullanılmıştır.

Nesneleri, bireyleri, durumları ya da yöntemleri (uyarıcı) puanlayıcılar tarafından belli bir kurala göre sıralayarak veri toplama, sosyal bilimlerde sıklıkla kullanılan bir yöntemdir. Ancak bu tür verilerin analiz edilmesinde çoğunlukla ilk sıraya yazılan uyarıcı dikkate alınmaktadır. Sıralama yargılarına dayalı olarak ölçekleme tekniği ise sadece birinci sıradaki uyarıcı değil yapılan tüm sıralamaları dikkate alarak verilerin çözümlenmesinde kullanılabilecek bir yöntemdir (Baykul & Turgut, 1992; Guilford, 1954). Sıralama yargılarına dayalı ölçekleme yöntemi, her bir uyarıcının kaçıncı sırada tercih edildiğine ilişkin Sıralama Yargıları Frekans Matrisi'nin oluşturulması ile başlamaktadır. Sıralama yargıları frekans matrisi üzerinden her bir uyarıcının diğer uyarıcılara göre ikili karşılaştırmada tercih edilme olasılığı $P_{j>k} = \sum_{i=1}^{n} \left[f_{ji} (f_{k<i} + f_{ki} / 2 \right]$ formülü ile hesaplanmakta ve uyarıcıların ikili karşılaştırmalarında birbirlerine göre tercih edilme olasılıklarını gösteren Sıralama Yargıları Olasılık Matrisi oluşturulmaktadır. Bu aşamadan sonra sıralama yargılarına dayalı yöntemlere ilişkin hesaplama aşamaları şu şekildedir:

1. Sıralama yargıları olasılık matrisindeki her bir birime karşılık gelen birim normal dağılım için z değerleri hesaplanarak Birim Normal Sapmalar Matrisi oluşturulur.
2. Her sütunun ortalaması alınarak ortalama z değerlerine ulaşılır.
3. En küçük ortalama z değeri sıfıra denk gelecek şekilde ötelenir.
4. Elde edilen ötelenmiş ortalama z değerleri her bir uyarıcıya ait ölçek değerlerini oluşturmaktadır (Anıl ve Güler, 2006; Baykul ve Turgut, 1992; Guilford 1954).

Sonuç ve Tartışma

Araştırma sonucunda üniversite öğrencilerinin tüm ihtiyaçlarını karşılamada aileyi öncelikli sosyal destek sistemi olarak gördükleri; güvenlik, ait olma-sevme ve saygı ihtiyaçlarının karşılanmasında sıralamanın değişmediği [aile (1), arkadaş (2), akraba (3), öğretmen/okul (4), toplum (5)]; fizyolojik ihtiyaçların karşılanmasında ise akrabanın arkadaştan daha çok tercih edildiği bulunmuştur.

Yılmaz, Yılmaz ve Karaca'nın (2008) yaptıkları çalışma da üniversite öğrencilerinin arkadaş ve özel insandan daha fazla aileyi sosyal destek olarak gördüklerini ortaya koymaktadır. Ailenin önemine vurgu yapan yönüne bakıldığında; Engin, Özen ve Bayoğlu, (2009) eğitim ve öğretim etkinliklerinde gerekli olan temel ihtiyaçların %91.3'ünün her zaman aileleri tarafından karşılandığını belirtmektedirler. Dwyer ve Cummings (2001), Kanada Üniversitesi'ndeki öğrencilerle yaptıkları çalışmada da aile ve arkadaşların sosyal desteği sağlamada önemli bir yere sahip olduğunu bulmuşlardır. Aile ve arkadaşların üniversite öğrencilerinin hayatında kalıcı bir şekilde var olmaları (Türkdoğan & Duru 2012); yapılan çalışmada da ihtiyaçların karşılanmasında sosyal destek sistemleri arasında ilk sıralarda yer almalarına açıklayıcı bir bakış açısı sunabilir. Bu bulgunun; aile ve arkadaş sosyal destek sistemlerinin insan hayatında daha az değişen, daha sabit özelliklere sahip olması ile de ilgili olduğu düşünülebilir. Ayrıca ailenin bireyin gelişiminde doğduğu andan itibaren destek sağlayıcı bir rolünün olması ailenin öncelikli tercih olmasına etki etmiş olabilir.

Güvenlik, ait olma ve saygı ihtiyaçlarının karşılanmasında arkadaşın akrabadan önce sıralamada yer alması Türk toplumunda akrabaların bireylerin hayatlarında kontrol edici bir rol üstlenmesinden (Aksoy, 2011) kaynaklanıyor olabilir. Çalışma grubunun içinde bulunduğu şartlar değerlendirildiğinde ise; üniversite ortamının bir sonucu olarak arkadaşları ile akrabalarından daha fazla temas ediyor olmaları bu bulguyu etkileyen faktörlerden birisi olarak değerlendirilebilir. Ayrıca üniversite öğrencilerinin yaşları itibariyle; Erikson'un psikososyal gelişim kuramına göre yakınlığa karşı yalıtılmışlık döneminde oldukları söylenebilir. Başkalarıyla ilişki kurmakta güçlük çeken gencin sağlıksız bir psikolojik yalnızlık içine girmesi muhtemeldir (Senemoğlu, 2013). Arkadaş bulmakta güçlük yaşayan üniversite öğrencilerinin daha fazla depresyona girmesi (Özdel, Bostancı, Özdel, Oğuzhanoğlu, 2002), arkadaş ilişkilerinin üniversite öğrencilerinin hayatında önemli bir yere sahip olduğunu göstermektedir. Arkadaşların üniversite öğrencilerinin yakınlık ihtiyacının sağlıklı bir şekilde giderilmesinde katkı sağladığı söylenebilir. Güvenlik, ait olma ve saygı ihtiyaçlarının karşılanmasında öğretmen/okul ve toplumun sıralamada sonlarda yer alması ise bu sosyal destek sistemlerinin bireyin daha mesafeli olduğu sosyal destek sistemleri olmasından kaynaklanıyor olabilir. Ancak özellikle saygı ihtiyacının karşılanması hususunda akrabanın öğretmen/okul ve toplumdan önce sıralamada yer alması araştırmanın önemli bir bulgusudur. Bu da Türk kültüründe akrabaya verilen önemin bir göstergesi olarak yorumlanabilir. Akrabanın sıralamada arkadaştan sonra yer alması ise üniversite öğrencilerinin içinde bulundukları yaşam dönemi gereğince daha çok arkadaşları ile zaman geçirmeleri ile ilişkili olabilir.

Fizyolojik ihtiyaçların karşılanmasında akrabanın, aileden sonra sıralamada yer alması ise öğrenciler tarafından arkadaşla kıyaslandığında; akrabaların maddi desteği aileden sonra sağlayabilecek önemli bir mekanizma olarak görülmesi olabilir. Üniversite öğrencilerinin edindikleri arkadaşların kendi yaş aralıklarında olması ve maddi desteklerini ailelerinden sağlıyor olmaları, üniversite öğrencileri arasında arkadaşın fizyolojik ihtiyaçları karşılamadaki sıralamasının düşmesinin nedenleri arasında yer alabilir.

Araştırmanın bir diğer bulgusu ise kendini gerçekleştirme ihtiyacının karşılanmasında akrabanın son sırada daha çok tercih edilerek öğretmen ve toplumun sıralamada akrabanın önüne geçmesidir. Öğretmenden öğrencisini dar bir bakış açısı ile değil bir bütün olarak görmesi beklenmektedir (Farmer, 1984). Bu sayede öğretmen öğrencinin potansiyelini açığa çıkarma konusunda destekleyici bir rol edinebilmektedir. Ercoşkun ve Nalçacı'ya (2005) göre öğretmenler uygun öğrenme ortamları oluşturarak öğrencinin kendini gerçekleştirme sürecine katkı sağlamaktadırlar. Öğretmeni ile etkili iletişim kurabilen öğrencilerin olumlu davranışlarını arttırmaları beklenir (Hoşgörür, 2006). Yapılan araştırmada diğer ihtiyaçlarla kıyaslandığında kendini gerçekleştirme ihtiyacının karşılanmasında öğretmenin sıralamada akrabadan öne geçmesi eğitim sistemi içerisinde öğretmene önemli bir değer atfedilmesi (Sünbül, 1996) ile ilişkili olarak değerlendirilebilir. Ayrıca öğretmenlerin öğrenciler için bir özdeşim kaynağı olması öğretmenin öğrenci için önemli bir faktör olması hususunu göstermektedir. Akrabaların kendini gerçekleştirme sürecinde son sırada yer alması ise çekirdek aile olarak yaşamanın getirdiği bir faktör olarak değerlendirilebilir. Büyük aile olarak yaşamada akrabalarla temas daha fazlayken çekirdek ailede akraba temasının sınırlı olması da bu sonucu ortaya çıkarmış olabilir.

Öğretmen ve okul denildiğinde ilk akla gelen kavramdan birisi ise akademik başarıdır. Nitekim Parickova (1982) da akademik başarının artmasını bireyin kendini gerçekleştirme düzeyini arttıran bir faktör olarak değerlendirmektedir (akt., Akbaş, 1989). Öğrenmeye devam etmek isteyen öğrenci için küçük bir başarı bile cesaretlendirici olabilmektedir (Crump, 1995). Yıldırım ve Ergene'nin (2003) yaptıkları çalışmada aile ve öğretmen desteğinin akademik başarıyı önemli ölçüde yordaması mevcut çalışmada da bireyin potansiyelini açığa çıkarma hususunda aile ve öğretmenin önemli role sahip oluşunu desteklemektedir. Ayrıca sosyal ilişkilerin geliştirilmesinin sosyal destek kaynaklarını etkili bir şekilde kullanmalarına yardımcı olacağı düşünüldüğünde (Terzi, 2008), okulun sosyal ilişkileri geliştirmedeki rolü de önem kazanmaktadır. Shaughnessy, Moffitt ve Cordova (2018) aile, öğretmenler ve psikolojik danışmanların öğrencilerin temel ihtiyaçlarının karşılanması konusunda hassas davranmaları gerektiği aksi durumun öğrencilerin okul hayatına olumsuz yansımalarının olabileceği görüşünü belirtmişlerdir.

Bu çalışma Sakarya Üniversitesi Eğitim Fakültesi'nde öğrenim gören 347 üniversite öğrencisi ile sınırlıdır. Gelecek araştırmalarda farklı demografik özelliklere sahip bireylere uygulama yapılması araştırmanın sonuçlarını değiştirilebilir. Özellikle yaşlı ve gençlerin ihtiyaçlarını hangi sosyal destek sistemlerinden karşıladıklarına ilişkin karşılaştırmalı bir çalışma yapılması önerilebilir. Yapılan araştırmada bireylerin ihtiyaçları Maslow'un ihtiyaçlar hiyerarşisinde yer alan beş temel ihtiyaç (fizyolojik, güvenlik, ait olma-sevme, saygı ve kendini gerçekleştirme) ve sosyal destek sistemleri aile, akraba, arkadaş, öğretmen-okul ve toplum olarak ele alınmıştır. Maslow'un ihtiyaçlar hiyerarşisine sonradan eklenen bilme ve anlama ile estetik ihtiyaçları da (İnceoğlu, 2004) dahil edilerek ya da farklı sosyal destek sınıflama sistemleri kullanılarak çalışma genişletilebilir. Ayrıca ilerleyen çalışmalarda ailelerin sahip olduğu sosyo-ekonomik koşullar da tespit edilerek aileyi ihtiyaçların doyurulmasında birinci yapan faktörler belirlenebilir. Bireylerin ihtiyaçlarını karşılama sürecinde mevcut durumlarının doğru biçimde tespit edilmesinin, bu ihtiyaçları karşılama konusundaki eksikliklerin giderilmesine önemli katkı sağlayacağı düşünülmektedir.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(4); 362-373

The Impact of Covariate Variables on Kernel Equating under the Non-equivalent Groups

Çiğdem AKIN ARIKAN *

Abstract

This study aims to use covariate variables correlated with the test scores instead of common items for non-equivalent groups with covariates (NEC) design in kernel equating. This study used the 2016 Monitoring and Evaluation of Academic Skills Project in Turkey. The study used data from 6,000 students, randomly selected from the Turkish Ministry of National Education’s current student data. Three thousand of the students took form A, and 3,000 of them took form B. The data include mathematics test scores and consist of 18 items, nine of which are the first items, and nine of which are anchor items. The equated scores from the NEC design were compared with equated scores from the non-equivalent group (NEAT) design. From the equating results, the root mean squared difference (RMSD) and standard error of equating (SEE) values were calculated. The results showed that NEC design could produce lower standard errors compared with the NEAT design, and the least RMSD was provided by NEAT PSE methods and NEC methods. The general result of this research is that test forms can be equated using covariates when there are no anchor items.

Key Words: NEC design, NEAT design, covariate variables, SEE, RMSD.

INTRODUCTION

In the last century, new equating methods and designs have been developed in test equating. One of these methods, kernel equating, was first defined by Holland and Thayer (1989) and then developed by von Davier, Holland, and Thayer (2004). Kernel equating is an equipercentile score equating technique in which discrete score probabilities are continuized, and score probabilities are equated (von Davier et al., 2006). In this regard, Kernel equating can be considered a developed form of traditional equating techniques. There are two reasons for this view. First, it makes data consistent by using presmoothing, yields smaller errors when compared to other methods by smoothing the transformation of data, and is less dependent on sample variability. Second, kernel equating can be applied to all designs and equating functions (von Davier et al., 2004). Kernel equating consists of five steps, namely, presmoothing, estimation of score probabilities, continuization, equating, and calculating the standard error of equating. Also, in kernel equating, both linear and equipercentile equating functions are used (von Davier et al., 2006).

In test equating, there are various group designs, such as single group design, equivalent group design, and nonequivalent groups with anchor test (NEAT) (Kolen & Brennan, 2014; von Davier et al., 2004). NEAT design is one of the most frequently used designs in the literature. Post-stratification equating (PSE) and chained equating (CE) that were used in this study and Levine observed-score linear equating methods are within the scope of NEAT design in kernel equating (von Davier et al., 2004). Two different test forms, namely, X and Y, in addition to the anchor test A, are taken by two different populations in NEAT design. For a detailed theoretical explanation of all methods, readers are encouraged to look at Chen and Holland (2010), von Davier et al. (2004), and von Davier, Fournier-Zajac, and Holland (2007). To estimate the distribution of X in group I and the distribution of Y in group II, the anchor test A is used by PSE. In this regard, the conditional distribution of X, given A, and the conditional distribution of Y, given A, constitute the population invariant. Afterward, it post-stratifies the distributions of both X and Y in a target population T (a synthetic population of Group I

* Assist. Prof., Ordu University, Faculty of Education, Ordu-Turkey, akincgdm@gmail.com, ORCID ID: 0000-0001-5255-8792

To cite this article:
Akın-Arıkan, Ç. (2020). The impact of covariate variables on kernel equating under the non-equivalent group design. *Journal of Measurement and Evaluation in Education and Psychology, 11*(4), 362-373. doi: 10.21031/epod.706835.
Received: 20.03.2020
Accepted: 15.11.2020

and Group II). In CE, the anchor is used as a part of a chain by linking X to A in group I and then A to Y in group II (von Davier et al., 2004).

In NEAT design, anchor items are used to adjust the differences in ability between the groups. However, anchor items might not appear in the forms of all the test programs or standardized tests. Additionally, test forms might not be equated since it is hard for groups that take different test forms to be equivalent in practice. For instance, if there are no anchor items in non-equivalent groups, significant covariates can be used instead of anchor items and the design is called the non-equivalent groups with covariates (NEC) design (Wiberg & Branberg, 2015). Wiberg and Branberg (2015) also used NEATNEC design, which is a mixture of the NEC design and the NEAT design in their research. NEC design is used in the post-stratification equating method (NEATPSE) of kernel equating and the populations of two groups are weighted to generate a synthetic population to equating the test scores (Andersson & Wiberg, 2017; von Davier et al., 2004). In fact, it is assumed that enhancing the correlation of the covariances used in NEC design with the test will result in similar numbers to that of NEAT CE and NEAT PSE (Wiberg & Branberg, 2015). When the literature was examined, it was seen that different variables (e.g. test scores and gender) were used as covariates (e.g. Branberg & Wiberg, 2011; Wiberg & Branberg, 2015; Yurtçu, 2018). The basic assumption of the NEC design is that these covariates can be used for the ability variability between two groups (Wiberg & Branberg, 2015). In the NEC design, the other critical point is this: for both groups the conditional distribution of the test scores with the covariates is the same (Wiberg & Branberg, 2015). For this assumption, the time, equating between test forms plays a critical role in the results. Therefore, bias can be avoided by adding a variable that affects its change over time to the equating model. (Wiberg & Branberg, 2015).

The literature about covariates in equating revealed that the number of studies are limited. First studies used different variables in test equating, paving the way for future studies to be conducted with covariates (e.g. Cook, Eignor, & Schmitt, 1990; Holland, Dorans, & Petersen 2007; Livingston, Dorans, & Wright, 1990). As for recent studies, Branberg (2010) equated test forms with covariates and claimed that it is possible to use covariates instead of anchor items. Branberg and Wiberg (2011), first of all, conducted a regression analysis between the test scores and variables to determine the covariates, which were education and gender in the real data. The study results showed that by correcting for variations in the test score distributions of covariates, test equating could be improved. Similarly, Wiberg and Branberg (2015) concluded that NEC design is more accurate than the equivalent group design in kernel equating. In line with this conclusion, using both covariates and anchor items resulted in the smallest standard error of equating over a large range of test scores. The research conducted by Gonzalez et al. (2015) revealed that the Bayesian non-parametric model for equating makes many assumptions that used to be vital for test equating unnecessary, demonstrating that even when there is no covariant, equating is possible. Wiberg and von Davier (2017) also stated the effect of covariates in various administrations would aid in the process of ensuring equal testing for test-takers.

Wallin and Wiberg (2017) investigated the manner in which propensity scores affect equation results when covariates are great in number by comparing the kernel equating methods in NEAT and NEC design. Their research indicated that using propensity scores in kernel post-stratification and kernel chained equating methods increases precision and leads to greater results compared to the equivalent group designs. Moreover, the research showed great similarities with the results of the anchor test design. Yurtçu (2018) used covariates to obtain scores equated by using non-parametric Bayes techniques. According to the research, this model is more informative compared to the traditional methods. Also, covariants can be used instead of anchor items and in some cases, this model has been found to give more accurate results. Likewise, the equated scores obtained through this model were closer to the target. The limited number of studies on the topic indicates a gap in the literature. Likewise, it is necessary to conduct Item Response Theory (IRT) studies, as well as testing kernel equating methods, which are new approaches to the topic. In Turkey, large-scale standardized tests generally are used when making important decisions, such as the selection and placement of students in any kind of educational program. The Monitoring and Evaluation of Academic Skills Study (ABIDE) of the Republic of Turkey Ministry of National Education (MoNE) in Turkey uses large-scale testing to assess the students' mental skills in topics such as math, science, and social studies

(MoNE, 2016). With the exception of the Monitoring and Evaluation of Academic Skills Project, which contains anchor items, all the other exams in Turkey lack anchor items.

The Purpose of the Study

Exams, such as language exams, academic personnel, and postgraduate education entrance exams, have various validity periods so that results from different years can be used to apply for a master's degree, Ph.D. degree, research assistant role, teaching assignments, etc. The fact that the test scores are comparable and interchangeable brings forward the topic of equating test forms. Through the NEC design, equating the test forms will be possible, even in cases where there are no anchor items. Furthermore, designing these sorts of tests will be a guide to equating methods and determining precautions to be taken in case of the existence or non-existence of anchor items. Accordingly, the purpose of the study is to compare the results obtained through the PSE and CE equating methods, which are among the NEAT and NEC kernel equating method designs, and to determine the effect of gender and socioeconomic level as covariates in test equating.

METHOD

In the context of the study, scores obtained from math sub-tests in the 2016 ABIDE were equated by using NEAT and NEC designs with kernel chained equipercentile, kernel post-stratification equipercentile, kernel chained linear, and kernel post-stratification linear methods. Seeing that the existing methods and techniques were verified through real data, it is possible to claim that the research is descriptive.

Sample

The population of the study comprised 38,000 students, from 16,118 schools and 48,091 branches, which were accessed via the ABIDE, with an approximate number of 400 students per city in Turkey (MoNE, 2016). In the scope of the study, data for 6000 students were used, randomly selected from the current student data of MoNE, 3000 of whom took form A, and 3000 of whom took form B. Among the students, 1292 (43.07%) who took form A were female, while 1708 (56.93%) of them were male. Of those completing form B, 1518 (50.6%) were female, whereas 1482 (49.4%) were male.

Process and Data Collection Instruments

Research data consist of math test scores for eighth-graders as a part of the ABIDE. There are 20 items in the math sub-test of the ABIDE. Nine of these items are primary items; nine of them are anchor items, and two of them are pilot items. Furthermore, the project consists of three forms (A, B, and C) and 12 booklets (A1-A4, B1-B4, C1-C4). Each form of the project consists of nine primary items. Form A is connected to form B and C with nine items, while form B is connected to form C with nine items. To put it in a different way, primary items exist in the booklets of forms A, B, and C. Booklet 1 in Form A and Form B were used for this study, and each booklet consists of 18 items, including anchor items. However, the anchor items were determined as external anchors and were not included in the total score. There are partially scored items in the booklet. Among the partially scored items, category 2 was recoded as 1 and was transformed into 1-0 data.

To determine the covariates, the correlation values between math success and student variables in the questionnaire were analyzed in a report prepared for the ABIDE. The socioeconomic index, which had the highest correlation value ($r = .37$, $p < .05$), was chosen as the anchor variable (MoNE, 2016). To categorize the socioeconomic level index, which is a continuous variable, three levels: low, middle, and high, were created through a two-step cluster analysis. Another covariant of the study was gender. Socioeconomic status (SES) was coded *low* = 1, *middle* = 2, *high* = 3, whereas gender was coded as *female* = 1 and *male* = 2.

For booklet A, the correlation values of the test and variables were as follows: the correlation of the anchor test and the test is (r = .51, p < .05); the correlation of the test and the socioeconomic variable is (r = .21, p < .05); the correlation of the categorical socioeconomic index and the test is (r = .19, p < .05); and the correlation of gender and the test is (r = .05, p > .05). As for booklet B, the correlation values were as follows: the correlation of the anchor test and the test is (r = .49, p < .05); the correlation of the test and the socioeconomic variable is (r = .16, p < .05); the correlation of the categorical socioeconomic index and the test is (r = .15, p < .05); and the correlation of gender and the test is (r = .02, p > .05). In other words, the relationship between the test and gender is insignificant for the two booklets, but the other relationships are significant. Although the relationship between the test and gender is not significant, it was used in other studies (Branberg & Wiberg, 2011; Gonzalez et al., 2015; Liou et al., 2001, and Yurtçu, 2018). It was also taken as a covariate in this study.

Data Analysis

Firstly, the test scores were equated with a NEAT design in kernel equating. Afterward, gender and socioeconomic level variables were set as covariates, and then the test scores were equated with an NEC design. To equate the tests, the R (R Core Team, 2013) package “kequate” was employed (Andersson et al. 2013). The standard error of equating (SEE) and error of equating (RMSD-root mean squared deviation/error) were used for the evaluation.

RESULTS

Before equating the tests, descriptive statistics for booklets A and B were obtained. The findings are listed in Table 1.

Table 1. Descriptive Statistics of the Booklets

Statistics	A-main test	A-anchor test	B-main test	B-anchor test
N	3000	3000	3000	3000
Mean	2.59	3.52	2.91	2.47
Standard Deviation	1.60	1.85	1.79	1.68
Skewness	0.48	0.38	0.51	0.54
Kurtosis	-0.14	-0.23	-0.02	-0.11

According to Table 1, the score distribution of booklets A, B, and anchor tests are right-skewed. As the kurtosis coefficients of the score distribution are negative, it can be argued that the distribution is platykurtic (negative kurtosis). The skewness and kurtosis coefficients are between -1.00 and +1.00, so the data indicates normal distribution. Additionally, the mean of the tests is revealed to be low.

Booklet A of the subtest on the Monitoring and Evaluation of Academic Skills Project was equated to Booklet B with PSE-EQ (EQ-equipercentile), PSE-L (L-linear), CE-EQ (equipercentile), CE-L (linear), NEC-EQ SEX (gender), NEC-L SEX, NEC-EQ SES (socioeconomic status), NEC-L SES, NEC-EQ-SEX-SES, and NEC-L-SEX-SES. Equated scores are listed in Table 2.

In kernel equating, the selection of bandwidths is important. If a large bandwidth is used, the equating function gets close to the linear equation, whereas in turn, if a small bandwidth is used, the equating function gets close to the linear equation (von Davier et al., 2004).The results obtained with bandwidths (h) were h_X = 0.51 and h_Y = 0.54 for NEC-EQ (SES), h_X = 1697.17 and h_Y = 1592.48 for NEC-L(SES) h_X = 0.51 and h_Y = 0.54 for NEC-EQ (SEX), h_X = 1681.76 and h_Y = 1601.33 for NEC-L(SEX), h_X = 0.51 h_Y = 01544.18 NEC-EQ (SEX-SES), h_X = 0.51 and h_Y = 0.53 for NEAT PSE-EQ, h_X = 1701.38 and h_Y = 01589.34 for NEAT PSE-L, h_X = 0.51 and h_Y = 0.53 for NEAT-CE EQ, and h_X = 1682.62 and h_Y = 1601.04 for NEAT-CE L.

Table 2. Equated Scores Derived from Different Methods

Booklet A	NEC-SEX-EQ	NEC-SEX-L	NEC-SES-EQ	NEC-SES-L	NEC-SEX-SES-EQ	NEC-SEX-SES-L	NEAT-PSE	NEAT-PSE-L	NEAT-CE	NEAT-CE-L
0	0.18	0.24	0.14	0.20	0.16	0.21	0.10	0.17	0.02	0.04
1	1.23	1.20	1.16	1.14	1.20	1.17	1.12	1.11	0.95	0.93
2	2.18	2.15	2.10	2.08	2.14	2.12	2.07	2.04	1.83	1.83
3	3.11	3.10	3.02	3.01	3.07	3.07	3.00	2.97	2.72	2.73
4	4.04	4.05	3.94	3.95	4.00	4.03	3.91	3.91	3.62	3.63
5	4.97	5.01	4.86	4.89	4.95	4.98	4.81	4.84	4.56	4.52
6	5.92	5.96	5.79	5.83	5.92	5.94	5.71	5.78	5.51	5.42
7	6.89	6.91	6.75	6.77	6.94	6.89	6.61	6.71	6.44	6.32
8	7.89	7.86	7.75	7.71	7.98	7.84	7.53	7.64	7.37	7.22
9	8.92	8.81	8.82	8.64	9.00	8.80	8.55	8.58	8.40	8.11

Raw scores for booklet A were between 0.00-9.00, and the equated scores, based on different equation designs and methods, were in the raw score range. Scores that were equated via NEC-EQ with covariate SEX, NEC-L, and NEC-EQ with two covariates were larger than the raw scores for booklet A in the 0.00-4.00 range and smaller than the raw scores in the 5.00-9.00 range. The scores equated via NEC-L with sex as the covariate were bigger than the raw scores for booklet A in the 0.00-5.00 range and smaller than the raw scores in the 6.00-9.00 range; the scores equated via NEAT PSE, NEC-EQ with covariate SES, and NEC-L with covariate SES, were larger than the raw scores for booklet A in 0.00-3.00 range and smaller than the raw scores in 4.00-9.00 range. The scores equated via NEAT CE were larger than the raw scores for booklet A at 0 and smaller than the raw scores in the 1.00-9.00 range. According to the findings, it is possible to claim that the degree of difficulty is not the same throughout the scale, and it changes depending on the forms. The difference between the equated scores and raw scores are given in Figure 1.

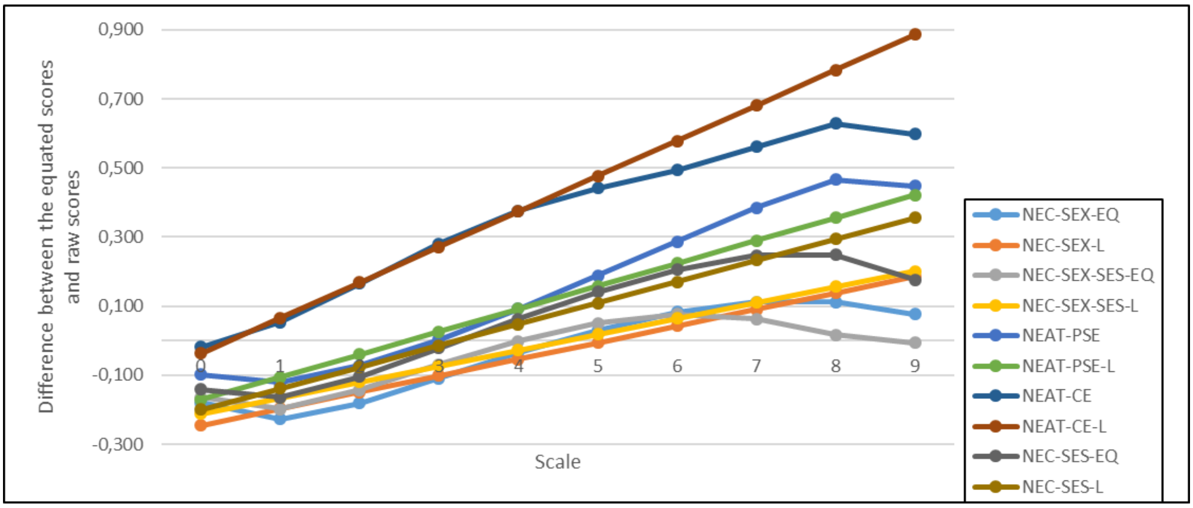


Figure 1. Difference Between Equated Scores and Raw Scores for NEC, NEAT PSE Equipercentile, NEAT PSE Linear, NEAT Chain Equipercentile, NEAT Chain Linear (*NEAT = non-equivalent groups with anchor test; PSE = poststratification; CE= Chain Equating equating; NEC = non-equivalent groups with covariates; EQ= equipercentile and L= linear, SES: socioeconomic status, SEX: Gender*)

In Figure 1, the raw scores and differences between the equated scores for the different equating methods and different equating data designs are displayed. For the NEAT design, PSE and CE with linear and equipercentile types; for the NEC design, equipercentile and linear with two covariates were used separately. The results show that the difference between equated scores and raw scores were smaller in cases where the NEC design was used than those of NEAT design. For linear equating, NEC design with sex and NEC design with sex-SES covariates gave similar results, whereas equipercentile equating yielded somewhat similar results, except for scores between 7-9. The NEC design (linear and equipercentile) with SES covariate gave similar results, except for scores between 8-9. The greatest gap between the scores equated with the NEC design and raw scores occurred when gender was used as a covariant. Also, NEAT CE methods (EQ and L) had large differences compared to other methods. Figure 2 shows the SEE values for the equating methods.

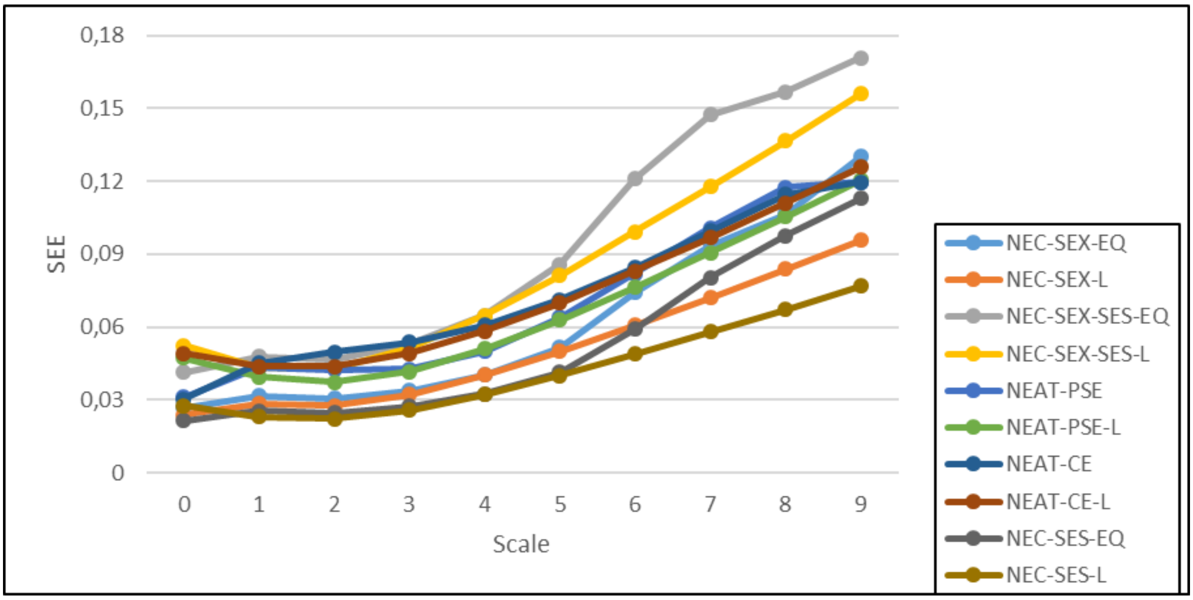


Figure 2. SEE Values for NEC, NEAT PSE Equipercentile, NEAT PSE Linear, NEAT Chain Equipercentile, NEAT Chain Linear (*NEAT = non-equivalent groups with anchor test; PSE = poststratification; CE= Chain Equating; NEC = non-equivalent groups with covariates; EQ= equipercentile and L= linear, SES: socioeconomic status, SEX: Gender*)

When there are few test-takers with very low results and very few with the highest result, the SEE should be larger at the lower end of the scale and at the upper end of the scale. But Figure 2 shows that this is not the case. At the upper end of the scale, SEE is large, while at the lower end of the scale, SEE is quite small. The fact that a lot of test-takers had low results could be the reason for this. Inspection of the SEE results for equating methods, NEAT PSE EQ, NEAT PSE L, NEAT CE L, NEAT CE EQ, NEC EQ with gender covariate gave similar SEE values. Also, all these methods yielded somewhat similar SEE results between 0-4. The SEE values were highest for NEC, with two covariates between 5-9. Another outstanding detail was that NEC with SES covariate (linear) gave the lowest SEE values throughout the score scale, and NEC with SES covariate (equipercentile) gave the lowest SEE values between 0-5. An RMSD coefficient was calculated to evaluate the random error involved in the equating methods. The resulting coefficients are given in Table 3.

Table 3 reveals that equal RMSD coefficients exist in scores equated with the NEAT PSE EQ, NEC SES EQ, and the NEC-SEX-SES-L methods. The smallest RMSD (0.10, 0.11, 0.12, and 0.13) coefficients were obtained from scores equated with the NEAT PSE, NEAT PSE EQ, NEC SES L, NEC SES EQ, and the NEC-SEX-SES-L method, while the largest RMSD coefficients were obtained through KE CE equating methods. It can be inferred that the maximum random error was provided by chained equating methods, whereas the least random error was yielded by NEAT KE PSE methods and NEC SES methods.

Table 3. RMSD Coefficient According to Equating Methods

Equating methods	RMSD
NEC-SEX-EQ	0.15
NEC-SEX-L	0.14
NEC-SEX-SES-EQ	0.13
NEC-SEX-SES-L	0.12
NEC-SES-EQ	0.12
NEC-SES-L	0.11
NEAT-PSE	0.12
NEAT-PSE L	0.10
NEAT-CE	0.27
NEAT-CE -L	0.28

DISCUSSION and CONCLUSION

In this research, the test forms were equated with the kernel-equating methods under the NEAT and NEC designs, and the equating results were compared according to SEE and RMSD coefficients. For the NEC design, the gender variable and socioeconomic index were used as covariates. After separately adding the covariates to the design, two covariates were added together, resulting in three different NEC designs. Equated scores obtained with kernel linear and kernel equipercentile equating techniques are in the raw scores range (0-9). The greatest gap between the raw scores and equated scores was seen in the NEAT CE methods, while the results of the other techniques were relatively similar to each other. The gap between the raw scores and equated scores was obtained smaller in NEC design, and scores obtained with NEAT PSE and NEC designs were similar to each other, as the PSE technique was used in the NEC design. This finding is consistent with the claims of Wiberg and Branberg (2015).

An inspection of the standard errors of the equating methods reveals that in the 0-4 range, standard errors of the methods are relatively similar and close; nevertheless, towards the middle and tail, the NEC-SES-L, NEC-SES-EQ, and NEC-SEX L equating methods show less standardized error. Conversely, the greatest standardized error is seen when the NEC design is used with two covariates. In their research, Wiberg and Branberg (2015) stated that NEC design shows greater standardized error compared to the NEAT design in the middle scale score range, while NEC, NEATCE, NEAT PSE, and NEATNEC techniques show similar SEE values throughout the score scale.

In this research, SEE values were relatively similar in the 0-4 score scale, while SEE values differed depending on the techniques for the 5-9 range. It is possible to state that the findings of the research are partially inconsistent with the findings of Wiberg and Branberg (2015). Sansivieri and Wiberg (2016) ascertained that using anchor test with covariates lessens the standard error in IRT-based tests equating with equivalent groups and NEAT design. It is possible to claim that this finding is consistent with the research of Sansivieri and Wiberg (2016). In kernel equating, the SEE values in the lower and upper parts of the score scale are generally higher compared to the middle part (von Davier et al., 2004; Wiberg & Branberg, 2015). However, in this research, the standard error was lesser at the lower tail of the score scale. This contrast is possibly caused by the high number of low scores. Branberg and Wiberg (2011) ascertained the fact that using covariates increases accuracy and decreases the standard error of the equation. However, it was revealed that the difference between covariant equating and using anchor items was small. This research revealed that for the NEAT and NEC designs, standard errors are similar in tail scores, while using only two covariates results in an increase in the standard error. Wiberg and Branberg (2015) stated that using more than one covariant causes increases in the sparse data for some cells. So, it is important to limit the number of categories, especially when using continuous variables as covariates. The cause of the increase in the standard error could be traced to the fact that there was sparsity in some cells, or the socioeconomic index being a continuous variable and not subcategorized meaningfully. The decrease in the sample number for the socioeconomic category within the gender category could be another reason for the increase in the standard error. Wallin and Wiberg (2017) suggested using propensity to avoid the problem of decline in the observation number for each category in NEC design.

Another finding of the research is that standard errors of linear equating are lower than those of equipercentile equating. This finding is consistent with research by Choi (2009), Liou, Cheng, and Johnson (1997), Mao (2006), Akın-Arıkan, and Gelbal (2018). The main reason for this is that the large h parameter value reduces the standard error. Additionally, this research revealed that random errors in NEAT CE methods are higher than other methods. The errors in NEAT PSE and NEC designs are also partially similar to each other. Usage of the PSE techniques in NEC design caused the similarity of the random errors. Also, a comparison of the techniques in NEC design indicates that using the socioeconomic level variable as the covariant leads to the lowest error value, whereas using the gender variable as the covariant causes the highest error value. The reason for this is the relationship between the covariant and the test. The correlation value of the gender variable and the test is statistically insignificant, where the SES variable has a significant low-level correlation. Despite higher correlation values between the anchor test and the test, the SES variable was able to define groups like anchor items. When the gender variable is used as the covariant, the error was high; however, adding the SES covariant to the gender variable lowered the error rate. Yurtçu (2018) argues that using two covariates is more effective than using anchor items in studies where researchers used covariates in equating.

The general result of this research is that test forms can be equated using covariates when there are no anchor items. Additionally, anchor tests might not be sufficient if the ability difference among the groups is high, the difficulty difference of the test forms is excessive, and the anchor tests are weak (Albano & Wiberg, 2019). Covariates could be used in such cases. Branberg (2010) states that covariates could be used instead of anchor items. Gonzalez et al. (2015) and Yurtçu (2018) used the Bayesian non-parametric model of covariates and stated that equating is possible even in cases where there are no anchor items. At this point, the ability of the covariant to explain the differences among the groups is critical. For this reason, inspecting the correlations and test scores is vital for the determination of the covariates (Branberg & Wiberg, 2011; Liou et al., 2001; Wiberg, 2015; Wiberg & Branberg, 2015).

Among the large-scale exams in Turkey, only the ABIDE has anchor items. Thanks to this study, it became obvious that the test forms could be used in cases where there are no anchor items, resulting in similar findings. To equate the scores of exams with more than one year of validity period, such as academic personnel exams, postgraduate education entrance exams, and public personnel selection exams, test forms must be equated with an equivalent group/random group design. However, in cases of equivalent groups, test forms could be equated with a covariant, as it is difficult to provide conditions for the groups to be equal. Similar research could be conducted comparing equivalent group design with NEC design and for different subtests of the ABIDE, such as Turkish language tests, science tests, etc. Moreover, Bayesian non-parametric models and kernel equating technique results can be compared.

Acknowledgements

The permission for the data was granted from the Ministry of National Education (Dated 24.12.2018 and No. 57750415-605.01-E.24790203).

REFERENCES

Akın-Arıkan, Ç., & Gelbal, S. (2018). A comparison of traditional and kernel equating methods. *International Journal of Assessment Tools in Education*, *5*(3), 417-427. doi: 10.21449/ijate.409826

Albano, A. D., & Wiberg, M. (2019). Linking with external covariates: examining accuracy by anchor type, test length, ability difference, and sample size. *Applied psychological measurement*, *43*(8), 597-610. doi: 10.1177/0146621618824855

Andersson, B., & Wiberg, M. (2017). Item response theory observed-score kernel equating. *Psychometrica*, *82*(1), 48-66. doi: 10.1007/s11336-016-9528-7

Andersson, B., Branberg, K., & Wiberg, M. (2013). Performing the kernel method of test equating with the package kequate. *Journal of Statistical Software*, *55*(6), 1-25. Retrieved from https://www.jstatsoft.org/article/view/v055i06

Branberg, B. (2010). *Observed score equating with covariates* (Statistical Studies No. 41). Umea: Umea Unıversity, Department of Statistics. Retrieved from https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A306427&dswid=8645

Branberg, K., & Wiberg, M. (2011). Observed score linear equating with covariates. *Journal of Educational Measurement, 48(*4), 419-440. doi: 10.1111/j.1745-3984.2011.00153.x

Chen, H., & Holland, P. (2010). New equating methods and their relationships with Levine observed score linear equating under the kernel equating framework. *Psychometrika*, *75*(3), 542-557. doi: 10.1007/S11336-010-9171-7

Choi, S. I. (2009). *A comparison of kernel equating and traditional equipercentile equating methods and the parametric bootstrap methods for estimating standard errors in equipercentile equating* (Unpublished doctoral thesis). University of Illinois at Urbana-Champaign.

Cook, L. L., Eignor, D. R., & Schmitt, A. P. (1990). *Equating achievement tests using samples matched on ability* (College Board Report No. 90-2). New York: College Entrance Examination Board.

Gonzalez, J., Barrientos, A. F., & Quintana, F. A. (2015). Bayesian non-parametric estimation of test equating functions with covariates. *Computational Statistics and Data Analysis, 89*, 222-244. doi: 10.1016/j.csda.2015.03.012

Holland, P. W., Dorans, N. J., & Petersen, N. S. (2007). Equating test scores. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 169-203). Oxford, UK: Elsevier.

Holland, P. W., & Thayer, D. T. (1989). *The kernel method of equating score distributions* (ETS RR-89-07). Princeton NJ: ETS.

Kolen, M. J. (1990). Does matching in equating work? A discussion. *Applied Measurement in Education, 3*(1), 97-104.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3nd. ed.). New York: Springer.

Liou, M., Cheng, P. E., & Johnson, E. G. (1997). Standard errors of the kernel equating methods under the common-item design. *Applied Psychological Measurement, 21*(4), 349-369.

Liou, M., Cheng, P. E., & Li, M. (2001). Estimating comparable scores using surrogate variables. *Applied Psychological Measurement, 25*(2), 197-207. doi: 10.1177/01466210122032000

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3*(1), 73-95.

Mao, X. (2006). *An investigation of the accuracy of the estimates of standard errors for the Kernel equating functions.* (Unpublished doctoral thesis). University of Iowa, Iowa City.

MoNE, Measurement Assessment and Examination Services General Directorate [Milli Eğitim Bakanlığı Ölçme Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü] (2016). Monitoring and evaluating of academic skills study, 8th students *report (ABIDE) [Akademik becerilerin izlenmesi ve değerlendirilmesi, 8. sınıflar raporu]*. Ankara: Republic of Turkey Ministry of National Education.

R Core Team. (2013). *R: A language and environment for statistical /computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/**.**

Sansivieri, V., & Wiberg, M. (2016). IRT observed-score equating with the nonequivalent groups with covariates design. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W.C. Wang (Eds.), *Quantitative psychology- 81st annual meeting of the psychometric society* (pp. 275-285). Asheville, NC: Springer.

von Davier, A. A., Fournier-Zajac, S., & Holland, P. W. (2007). *An equipercentile version of the Levine linear observed-score equating function using the methods of kernel equating.* (Research Report No: RR-87-31). Princeton, NJ: Educational Testing Service.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York: Springer Verlag.

von Davier, A. A., Holland, P. W., Livingston, S. A., Casabianca, J., Grant, M. C., & Martin, K. (2006). *An evaluation of the Kernel equating method. A special study with pseudotests constructed from real test data* (Research Report No: RR-06-02). Princeton, NJ: Educational Testing Service.

Wallin, G., & Wiberg, M. (2017). Non-equivalent groups with covariates design using propensity scores for kernel equating. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W. C. Wang (Eds.), *Quantitative psychology – 81st annual meeting of the psychometric society* (pp. 309-319). Asheville, NC: Springer.

Wiberg, M. (2015). Anote on equating test scores with covariates. In E. Frackle-Fornius (Ed.), *Festschrift in honor of Hans Nyquist on the occasion of his 65th birthday* (pp. 96-99). Stockholm, Sweden: Department of Statistics, Stockholm University.

Wiberg, M., & von Davier, A. A. (2017). Examining the impact of covariates on anchor tests to ascertain quality over time in a college admissions test, *International Journal of Testing, 17*(2), 105-126. doi: 10.1080/15305058.2016.1277357

Wiberg. M., & Branberg, K. (2015). Kernel equating under the non-equivalent groups with covariates design. *Applied Psychological Measurement, 39*(5), 349-361. doi: 10.1177/0146621614567939

Yurtçu, M. (2018). *Parametrik olmayan bayes yöntemiyle ortak değişkenlere göre yapilan test eşitlemelerinin karşilaştirilmasi* (Yayınlanmamış doktora tezi). Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Eşdeğer Olmayan Gruplarda Ortak Değişkenlerin Kernel Eşitlemeye Etkisi

Giriş

Yaklaşık yüz yıldır, test eşitlemede yeni yöntemler ve desenler geliştirilmiştir. Bu yöntemlerden biri olan Kernel eşitleme yöntemi, ilk olarak Holland ve Thayer (1989) tarafından tanımlanmış ve daha sonra von Davier, Holland ve Thayer (2004) tarafından geliştirilmiştir. Kernel eşitleme, kesikli puan dağılımlarının sürekli dağılımlara dönüştürerek puan dağılımlarının eşitlendiği bir eşit yüzdelikli gözlenen puan eşitleme yöntemidir (von Davier ve diğerleri, 2006). Kernel eşitleme beş basamaktan oluşur: ön düzgünleştirme, puan dağılımlarının kestirilmesi, süreklileştirme, eşitleme ve eşitlemenin standart hatasının hesaplanmasıdır (von Davier ve diğerleri, 2004). Kernel eşitleme doğrusal ve eşit yüzdelikli eşitleme fonksiyonlarını içerir (von Davier ve diğerleri, 2006).

Test eşitlemede; tek grup deseni, eşdeğer grup deseni, dengelenmiş grup deseni ve denk olmayan gruplarda ortak madde deseni (NEAT) gibi birçok farklı grup deseni bulunmaktadır (Kolen ve Brennan, 2014; von Davier ve diğerleri, 2004). Alanyazında en sık kullanılan desenlerden biri NEAT desendir. Kernel eşitlemede NEAT deseninde; son tabakalama (PSE), Levine gözlenen puan doğrusal, zincirleme eşitleme (CE) yöntemleri kullanılmaktadır (von Davier ve diğerleri, 2004). NEAT deseninde iki farklı grup vardır ve bu gruplar, iki farklı test formu X ve Y ve ortak test olan A testini alır. PSE, grup I'deki X dağılımını ve grup II'deki Y dağılımını tahmin etmek için ortak test olan A'yı kullanır. A verilen X'in koşullu dağılımının ve A verilen Y'nin koşullu dağılımının popülasyonun değişmez olduğunu varsayar. CE'de ise ortak test zincirin bir parçası olarak kullanılır: ilk önce grup I üzerinden X testini A'ya, sonra da grup II üzerinden A ortak testini Y testine bağlar.

NEAT deseninde, gruplar arasındaki yetenek farkını ayarlamak için ortak maddeler kullanılmaktadır. Ancak bütün test programları veya standartlaştırılmış testlerde ortak maddeler test formlarında yer almayabilir. Ayrıca farklı test formlarını alan grupların eşdeğer olması da uygulamada çok zor olduğundan, eşdeğer grup desenine göre test formları eşitlenmeyebilir. Bu durumda, eğer denk olmayan gruplarda ortak madde yer almıyorsa anlamlı ortak değişkenler ortak maddeler yerine kullanılabilir (Wiberg & Branberg, 2015). Eğer ortak maddeler yerine ortak değişkenler kullanılıyorsa bu desen denk olmayan gruplarda ortak değişken deseni (NEC) adını alır.

Eşitlemede ortak değişkenler ile ilgili alan yazın incelendiğinde, sınırlı sayıda çalışma yapıldığı görülmüştür. Yurtdışında yapılan ilk çalışmalarda test eşitlemede farklı değişkenler kullanılarak ilerde ortak değişkenlerle yapılacak araştırmalara ışık tutulmuştur (Cook, Eignor & Schmitt, 1990; Holland, Dorans & Petersen, 2007; Kolen, 1990; Livingston, Dorans, & Wright, 1990). Son yıllarda yapılan çalışmalara baktığımızda, Branberg (2010) çalışmasında ortak değişkenleri kullanarak test formlarını eşitlemiştir ve ortak değişkenlerin ortak maddeler yerine kullanılabileceğini ifade etmiştir.

Bu bağlamda bu çalışmanın amacı, NEAT ve NEC desenlerinde Kernel eşitleme yöntemlerinden PSE ve CE doğrusal ve eşit yüzdelikli eşitleme yöntemleri ile elde edilen sonuçların karşılaştırılması ve cinsiyet ve sosyoekonomik düzey ortak değişkenlerin test eşitlemeye etkisini belirlemektir.

Yöntem

Bu çalışma kapsamında, 2016 ABİDE projesi kapsamında uygulanan matematik alt testlerinden elde edilen puanlar NEAT ve NEC desenlerine göre Kernel zincirleme eşit yüzdelikli, Kernel son tabakalama eşit yüzdelikli, Kernel zincirleme doğrusal ve Kernel son tabakalama doğrusal eşitleme yöntemleri kullanılarak eşitlenmiştir. Bu araştırmada, var olan yöntem ve tekniklerin gerçek veri üzerinden sınanması yapıldığından araştırma betimsel araştırmadır.

Bu çalışma kapsamında, Millî Eğitim Bakanlığı'ndan var olan öğrenci verilerinden rastgele olarak seçilen A formunu alan 3000 ve B formunu alan 3000 öğrenci olmak üzere toplamda 6000 öğrenciye ait veriler kullanılmıştır. A formunu alanların 1292'si (%43.07) kız ve 1708'i (%56.93) erkek; B formunu alanların 1518'i (%50.6) kız ve 1482'si (%49.4) erkek öğrencilerden oluşmaktadır.

NEC desende ortak değişkenleri belirleyebilmek amacıyla ABİDE projesi kapsamında hazırlanan raporda matematik başarısı ile öğrenci anketinde yer alan değişkenler arasındaki korelasyon değerleri incelenmiş ve bu değişkenlerden korelasyon değeri en yüksek olan sosyoekonomik indisi (r = .37, p < .05) ortak değişken olarak seçilmiştir (MEB, 2016). Sürekli bir değişken olan sosyoekonomik düzey indisini kategorilere ayırmak için iki aşamalı kümeleme analizi kullanılarak düşük, orta ve yüksek olmak üzere üç düzey oluşturulmuştur. Çalışmadaki bir diğer ortak değişken ise cinsiyet (sex) değişkeni olarak belirlenmiştir (Branberg & Wiberg, 2011; Gonzalez, Barrientos, & Quintana, 2015; Liou, Cheng, & Li, 2001, Yurtçu, 2018). Sosyoekonomik statü (ses); *düşük* = 1, *orta* = 2 ve *yüksek* = 3 ve cinsiyet değişkeni ise *kadın* = 1 ve *erkek* = 2 olarak kodlanmıştır. Testlerin eşitlenmesi için R programında (R Core Team, 2013) yer alan "kequate" paketi (Andersson, Branberg, & Wiberg, 2013) kullanılmıştır. Değerlendirme kriterleri olarak, eşitleme yöntemleri için eşitlemenin standart hatası (SEE) ve eşitleme hatası (RMSD) kullanılmıştır.

Sonuç ve Tartışma

Bu araştırmada, Kernel eşitleme yöntemleriyle NEAT ve NEC desenlerinde test formları eşitlenerek SEE ve RMSD katsayılarına göre eşitleme yöntemleri karşılaştırılmıştır. NEC desende ortak değişkenler olarak cinsiyet değişkeni ve sosyoekonomik indisi kullanılmıştır. Ortak değişkenler ayrı ayrı desene eklendikten sonra, iki ortak değişken birlikte eklenerek üç farklı NEC deseni oluşturulmuştur. NEAT ve NEC desenlerinde on ayrı eşitleme puanı elde edilerek, sonuçlar karşılaştırılmıştır.

Kernel doğrusal ve Kernel eşit yüzdelikli eşitleme yöntemleriyle elde edilen eşitlenmiş puanların, ham puan ranjında olduğu (0-9 aralığında) görülmüştür. Ham puan ile eşitlenmiş puanlar arasındaki en büyük farklılığın ise NEAT CE yöntemlerinde olduğu, diğer yöntemlerin kısmen birbirine daha yakın olduğu görülmüştür. Ayrıca NEC desende ham puan ile eşitlenmiş puanlar arasındaki farkın az olduğu ve NEC desende PSE yöntemi kullanıldığından NEAT PSE ile NEC desenlerinden elde edilen eşitlenmiş puanların benzer olduğu sonucu elde edilmiştir. Elde edilen bu bulgu Wiberg ve Branberg (2015) tarafından ulaşılan bulgularla tutarlıdır.

Eşitleme yöntemlerine ilişkin eşitlemenin standart hataları incelendiğinde; 0-4 puan aralığında yöntemlerin kısmen birbirine yakın veya benzer standart hatalara sahip olduğu; ancak orta puan ve uç puanlara doğru gidildikçe NEC-SES-L, NEC-SES-EQ ve NEC-SEX L eşitleme yöntemlerinin daha düşük standart hata verdiği görülmektedir. Ancak NEC desenin iki ortak değişken ile birlikte olduğu eşitleme deseninde en yüksek standart hata elde edilmiştir. Wiberg ve Branberg (2015) çalışmasında orta ölçek puan aralığında NEC desenin NEAT desenden daha büyük standart hataya sahip olduğu, ancak bütün puan ölçeği boyunca NEC, NEATCE, NEAT PSE ve NEATNEC yöntemlerinin benzer SEE değerlerine sahip olduğu sonucuna ulaşılmıştır. Bu çalışmada ise 0-4 puan ölçeğinde eşitleme yöntemlerinden elde edilen SEE değerleri kısmen benzer iken, 5-9 puan aralığında SEE değerlerinin yöntemlere göre farklılaştığı sonucuna ulaşılmıştır. Elde edilen bu bulgunun Wiberg ve Branberg (2015) tarafından ulaşılan bulgularla kısmen tutarlı olmadığı söylenebilir. Elde edilen bu bulgu Sansivieri ve Wiberg'in (2016) MTK'ya dayalı test eşitleme yöntemlerinde eşdeğer grup ve NEAT desenlerinde ortak test ile birlikte ortak değişkenler kullanıldığında standart hatanın azaldığı bulgusuyla tutarlı olduğu söylenebilir.

Ayrıca NEC deseninden elde edilen yöntemleri karşılaştırırsak, sosyoekonomik düzey değişkeninin ortak değişken olarak kullanıldığı desende en düşük hata değerinin olduğu, en yüksek ise cinsiyet değişkeninin ortak değişken olarak kullanıldığı desende olduğu bulunmuştur. Bunun nedeni ise ortak değişkenler ile test arasındaki ilişkidir. Cinsiyet değişkeni ile test arasındaki korelasyon değeri istatistiksel olarak anlamsız iken, SES değişkeni ile düşük düzeyde anlamlı bir ilişki bulunmaktadır. Ortak test ile test arasındaki korelasyon değeri daha yüksek olmasına rağmen, ses değişkeni ortak maddeler gibi gruplar arasındaki farkları açıklayabilmiştir. Cinsiyet değişkeninin ortak değişken alındığı durumda hata yüksek iken, cinsiyet değişkenine ek olarak SES ortak değişkeninin eklenmesi de hatayı azalmıştır. Yurtçu'nun (2018) Bayes modelde ortak değişkenler ile eşitleme yaptığı çalışmada, iki ortak değişken kullanılmasının ortak maddelerden daha etkili olduğu sonucunu elde etmiştir.

Türkiye'de yapılan geniş ölçekli sınavlar göz önüne alındığında sadece ABİDE projesinde ortak maddeler yer almaktadır. Bu proje kapsamında ortak maddeler olmadığında ortak değişkenler de kullanılarak test formlarının eşitlenebileceği ve elde edilen sonuçların birbirine yakın olduğu elde edilmiştir. Bir yılda fazla geçerliği olan geniş ölçekli sınavların (KPSS, ALES gibi) test puanları eşitlenmek istendiğinde ise bu test formlarında ortak maddeler olmadığından eşdeğer grup/random grup desenine göre test formları eşitlenmelidir. Ancak bu durumda da grupların eşdeğer olma şartının sağlanması çok zor olduğundan, test ile ortak değişkenler arasındaki ilişki göz önüne alınarak, test formları eşitlenebilir. Benzer bir çalışma ABİDE projesinde yer alan Türkçe, Fen bilgisi gibi farklı alt testler için ve eşdeğer grup deseni ile NEC desen karşılaştırılarak yapılabilir. İleride yapılacak bir araştırmada NEC desende Parametrik olmayan Bayes modelleri ile Kernel eşitleme yöntemlerinden elde edilen sonuçlar karşılaştırılabilir.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(4); 374-404

A Meta-Analytic Reliability Generalization Study of the Oxford Happiness Scale in Turkish Sample

Vildan ÖZDEMİR * Yıldız YILDIRIM ** Şeref TAN ***

Abstract
The purpose of this study was to analyze the meta-analytical reliability generalization of short form and long form of the Oxford Happiness Scale (OHS) for Turkish sample. In addition, how different moderator variables affect reliability coefficients was examined. A number of criteria have been set to determine the studies to be included in meta-analysis. Of 95 Cronbach's Alpha coefficients obtained from 92 studies that were selected according to criteria were included in the meta-analysis. In the data analysis, reliability generalization based on meta-analysis was used. The effect of moderator variables on variability in reliability estimations as effect size was examined by Analog ANOVA. As a result of the research, it was found that the mean alpha was .81 for overall studies; .76 for the short form and .87 for the long form of OHS. In addition, it was concluded that number of items had a statistically significant effect on the reliability estimation in terms of heterogeneity of true effect sizes, and sample type had a statistically significant effect on the reliability estimation for OHS (long-form). But sample type had no effect on the reliability estimation for OHS-S (short-form), and field of study had no effect for both short and long form reliability estimates.

Key Words: Reliability generalization, meta-analysis, Oxford happiness scale.

INTRODUCTION

The place and importance of measurement and assessment in education and psychology are indisputable. Accordingly, education and psychology are unthinkable without the field of measurement and evaluation. Two conceptions underlie the field of measurement and evaluation: these are *reliability* and *validity*. The aim of the classical test theory is to present a model to estimate the accuracy of test score measures. And the accuracy is related to the reliability of the test (McDonald, 1999). In short, reliability is the degree of being free from random error of measures. In addition, reliability means consistency of the scores received by the same individuals participating in the same or equivalent tests (Anastasi, 1982). A number of calculations are required to interpret reliability. At this point, the concepts of reliability index and reliability coefficient appear. While the reliability index refers to the relationship between observed scores and true scores, the reliability coefficient refers to the relationship between the scores from the parallel forms. Based on the mathematical relationship between the two concepts, it can be said that the reliability coefficient is the ratio between the true score variance and the observed score variance (Crocker & Algina, 2008). There are formulas suggested by researchers in the calculation of the reliability coefficient. Some coefficients require a single test administration, while others require more than one test administration. One of the most useful characteristics of internal consistency calculations is that it is based on only single test administration (Kline, 2005). Some of the formulas that used in calculating the reliability coefficient in the context of internal consistency are as follows: KR-20, KR-21 (Kuder & Richardson, 1937), Guttman Lambda ($\lambda 3$) / or Cronbach's Alpha (α) (Cronbach, 1951; Guttman, 1945), Kristof's coefficient (Kristof, 1963), Stratified alpha (Cronbach, Schönemann, & McKie, 1965), Heise and

* Res Assist, Aksaray University, Faculty of Education, Aksaray-Turkey, vildanbagci@gmail.com, ORCID ID: 0000-0002-9051-8860
** Res Assist, Aydin Adnan Menderes University, Faculty of Education, Aydin-Turkey, yildizyldrm@gmail.com, ORCID ID: 0000-0001-8434-5062
*** Prof. PhD., Gazi University, Gazi Education Faculty, Ankara-Turkey, sereftan4@yahoo.com, ORCID ID: 0000-0002-9892-3369

To cite this article:
Özdemir, V., Yıldırım, Y., & Tan, Ş. (2020). A meta-analytic reliability generalization study of the Oxford Happiness Scale in Turkish sample. *Journal of Measurement and Evaluation in Education and Psychology, 11*(4), 374-404. doi: 10.21031/epod.766266.

Received: 08.07.2020
Accepted: 29.11.2020

Bohrnstedt's Omega (Ω) (Heise & Bohrnstedt, 1970), Armor's theta (θ) (Armor, 1973), Raju's Beta (β) (Raju, 1977), Revelle's Beta (β) (Revelle, 1979), Feldt-Gilmer coefficient (Gilmer & Feldt, 1983), McDonald's Omega (ω) (McDonald, 1985), and Angoff-Feldt coefficient (Feldt & Brennan, 1989).

The reliability and validity of the scores obtained from the measurement tools must be investigated absolutely (Crocker & Algina, 2008). Among the previous coefficients, the Cronbach's Alpha is the most frequently used coefficient in the literature for analyzing and interpreting internal consistency reliability. However, as with all coefficients, Cronbach's Alpha differs from research to research even though the same scale is used because it is a sample dependent coefficient. For example, Cronbach's alpha was .29 in one of the studies in which the Oxford Happiness Scale-Short Form was used and in which the sample was chosen from university students (Taşdibi-Ünlü, 2019), while it was found .97 in another study (İlhan & Güler, 2017). The reliability coefficients vary depending on the variation of the sample characteristics: sample size, administration conditions, time of administration, etc. Such differences in the studies required the generalization of reliability. Reliability generalization (RG) based on meta-analysis was first made by Vacha-Haase (1998). According to Vacha-Haase, the RG analyzes the amount and sources of the variability of the reliability coefficients in different measurements and studies. In other words, the RG study examines whether the reliability coefficient differs between studies. When the literature is reviewed, there are a lot of RG studies based on meta-analysis (e.g. Barnes, Harp, & Jung, 2002; Beretvas, Meyers, & Leite, 2002; Bornmann, Mutz, & Daniel, 2010; Li & Bagger, 2007; Nilsson, Schmidt, & Meek, 2002; Shields & Caruso, 2003; Shields & Caruso, 2004; Vacha-Haase & Thompson, 2011; Vicent, Rubio-Aparicio, Sánchez-Meca, & Gonzálvez, 2019). On the other hand, when the Turkish literature is examined, no RG studies related to a specific scale were found. Some of these studies which have examined different study characteristics that affect the mean reliability estimation in literature can be summarized as follows:

Aguayo, Vargas, Emilia, and Lozano (2011), Capraro and Capraro (2002), and Graham, Liu, & Jeziorski (2006) examined the effect of sample characteristics on reliability estimates, and their results showed that reliability coefficients were dependent on sample characteristics. Also, Caruso (2000) aimed an RG analysis of NEO Personality Scales and examined the effect of sample characteristics. He founded that there was a significant difference between the reliability coefficients for sample type for agreeableness subscale scores. Similar to this result, Caruso, Witkiewitz, Belcourt-Dittloff, and Gottlieb (2001), Shields and Caruso (2004), and Yin and Fan (2000) found that sample type was a statistically significant predictor for reliability. In contrast to these studies, Hess, McNab, and Basoglu (2014), Thompson and Cook (2002), and Wallace and Wheeler (2002) found that the mean reliability estimates were invariant across different sample types. Additionaly, Wallace and Wheeler (2002) examined whether language was related to reliability estimates. In their results there were no statistically significant differences for different languages. However, the results indicated that coefficient alpha estimates were affected by the language in Wheeler, Vassar, Worley, and Barnes's (2011) study.

Graham, Diebels, and Barnow (2011), Henson, Kogan, and Vacha-Haase (2001), Hess et al. (2014), and Nilsson et al. (2002) examined how the length of the test differentiates reliability coefficients in meta-analytic RG, and they concluded that as the length of the test increase, the reliability estimates increase. In addition to these results, Caruso (2000) concluded that the most important factor was the scale length for reliability coefficient. However, in another study, it was found that the mean reliability was predicted higher as the number of items in the scale decreased (Hanson, Curry, & Bandalos, 2002).

Sample size, which can be another possible source of variability for reliability estimation, was examined by some researchers (Hanson et al., 2002; Henson et al. 2001; Viswesvaran & Ones, 2000). Hanson et al. (2002) explored a correlation between reliability estimation and sample size or gender homogeneity. They reported that reliability estimates were related to the sample size of the client or therapist. The correlations between sample size and reliability estimates fluctuated in both direction and size for all subscales in Henson et al's. (2001) study. In contrast to these results, Viswesvaran and Ones (2000) found that there were no correlations between sample size and reliability estimations.

The other moderator variables, which are age of research participants, type of research design, testing conditions, gender, sexual orientation, ethnicity and marital status, and standard deviation of the subscale scores were also examined by researchers (Barnes et al., 2002; Capraro & Capraro, 2002; Graham et al., 2006; Graham et al., 2011; Vicent et al., 2019, and Wheeler et al., 2011). While the reliability estimates did not differ by the gender, sexual orientation, ethnicity and marital status in Graham et al.'s (2006) study, testing conditions, type of research design, aim of the study (psychometric or applied), and the standard deviation of the subscale scores were found as the sources of variability in reliability coefficients (Barnes et al. 2002; Capraro & Capraro, 2002; Vicent et al., 2019, and Wheeler et al., 2011). Also, the results of these studies showed that reliability estimates were sensitive to the age of the sample (Barnes et al., 2002; Graham et al., 2011; Vicent et al., 2019; Yin & Fan, 2000).

When the studies in the literature were examined, there were studies which analyzed the reliability generalization, and also whether the reliability coefficient differs according to variables such as test length (or the number of items), sample size, sample type, gender, reliability coefficient type, study language, race, marital status, age, etc. In the results of some of these studies, it was seen that variables such as number of items and sample type were found as sources of variability in reliability (e. g. Caruso, 2000; Caruso et al., 2001; Hanson et al., 2002; Henson et al., 2001; Hess et al., 2014; Nilsson et al., 2002; Shields & Caruso, 2004; Yin & Fan, 2000). In contrast, some studies concluded that these variables did not affect the reliability coefficient (e.g., Graham et al., 2011; Hess et al., 2014; Thompson & Cook, 2002; Wallace &Wheeler, 2002). As seen in previous studies, the reliability of the measures obtained with different scales can affected by different variables. In this study, by examining these studies, a meta-analytic RG analysis was carried out for the Turkish sample. And similar to the literature, it was investigated how general the reliability coefficients are in different number of items, sample types, and fields of study and whether the reliability coefficients were affected by these variables. Within the scope of the study, it was aimed to analyze the reliability generalization of the long and short forms of the Oxford Happiness Scale (OHS) (Argyle, Martin, & Lu, 1995; Hills & Argyle, 2002). The reason for choosing this scale in the study was that the studies in the field of positive psychology have increased in recent years, and happiness is one of the concepts that are frequently researched in the field of positive psychology (Compton & Hoffman, 2019). Also, considering that the feeling of happiness has an effect on many aspects of individuals' lives, it is extremely important to measure the structure of happiness reliably. When both Turkish and non-Turkish literatures were examined, it was seen that the OHS is frequently used to measure happiness (e.g. Demir, 2020; Francis & Crea, 2018; Francis, Ok, & Robbins, 2017; Lin, Imani, Griffiths & Pakpour, 2020; Okur & Totan, 2019; Yıldırım & Sezer, 2020). Considering that the reliability values of the studies using the OHS in the Turkish literature have been in a wide range (.29 - .97), these differences should be investigated, and the reliability should be generalized for the Turkish sample. It is thought that RG studies can contribute as important sources of information for test administrators and researchers by Vacha-Haase, Henson, and Caruso (2002). In line with all this, it is important to bring this study into Turkish literature and the field of education and psychology.

METHOD

This meta-analysis study was performed according to the PRISMA (Liberati et al., 2009) guidelines. According to that, two authors searched the databases independently, identified the studies by screening the titles and the abstracts, removed the duplicates, and assessed the full-text articles for including in meta-analysis. This section includes data collection tools, sample, coding of study characteristics, and data analysis.

Data Collection Tool

The studies that were searched at the databases of Google Scholar, YOK (Higher Education Institution in Turkey) national thesis/dissertation center, EBSCOhost via Gazi University Central Library, and finally Aydin Adnan Menderes University Library's databases (e.g. BMJ, Dergipark, DOAJ, Clinical

Key, SAGE, Science Direct, Springer Link, Taylor & Francis etc.) and published between 2011 and 2020, and which used the long and short forms of the OHS were included in this review.

Oxford happiness scale

In the psychology, educational and social sciences, different measurement tools have been developed in order to measure happiness according to the increase in the studies on the concept of happiness. One of the most commonly used measurement tools in measuring happiness is the Oxford Happiness Inventory (OHI). This scale was developed by Argyle, Martin, and Crossland (1989) and Argyle et al. (1995).

OHI has been developed similarly to the format of the Beck Depression Inventory. The inventory has consisted of 29 personal well-being items by reversing 20 items from Beck Depression inventory and adding nine items that reflect different aspects of happiness. The cross-cultural comparison of OHI has been made by applying to students in Australia, Canada, and America (Francis, Brown, Lester, & Philipchalk, 1998). At the same time, it has been adapted for many different cultures such as Israel and China (Francis & Katz, 2000; Lu & Shih, 1997). However, since this inventory was developed by applying it to clinical patients, it was observed that individual responses were directed towards one of the two main items when administered to non-patients. The means for a substantial portion of items could be below the corresponding standard deviations. This showed that the responses could be distributed uniformly, and the items might not be able to fully contribute to the measurement of happiness. To overcome these situations, Hills and Argyle (2002) revised the inventory and constituted the OHS.

OHS consists of 29 items which are 6-point Likert-scale, and these points are within the range of *strongly agree-strongly disagree*. Half of the scale items are reversed. Thus, it is thought to decrease the possibility of individuals to respond harmoniously or biased. In addition, in the same study, an 8-item short form of OHS was developed for situations when setting was limited (Hills & Argyle, 2002).

The adaptation study of OHS to Turkish was conducted by Doğan and Sapmaz (2012). They examined the psychometric properties of the scale by implementing 491 university students. While the validity of OHS was investigated by criterion-related validity methods and exploratory and confirmatory factor analyses (EFA, CFA), the reliability was investigated by internal consistency, split-half, and composite reliability methods. Accordingly, the Cronbach’s Alpha coefficient and composite reliability coefficient were found .91, and the reliability coefficient obtained by the split-half method was found .86.

When the validity studies were examined, as a result of the EFA, a single-factor structure was obtained, as it was in its original form. It was concluded that the single-factor structure of the scale was preserved with CFA. The findings revealed that the Turkish form of OHS showed similar psychometric features to its original form.

The short form of OHS, which consists of eight items, was adapted to Turkish by Doğan and Çötok (2011). They applied the scale to 532 university students and evaluated the psychometric properties via EFA and CFA, internal consistency, and test-retest methods. In the item analysis, item 4 was excluded from the scale because the item-total correlation value was less than .30. The reliability and validity analyses after this stage were made with the remaining seven items. Cronbach’s Alpha coefficient calculated from the data obtained from 321 students were found as .74. In the test-retest reliability study, OHS-S was applied to 81 students at two-week intervals, and the correlation was found .85 between the two administrations.

The EFA was showed that the scale has a single-factor structure as its original form does. It was concluded that the single-factor structure of the scale was confirmed by CFA. As a result, it was determined that OHS-7 was a valid and reliable measurement tool to measure the happiness of Turkish students. In this review, studies administering the long or short form of the OHS, which was adapted to Turkish and was analyzed in terms of validity and reliability, were searched.

Search Strategy

The process for selecting studies was given in Figure 1 (Moher, Liberati, Tetzlaff, Altman, & The PRISMA Group, 2009). When Figure 1 was examined, it was seen that the first stage was searching. The search of articles and thesis was carried out in the following databases: Google Academic, YOK national thesis/dissertation center, all databases in Gazi University Central Library, and Aydin Adnan Menderes University Library (Databases were presented in Appendix A). The used keywords were "oxford happiness" and "oxford mutluluk". The search was carried out spanning the years 2011 to 2020 because the short and long forms of the OHS were adapted to Turkish in 2011 and 2012 respectively. When the specified databases were searched with keywords, it was seen that there were 6906 studies in total. First, the studies without Turkish sample groups were eliminated, and double coding was avoided. In addition, for studies involving more than one reliability coefficient, each coefficient was coded separately. Therefore, a total of 206 studies were coded from 6906 studies. Later, these 206 studies were examined according to inclusion criteria [i) They must be published in specified databases, ii) Cronbach's Alpha reliability coefficients must be reported or can be calculable iii) They must include a sample group, sample size and scale form/number of item of study, iv) the sample group of study must consist of Turkish people and v) the language of the studies must be English or Turkish]. A summary of the phases of the meta-analysis was shown in Figure 1.

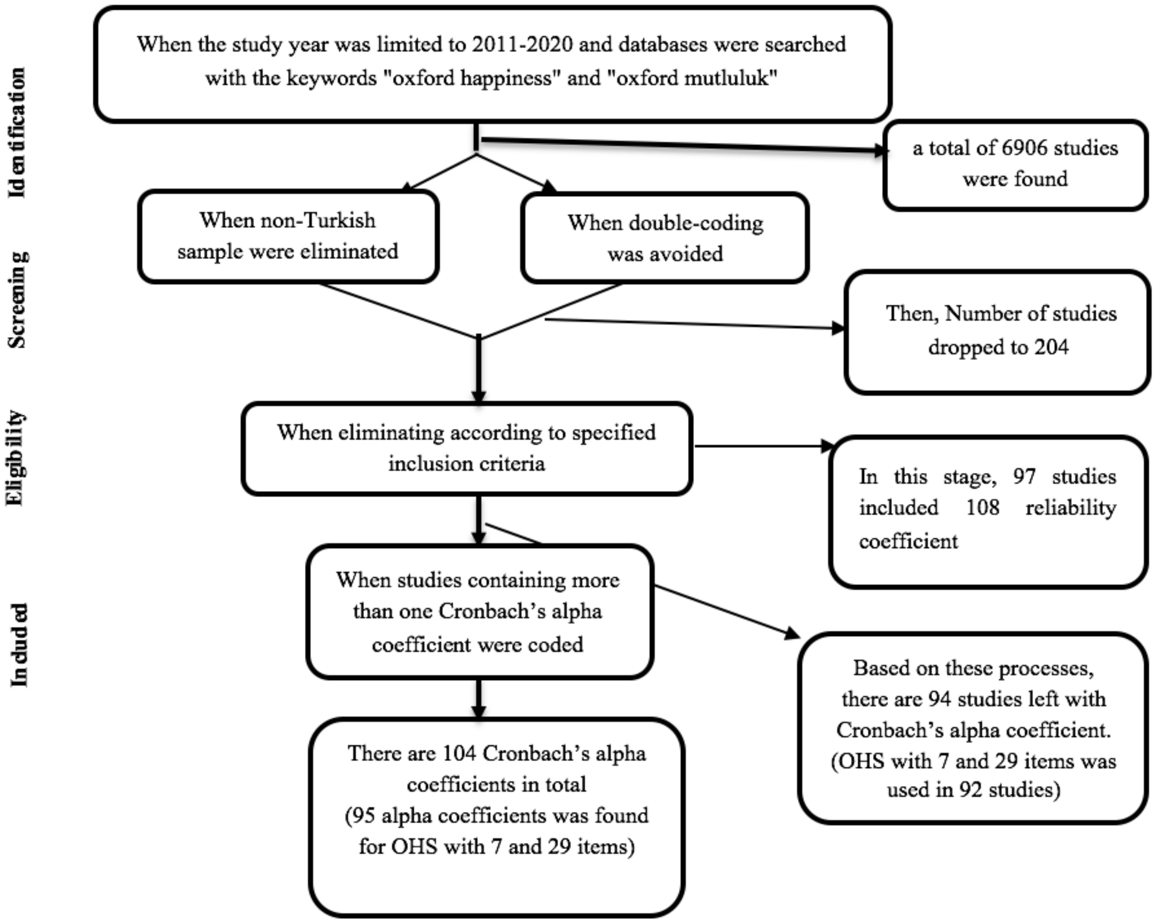

Identification
Screening
Eligibility
Included
When the study year was limited to 2011-2020 and databases were searched with the keywords "oxford happiness" and "oxford mutluluk"
a total of 6906 studies were found
When non-Turkish sample were eliminated
When double-coding was avoided
Then, Number of studies dropped to 204
When eliminating according to specified inclusion criteria
In this stage, 97 studies included 108 reliability coefficient
When studies containing more than one Cronbach's alpha coefficient were coded
Based on these processes, there are 94 studies left with Cronbach's alpha coefficient. (OHS with 7 and 29 items was used in 92 studies)
There are 104 Cronbach's alpha coefficients in total (95 alpha coefficients was found for OHS with 7 and 29 items)

Figure 1. PRISMA Flowchart

After these phases of meta-analysis, the study group consisted of 92 studies of which 27 were thesis, and 65 were articles in accordance with the criteria determined by the researchers in this study. And 95 Cronbach's Alpha coefficients were obtained from 92 studies that were presented in Appendix B. The selected studies were read and classified by two authors.

When Table 1, which has shown descriptive features of the studies included in the study, was examined, it was seen that seven studies were published between 2011-2015, and 85 studies were published between 2016-2020. Twenty studies were in English, and 72 studies were in Turkish. The short form was used in 56 of the studies and the long form of the scale in 36 of them. In addition, the sample type was coded as student for 50 studies and non-student for 42 studies. And, there were four studies with sample sizes ≤ 100, 15 studies with sample sizes between 100 and 200, and 73 studies with sample sizes > 200. Finally, the field of study was examined; it was observed that most studies (63) were in the field of social sciences. Also, it was seen that least studies (11) were in the field of sport sciences. Lastly, there were 18 articles or thesis for psychology/health sciences.

Table 1. Frequencies of the Studies According to Study Characteristics

Descriptive Variables	Categories	Number of Studies	Number of Cronbach's α
Number of items	7	56	58
	29	36	37
Type of Sample	Student	50	51
	Non-Student	42	44
Sample Size	≤100	4	6
	>100 and ≤200	15	16
	>200	73	73
Year of Study	2011-2015	7	7
	2016-2020	85	88
Language of Publication	Turkish	72	73
	English	20	22
Type of Publication	Article	65	68
	Thesis	27	27
Field of Study	Social Sciences (SS)	63	63
	Psychology (P)/Health Sciences (HS)	18	18
	Sport Sciences (SPS)	11	13
Total		92	95

Coding of Study Characteristics

After selecting the studies according to the inclusion criteria to the meta-analysis, the following sample and study characteristics were recorded by the researchers: (i)name of the article or thesis, (ii)name of the author(s) who conducted the study, (iii)year of the article or thesis, (iv)publication language of the study, (v)type of the study (article/thesis), (vi)type of the scale (the short form/the original form), (vii) reliability coefficient, (viii)type of reliability, (ix)sample size/the number of participants in the sample, (x)the number of items on the scale, (xi)fields of study and (xii)participant characteristics.

A total of 108 reliability coefficients were obtained from 97 studies. Of the 108, 104 were coefficient alpha; four coefficients were test-retest reliability, split-half reliability, and composite reliability estimates. However, the present study didn't characterize the scores by reliability type because of the small number of the reliability estimates differing from the coefficient alpha. Also, in some studies, it was observed that the item was removed or not used completely, and studies indicating a different number of items from the 7 and 29 items in the original scale forms were excluded from the study. Therefore, 92 studies remained when the studies that did not use all of the items were eliminated, and 95 alpha coefficients were obtained from these studies. Finally, 95 coefficient alpha values were analyzed for reliability generalization.

The inter-coder reliability was also examined for the data coded by the two authors according to the determined variables and criteria. The inter-coders reliability was calculated by the percent of agreement and Krippendorff's Alpha coefficient. For this, two coders coded for the same 10 studies and 11 reliability coefficients. These statistics were analyzed by SPSS 23 and SPSS macro that was developed by Hayes and Krippendorff (2007) and used for Krippendorff's Alpha coefficient. As a result of the analyses, the percent of agreement was .95, and the Krippendorff's alpha coefficient was

.94. These values were an indication that the inter-coder reliability is high. Krippendorff (2004) suggested that Krippendorff's Alpha coefficient should be at least .80, and he stated that alpha ≥ .667 is acceptable. Accordingly, inter-coder reliability is considered appropriate. Also, conflicts between the coders were examined by authors, and it has been determined that it was caused by the use of the keyboard. These conflicts that were detected were resolved.

Data Analysis

Reliability generalization studies provide reliability predictions to make a comparison between studies. In addition, it also examines the potential causes of variability in score reliability across studies (Graham et al., 2006). In this RG study, the generalizability of Cronbach's Alpha coefficients was investigated. Cronbach's Alpha is the square of the correlation because the reliability coefficients are variance-accounted statistics (Thompson & Vacha-Haase, 2000). Since the distribution of correlations isn't normal and has problematic standard errors, they must be transformed. Therefore, the raw alpha coefficients were transformed by *Fisher z-transformation*. Although Fisher's z-transformation was suggested for reliability coefficients calculated as Pearson correlation (e.g., test-retest, parallel forms) (Sánchez-Meca, López-López & López-Pina, 2013), recent studies have shown that Fisher z performed well and was very similar to other transformations in terms of empirical coverage probability (Romano, Kromrey, & Hibbard, 2010).

The random effects model (REM) which assumes that between-studies variance has been estimated greater than zero was used because of considering that the studies included in the research were obtained from different samples, fields, and years. Also, REM has been more realistic for real world applications (Field, 2003). In RG studies, there are a few heterogeneity estimators that are used for REM. Some of these estimators are Hunter-Schmidt, Hedges, DerSimonian and Laird, and the estimator based on maximum likelihood estimation (Maximum Likelihood-ML, Restricted ML-REML). In this study, the between-study variance, τ^2, was estimated by DerSimonian and Laird.

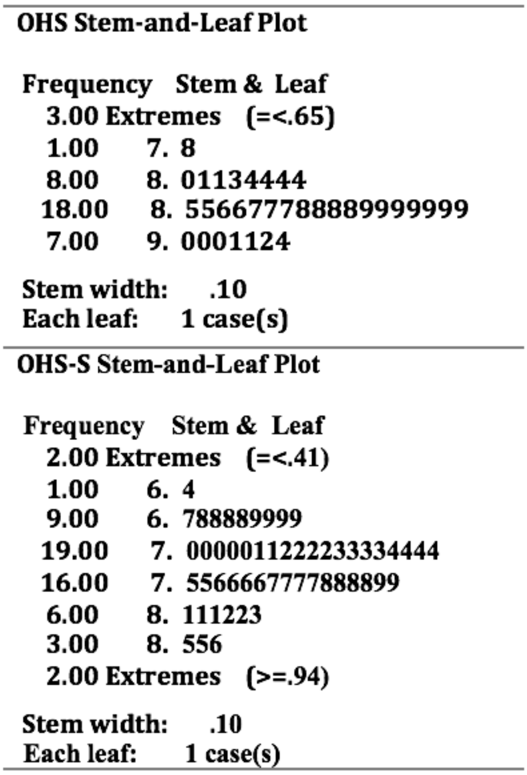
The heterogeneity of Cronbach's Alphas was assessed by calculating the I^2 index as a function of Q statistic. The Q statistic was applied to test the assumption of homogeneity among the alpha coefficients. I^2 index is a possible measure of the amount of heterogeneity (Higgins & Thompson, 2002). It can be thought that I^2 values, which are approximately 25%, 50%, and 75%, reflect low, moderate, and large heterogeneity, respectively (Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006).

To interpret the results, the mean effect sizes, their lower and upper confidence intervals obtained with Fisher z-transformation were back-transformed to the original metric of alpha coefficient. The predicted alpha coefficients were evaluated according to the .70 criterion level determined by Nunnally and Bernstein (1994). Values of .70 and above indicate that there is sufficient reliability for the internal consistency of the scale. The effect of the moderator variables on the variability of the reliability estimates was performed through Analog ANOVA. These moderator variables are type of scale (OHS, OHS-S), type of sample (student, non-student), and field of study (social sciences, psychology/health sciences, sport sciences). In addition, the variables of sample type and study field were analyzed as moderators separately for both OHS and OHS-S.

Lastly, publication bias was assessed by Egger's regression test (Egger, Smith, Schneider, & Minder, 1997), Begg and Mazumdar's rank correlation test (Begg & Mazumdar, 1994), Duval and Tweedie Trim and Fill (Duval & Tweedie, 2000a, 2000b) test for funnel plot asymmetry, fail-safe N method. Jamovi and Comprehensive Meta-Analysis V3 free trial (Retrieved from www.meta-analysis.com) was used for statistical analyses.

RESULTS

In this study, a meta-analysis of 95 Cronbach's Alpha coefficients was performed from moderator variables determined by examining literature. The distribution of alpha values in primary studies separately for each scale type is shown in Figure 2. Without the weighting factor, the average reliabilities of the alpha coefficients are .85 (SD = 0.08) and .74 (SD = 0.10) for OHS and OHS-S, respectively. The kurtosis and skewness coefficients are -2.50 (SE = 0.39), 6.73 (SE = 0.76) for OHS; and -1.95 (SE = 0.31), 9.13 (SE = 0.62) for OHS-S.



```
OHS Stem-and-Leaf Plot

 Frequency    Stem &  Leaf
    3.00 Extremes    (=<.65)
    1.00       7.  8
    8.00       8.  01134444
   18.00       8.  556677788889999999
    7.00       9.  0001124

 Stem width:      .10
 Each leaf:       1 case(s)

OHS-S Stem-and-Leaf Plot

 Frequency    Stem &  Leaf
    2.00 Extremes    (=<.41)
    1.00       6.  4
    9.00       6.  788889999
   19.00       7.  0000011222233334444
   16.00       7.  5566667777888899
    6.00       8.  111223
    3.00       8.  556
    2.00 Extremes    (>=.94)

 Stem width:      .10
 Each leaf:       1 case(s)
```

Figure 2. Distributions of Alpha Coefficients for OHS and OHS-S

Table 2 given below presents descriptive results for the estimates of alpha coefficients for general and moderator variables which are back-transformed to the alpha coefficient's original metric. Table 2 also shows 95% confidence interval for the estimated mean Cronbach's Alpha and the highest and lowest alpha values of the studies which constitute the RG meta-analysis.

As shown on the bottom line in Table 2, the reliability or the mean effect size of total OHS scores yielded a mean coefficient of .81 while the lower limit was .78 and the upper limit was .82 in 95% confidence interval. In addition, the reliability of total scores ranged from .29 to .97. Although there was a wide distribution of reliability estimates, the mean reliability estimate and limits of the confidence intervals are acceptable score reliabilities. For total reliability estimates, it can be said that they tend to be large and heterogeneous.

Table 2. Reliability Estimates of Oxford Happiness Scores across Studies for Different Moderator Variables

					95 % Confidence Interval			
Category		*df*	Mean α (SD)	Z Value	Lower	Upper	Min.	Max.
Type of Scale								
OHS		36	0.87* (0.51)	35.98	0.85	0.88	0.50	0.94
OHS-S		57	0.76* (0.50)	34.09	0.73	0.78	0.29	0.97
Type of Sample								
Student	OHS	13	0.89* (0.24)	39.55	0.87	0.91	0.82	0.94
	OHS-S	36	0.75* (0.37)	25.42	0.71	0.78	0.29	0.97
Non-Student	OHS	22	0.85* (0.45)	30.03	0.83	0.87	0.54	0.92
	OHS-S	20	0.77* (0.34)	19.87	0.72	0.81	0.68	0.94
Field of Study								
Social Sciences (SS)	OHS	25	0.87* (0.33)	35.38	0.85	0.88	0.54	0.91
	OHS-S	38	0.77* (0.39)	25.54	0.73	0.80	0.29	0.97
Psychology (P)/Health Sciences (HS)	OHS	3	0.89* (0.10)	15.48	0.85	0.92	0.80	0.92
	OHS-S	12	0.74* (0.21)	13.91	0.67	0.79	0.41	0.85
Sport Sciences (SPS)	OHS	6	0.86* (0.38)	16.11	0.81	0.89	0.64	0.94
	OHS-S	5	0.72* (0.25)	8.67	0.61	0.81	0.68	0.74
Total		94	0.81* (0.72)		0.78	0.82	0.29	0.97

Notes. Min.= minimum94; Max.= maximum; OHS = Oxford Happiness Scale (original form which has consisted of 29 items); OHS-S= The short form of OHS (which has consisted of 7 items) and estimates use a random-effects model.
*$p < .05$

Table 2 also presents the mean alpha coefficients obtained for moderator variables. When the mean alpha coefficient was analyzed according to the type of scale, the mean alpha from the OHS-S was found .76 with a lower limit of .73 and an upper limit of .78 in 95% confidence interval. The mean effect size for the OHS was found .87 while the lower limit was .85, and the upper limit was .88 in 95% confidence interval. The reliability scores ranged between .50-.94 for the OHS and .29-.97 for OHS-S. The reliability scores range showed that especially OHS-S had lower coefficients than the OHS. The minimum reliability coefficients were below .70 for both types of scale (Nunnally & Bernstein, 1994). Again, the mean effect sizes and their 95% confidence interval limits were at an acceptable level for both types of scale. When the mean effect sizes were examined for two types of scale, the mean alpha coefficient for the OHS-S was smaller than OHS.

When Table 2 was examined according to characteristic of sample, the mean effect size estimates were higher in non-student sample for OHS-S. Despite that, the mean alpha value was higher in student sample for OHS. For OHS, the mean effect sizes were .89 and .85, respectively, in student sample and non-student sample. On the other hand, the mean effect sizes were found .75 and .77 respectively in student and non-student sample for OHS-S. For both types of sample there were wide distributions of reported alpha coefficients except OHS in student sample. The lowest reported alpha coefficient (α = .29) was in student sample. And so, the minimum mean effect size was calculated as .75 (95%CI, .71-.78) in this sample.

With regard to field of study, it was seen that the mean alpha estimates were reported for three categories. The mean effect sizes obtained with the alpha coefficients of OHS were for Social Sciences (α = .87, 95%CI, .85-.88), Psychology/Health Sciences (α = .89. 95%CI, .85-.92), and Sport Sciences (α = .86, 95%CI, .81-.89). Also, the mean effect sizes of OHS-S were .77 (95%CI, .73-.80) for Social Sciences, .74 (95%CI, .67-.79) for Psychology/Health Sciences, and .72 (95%CI, .61-.81) for Sport Sciences. According to these results, the mean alpha estimate for the field of Social Science was greater than the other fields for OHS-S. For OHS-S, the mean alpha estimates were almost close for all categories of field of study. For OHS, although the reliability estimates were close, the highest mean alpha value was in the field of psychology/health sciences.

In this study, the heterogeneity of Cronbach Alpha values was investigated. So, I^2 index for the amount of heterogeneity and Q test of homogeneity for the total scale were calculated. According to the results, the Q test was statistically significant with high heterogeneity coefficients. The estimates of Q for the scale was $Q_{\text{Total}}(94) = 2639.66$, $p < .001$. The between-study variance, τ^2, was estimated 0.08 by DerSimonian and Laird method. The I^2 index indicated that in 96.44%, the reliability coefficients had a large variability among the true effect size estimates. In the next step, the effect of the sub-group moderator variables was examined.

The Analog ANOVA was performed to examine the effect of the moderator variables on the variability of the reliability estimates. Whether the mean alpha coefficients differ according to the type of the scale was analyzed with the Analog ANOVA. The result was presented in Table 3.

Table 3. The Results of Analog ANOVA for Type of Scale

Moderator Variable	Categories	*Q* Statistics	df(*Q*)	I^2
Type of Scale	OHS	$Q_{\text{OHS}} = 415.63$*	36	91.34%
	OHS-S	$Q_{\text{OHS-S}} = 1088.35$*	57	94.76%
		$Q_{\text{within}} = 1503.98$*	93	
		$Q_{\text{between(FEM)}} = 1135.69$*	1	
		$Q_{\text{between(REM)}} = 50.75$*	1	
		$Q_{\text{total}} = 2639.66$*	94	

*$p < .05$

As shown in Table 3, Q_{total} of coefficient alpha values was found 2639.66 ($p = .00$), and it was statistically significant. Therefore, it can be said that the true variance estimate of reliability coefficients was statistically significant for all of the studies. In addition, it can be said that the variance within groups was statistically significant at the level of $p < .05$ since Q_{within} was 1503.98 ($p = .00$). When the difference between the groups was examined, it was seen that the $Q_{\text{between(REM)}}$ value was 50.75 ($p = .00$), and this value was significant. Accordingly, it can be said that the alpha coefficient was related to the scale type. When the variance in which the scale type explained for the alpha coefficient was examined, it was found that (1135.69/2639.66) .43 proportion of the true variance or 43.02% of the true variance was explained by the scale type. Based on this value, it can be said that the proportion of explaining the variance in the alpha coefficient of the scale type alone is high. When heterogeneity was examined for different scale types separately, heterogeneity was high in both scale types because *Q* statistics (Q_{OHS} and $Q_{\text{OHS-S}}$) was statistically significant at the level of $p < .05$, and I^2 were 94.76% and 91.34% for OHS-S and OHS, respectively. This may be due to lack of classification according to other variables ignored in this *Q* test. Some of these variables can be administration conditions, sample size, administration year, research type, etc. Due to the significant variance between the scale forms, it would be more meaningful to examine the effect of the moderator variables separately for OHS-S and OHS. As seen in Table 3, Q_{total} for OHS-S form data is also significant which means weighted sum of squares is much more than expected (df, k-1) by random, within study, variation. Forest plots for long form and short form were presented separately in Figure 3 and Figure 4.

When Figure 3 and Figure 4 were examined, it can be said that the reliability coefficients of all individual studies were statistically significant. In addition, it can be stated that the reliability coefficients were generally in the range of .80-1.00 and .60-1.00 for OHS and OHS-S, respectively.

Table 4. The Results of Analog ANOVA for Type of Sample in OHS-S

Moderator Variable	Categories	*Q* Statistics	df(*Q*)	I^2
Type of Sample	Student	$Q_{\text{student}} = 588.52$*	37	93.88%
	Nonstudent	$Q_{\text{nonstudent}} = 394.30$*	21	94.93%
		$Q_{\text{within}} = 982.82$*	56	
		$Q_{\text{between(FEM)}} = 105.53$*	1	
		$Q_{\text{between(REM)}} = 0.57$	1	
		$Q_{\text{total}} = 1088.35$*	57	

*$p < .05$

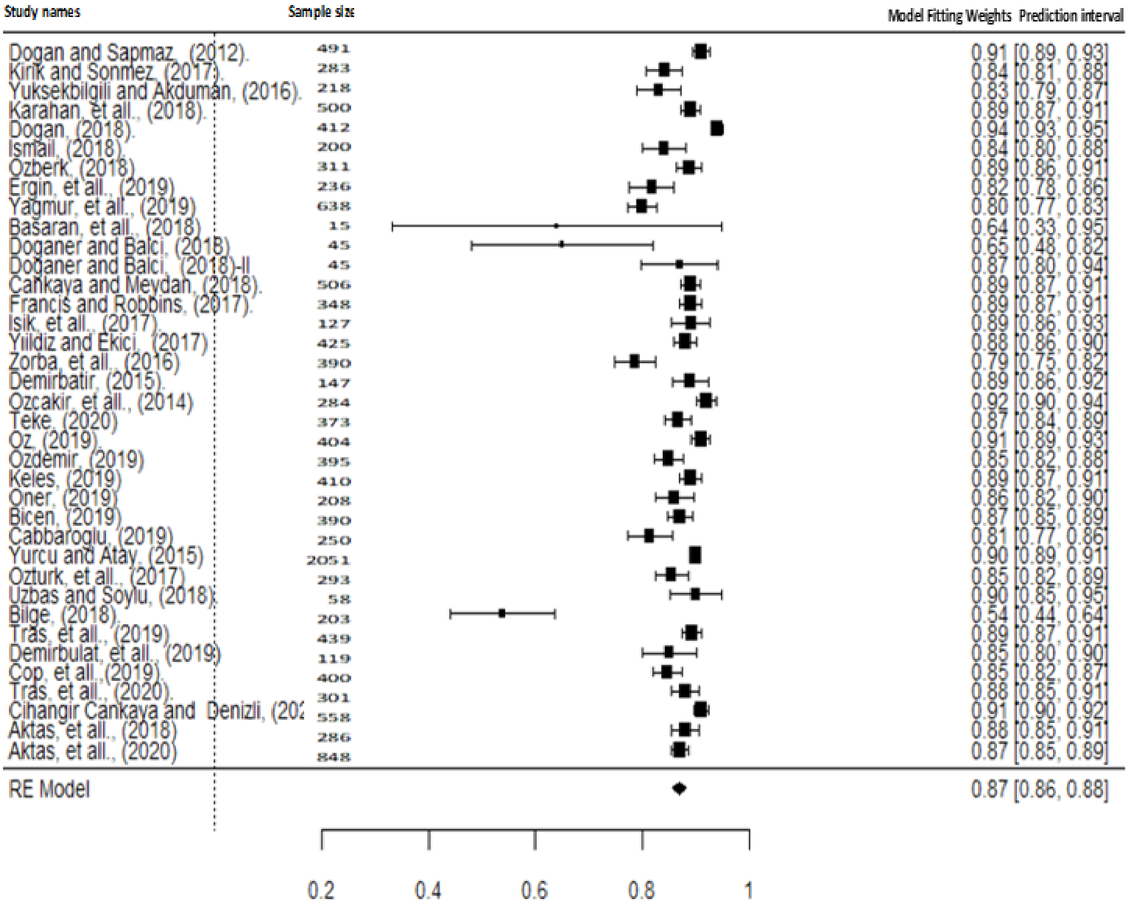


Figure 3. Forest Plot for OHS

The significance of the difference of alpha coefficients according to the type of the sample was analyzed by Analog ANOVA for OHS-S. When Table 4 was examined, it can be seen that the Q_{total} of coefficient alpha values was found 1088.35 (p = .00), and it was statistically significant. Therefore, it can be said that the variance was statistically significant for all of the studies that used OHS-S. In addition, it can be said that the variance within groups was statistically significant at the level of $p <$ .05 since the Q_{within} was 982.82 (p = .00). When the difference between the groups was examined, it was seen that the $Q_{\text{between(FEM)}}$ value was 105.53 (p < .05). Accordingly, whether the sample consists of the students or not did have a statistically significant effect on the variability of alpha coefficient when FEM was used. In this case (105.53/1088.35), .10 proportion of variance or 10% of the true variance was explained by sample groups. However, this group difference could be overcome by using the REM analysis. As can be seen in Table 4, when the REM approach was used in the analysis, this variance was not significant anymore. Accordingly, whether the sample consists of students or not didn't have a statistically significant effect on the variability of alpha coefficients when REM was used. So, it can be said that the alpha coefficient was not related to the sample type of OHS-S for REM. When heterogeneity was examined for different sample types, it was high in both sample types for studies that used OHS-S. Because Q statistics (Q_{student} and $Q_{\text{nonstudent}}$) was statistically significant at the level of p < .05, and I^2 were respectively 93.88% and 94.93% for student sample and non-student sample. This may be due to the lack of classification according to other variables ignored in this Q test. Some of these variables can be study field, administration conditions, sample size, administration year, research type, etc. Some further studies are needed to explain the remaining heterogeneity in OHS-S form reliability estimates.

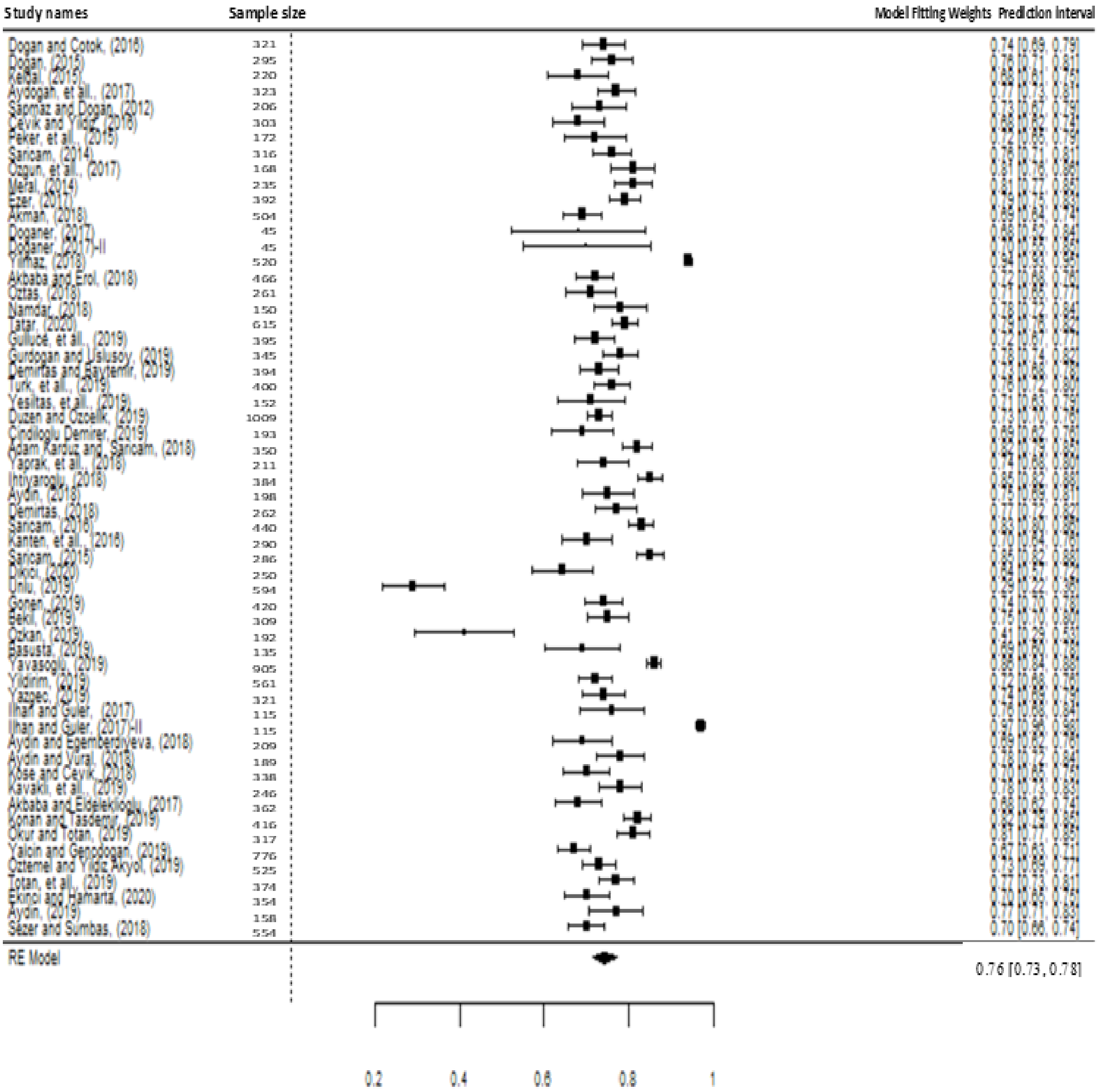


Figure 4. Forest Plot for OHS-S

Table 5. The Results of Analog ANOVA for Type of Sample in OHS

Moderator Variable	Categories	*Q* Statistics	df(*Q*)	I^2
Type of Sample	Student	$Q_{\text{student}} = 76.61$*	13	83.03%
	Nonstudent	$Q_{\text{nonstudent}} = 292.70$*	22	92.48%
		$Q_{\text{within}} = 369.31$*	35	
		$Q_{\text{between(FEM)}} = 46.32$*	1	
		$Q_{\text{between(REM)}} = 7.90$*	1	
		$Q_{\text{total}} = 415.63$*	36	

*$p < .05$

Table 5 presents the significance of the difference of alpha coefficients according to the type of sample for OHS. With regard to Table 5, Q_{total} of coefficient alpha values was found 415.63 (p = .00), and it was statistically significant. In addition, when we examined the variance within the groups, the studies separated by sample type were also heterogeneous in within groups. Q_{within} was 369.31, and p-value was 0.00 ($p < .05$). When the difference between the groups was examined, it was seen that the $Q_{\text{between(FEM)}}$ value was 46.32 ($p < .05$). Accordingly, whether the sample consists of students or not did have a statistically significant effect on the variability of alpha coefficients when FEM was used. In this case (46.32/415.63), .11 proportion of true variance or 11.14% of the true variance was explained by means of sample groups for OHS. Also, this group difference couldn't be overcome by using the REM analysis because, as can be seen in Table 5, when the REM approach was used in the analysis, $Q_{\text{between(REM)}}$ value was 7.90 ($p < .05$), and this value was significant. Accordingly, whether the sample consists of students or not had a still statistically significant effect on the variability of alpha coefficients when REM was used. Therefore, although the heterogeneity was significant in the within groups, it can be said that the mean alpha values of the studies separated according to the sample type differed significantly from each other. And it can be said that the alpha coefficient was related to the sample type for OHS. In addition, based on the proportion of true variance value, it can be said that the proportion of explaining the true variance in the alpha coefficient of the sample type alone is low. When heterogeneity was examined for different sample types, heterogeneity was high in both sample types. Because Q statistics (Q_{student} and $Q_{\text{nonstudent}}$) was statistically significant at the level of $p < .05$, and I^2 values were respectively 83.03% and 92.48% for sample of student and sample of nonstudent. This may be due to the lack of classification according to other variables ignored in this Q test. Some of these variables can be field of study, administration conditions, sample size, administration year, research type, etc. Some further studies are needed to explain the remaining heterogeneity in OHS long-form reliability estimates.

Table 6. The Results of Analog ANOVA for Field of Study in OHS-S

Moderator Variable	Categories	Q Statistics	df(Q)	I^2
Field of Study	Social Sciences	$Q_{\text{Social}} = 948.20$*	38	95.99%
	Psychology/Health Sciences	$Q_{\text{Psychology/Health}} = 129.54$*	12	90.74%
	Sport Sciences	$Q_{\text{Sport}} = 1.31$	5	0.00%
		$Q_{\text{within}} = 1079.06$*	55	
		$Q_{\text{between(FEM)}} = 9.29$*	2	
		$Q_{\text{between(REM)}} = 1.17$	2	
		$Q_{\text{total}} = 1088.35$*	57	

*$p < .05$

The Analog ANOVA was performed to examine whether alpha coefficients showed a statistically significant difference according to field of study for OHS-S. As seen in Table 6, Q_{total} of coefficient alpha values was found 1088.35 ($p = .00$), and it was statistically significant. Therefore, it can be said that variance was statistically significant for all of the studies that used OHS-S. In addition, it can be said that the variance within groups was statistically significant at the level of $p < .05$ since Q_{within} was 1079.06 ($p = .00$). When the difference between the groups was examined, it was seen that the $Q_{\text{between(FEM)}}$ value was 9.29 ($p < .05$). Accordingly, whether the sample consists of students or not did have a statistically significant effect on the variability of alpha coefficients when FEM was used. In this case (9.29/1088.35), a .01 proportion of the true variance or 1.00% of the true variance was explained by field of study for OHS-S. However, this group difference can be overcome by using the REM analysis. As can be seen in Table 6, when the REM approach was used in the analysis, this variance was not significant anymore. Accordingly, whether the sample consists of students or not didn't have a statistically significant effect on the variability of alpha coefficients when REM was used. So, it can be said that the alpha coefficient wasn't related to the field of study for OHS-S for REM. When heterogeneity was examined for different fields of study, heterogeneity was high in social sciences and psychology/health sciences for studies that used OHS-S because Q statistics (Q_{Social} and $Q_{\text{Psychology/Health}}$) was statistically significant at the level of $p < .05$ and I^2 were respectively 95.99% and 90.74% for the field of social sciences and psychology/health sciences. As mentioned before, this may be due to the lack of classification according to other variables ignored in this Q test. Some of these variables can be sample type, administration conditions, sample size, administration year, research type, etc. Some further studies are needed to explain the remaining heterogeneity in OHS-S form reliability estimates. In addition, although heterogeneity was high in the fields of social sciences and psychology/health sciences, it was observed that there was low heterogeneity in the field of sports sciences because Q_{Sport} = 1.31 (p = 0.93) was not statistically significant at the level of $p < .05$, and I^2 was 0.00% for field of sport sciences.

Table 7. The Results of Analog ANOVA for Field of Study in OHS

Moderator Variable	Categories	Q Statistics	df(Q)	I^2
Field of Study	Social Sciences	Q_{Social} = 239.02*	25	89.54%
	Psychology/Health Sciences	$Q_{\text{Psychology/Health}}$ = 82.50*	3	96.36%
	Sport Sciences	Q_{Sport} = 91.12*	6	93.42%
		Q_{within} = 412.64*	34	
		$Q_{\text{between(FEM)}}$ = 2.99	2	
		$Q_{\text{between(REM)}}$ = 1.83	2	
		Q_{total} = 415.63*	36	

*$p < .05$

The Analog ANOVA was performed to examine whether alpha coefficients showed a statistically significant difference according to field of study for OHS. As seen in Table 7, Q_{total} of coefficient alpha values was found 415.63 (p = .00), and it was statistically significant. Therefore, it can be said that the variance was statistically significant for all of the studies that used OHS. In addition, it can be said that the variance within groups was statistically significant at the level of $p < .05$ since Q_{within} was 412.64 (p = .00).

When the difference between the groups was examined, it was seen that the $Q_{\text{between(FEM)}}$ value was 2.99 ($p > .05$). Also, $Q_{\text{between(REM)}}$ value was 1.83 (p = 0.40), and this value wasn't statistically significant. Accordingly, whether the field of study is social sciences, psychology/health sciences, and sport sciences or not did not significantly affect the variability in alpha coefficients for both models. So, it can be said that the alpha coefficient was not related to the field of study for OHS. Already, (2.99/415.63) the 0.01 proportion of true variance or 1.00% of the true variance was explained by study fields for OHS. When heterogeneity was examined for different fields of study, it was high in social sciences, psychology/health sciences, and sport sciences for studies that used OHS-S since Q statistics (Q_{Social}, $Q_{\text{Psychology/Health}}$ and Q_{Sport}) was statistically significant at the level of $p < .05$, and I^2 values were respectively 89.54%, 96.36% and 93.42% for the fields of social sciences, psychology/health sciences, and sport sciences. This may be due to the lack of classification according to other variables ignored in this Q test. Some of these variables can be sample type, administration conditions, sample size, administration year, research type, etc. Some further studies are needed to explain the remaining heterogeneity in OHS long-form reliability estimates. Also, it can be said that the absence of a significant difference between these fields supports the high level of heterogeneity among the groups. As mentioned before, the scale type had a large variability source for the alpha coefficient. On the other hand, the moderator variables which were sample type and field of study weren't seen as the important sources of variability in the alpha coefficients. The reason for the high level of heterogeneity in the same fields of study, sample types, and scale types is that the studies come from different universes.

In this study, the meta-analysis was performed with only published articles and theses. Since the published studies generally have a high or significant effect size, taking only these studies into the meta-analysis may cause publication bias. Therefore, the publication bias was examined by rank correlation and regression test for funnel plot asymmetry and classic fail-safe N method. In the fail-safe N method, assuming the main effect of the studies to be added is zero, it is calculated how many studies are to be added so that the p-value isn't significant. And the calculated number's name is fail-safe N. If only a few studies are needed, there may be a concern that the effect is actually zero (Borenstein, Hedges, Higgins, & Rothstein, 2013). The fail-safe N was calculated as 4975 ($p < .01$). According to these results, it was seen that the number of studies to be added was quite high so that not the summary effect was significant. The other approach, funnel plot asymmetry is seen in Figure 4.

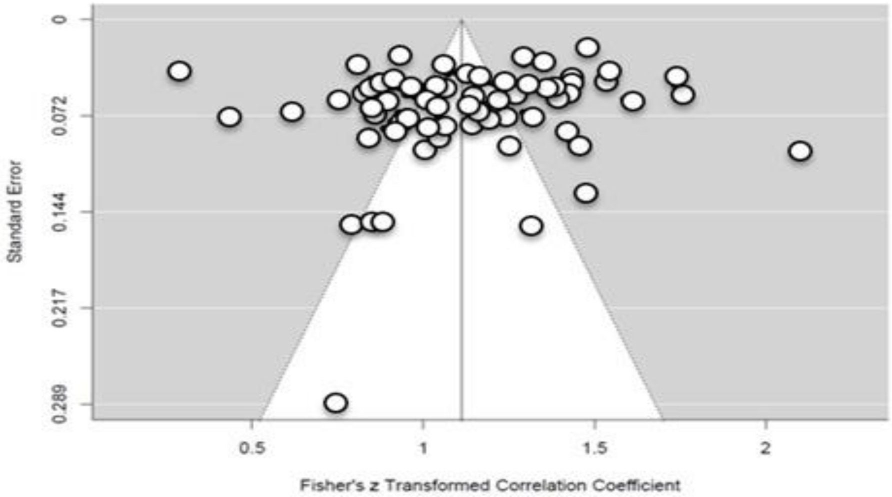


Figure 4. Funnel Plot for All Alpha Coefficients

In the funnel plot, studies are expected to be distributed symmetrically around the summary effect size. Although the studies were seen approximately symmetrically distributed to the right and left of the summary effect size, this interpretation is subjective (Borenstein et al., 2013). The rank correlation and regression tests were performed for more an objective interpretation. According to Egger's regression test, the regression intercept was not significant (intercept = -2.94, *p* = .09). The hypothesis was accepted to show that the regression constant didn't deviate from zero significantly. Begg and Mazumdar's rank correlation test also contributed to the lack of asymmetry in the funnel plot. According to that, Kendall's tau was not significant (Kendall's tau = -0.057, *p* = .41). It can be interpreted that there wasn't an asymmetry in the funnel diagram. In addition, according to Duval and Tweedie Trim and Fill test, there was no difference between the observed effect size and true effect size which was created to correct the effect caused by publication bias. As a result of the general symmetrical distribution of studies on both sides of the overall effect size, the difference was found zero. So the statistical tests for funnel plot asymmetry did not show any evidence of publication bias. Therefore, it can be said that all results were not likely to be the result of publication bias.

DISCUSSION and CONCLUSION

In this study, a meta-analytical reliability generalization analysis was conducted on OHS and OHS-S. In addition, it was investigated whether Cronbach's Alpha was affected by sample type, scale type, and study field. The results of this study showed that the mean Cronbach's Alpha coefficients obtained from both the OHS and OHS-S were at an acceptable level. The fact that these coefficients are high is an indication of the usability of the scale by both practitioners and researchers. When it was examined whether the reliability coefficient was affected by the scale type, a significant difference was observed between the two scale forms according to both REM and FEM. This difference was observed for the favor of OHS. Accordingly, it can be said that the measures obtained with the OHS, in general, are more reliable. In general, it is thought that more sensitive and more reliable measurements will be made as the number of items increases. And the results of this study support this idea. When the results of other studies were examined, it was seen that similar results were found. For example, Henson et al. (2001) and Nilsson et al. (2002) observed that reliability was higher for the long-form. Henson et al. (2001) stated that as the length of the test increase, the reliability estimates increase in all subscales except one. Also, Nilsson et al. (2002) found that the CDMSE long form's reliability coefficients were higher than the short form's. In contrast, Hanson et al. (2002), Hess et al. (2014), and Vacha-Haase (1998) observed that reliability was higher for the short form. Hanson et al. (2002) observed that the mean reliability coefficient obtained from the short form was slightly higher for both client and therapist versions of the Working Alliance Inventory. Hess et al. (2014) and Vacha-Haase (1998) similarly concluded that the measures obtained from the short form were more reliable in a result of the RG analysis. As can be seen, while the effect of test length on reliability varies in the studies, it was observed that the reliability coefficient increases significantly as the test length increases for OHS in the Turkish sample. In the tests, it is recommended that the length of the test is as short as possible in terms of usability and it is long enough for acceptable reliability (McDonald, 1999). In this regard, the OHS is considered to be appropriate in terms of usability as it will not take much time to respond. Also, as a result of this study, it is obvious that the mean alpha coefficient of OHS for the Turkish sample is higher than OHS-S. In line with all of these, administering OHS instead of OHS-S may be more suitable for reliability for the Turkish sample. However, this situation may vary with the variance explained by the scale, the properties of the administration group, administration conditions, etc.

When the effects of the sample type on reliability were examined, it was seen that reliability was significantly different for students and non-students for the OHS-S according to FEM. But the true variance explained by sample groups was low for OHS-S. Although there was a significant effect in FEM, this group difference could be overcome by REM analysis. Therefore, researchers and practitioners may be advised to use REM analysis for such group differences. As a result, it was seen that the reliability wasn’t significantly different for students and non-students for the OHS-S when REM was used. When the effects of the sample type on reliability were examined for OHS, it was seen that the reliability was significantly different for students and non-students for the OHS, according to both REM and FEM. Similar to these results, as it was in REM for OHS-S, Hess et al. (2014), Thompson and Cook (2002), Wallace and Weller (2002) found that reliability wasn’t affected by the sample type. While Hess et al. (2014) separated sample types as student and professional, it was observed that Thompson and Cook (2002) distinguished as undergraduate, graduate, and faculty. And it was stated that there was no variability between the reliability coefficients of the groups in both studies. Also, Graham et al. (2011) concluded that the relationship between the proportion of college students in the sample and the reliability coefficient regarding the scores obtained with Locke–Wallace Marital Adjustment Test (LWMAT), Kansas Marital Satisfaction Scale (KMS), Quality of Marriage Index (QMI), and Marital Opinion Questionnaire (MOQ) wasn’t significant. Contrary to these studies and similar to the results of OHS long-form Caruso et al. (2001), Vacha-Haase (1998), and Yin and Fan (2000) concluded that the sample type (student/non-student) affected reliability coefficients. As can be seen, while the effect of sample type on reliability varies in the studies, it was observed that the reliability coefficient didn't differ in student or non-student samples for OHS-S according to REM analysis in this study. The difference between the alpha coefficients is almost negligible, with about two per thousand for OHS-S. Also, the mean alpha coefficients were found to be high in both groups. These results are indicators of the availability of the OHS-S for both students and non-students for Turkish sample. In addition, while the effect of sample type on reliability varies in the studies, it was observed that the reliability coefficient differs for student or non-student samples for OHS in this study according to FEM and REM. This difference between the alpha coefficients was about four per thousand. But the proportion of explaining the true variance in the alpha coefficient of the sample type alone was quite low. Therefore, this difference wasn’t at an important level. In addition, the mean alpha for student sample was higher than non-student sample, and the mean alpha coefficients were at an acceptable level for both OHS and OHS-S. The development of the OHS by applying it to students may be a factor in this. So, OHS is more suitable for students, but it can be used for both sample types. To summarize, it can be said to researchers and practitioners that both OHS and OHS-S can be used for both student groups and non-student groups for Turkish sample.

Finally, in the scope of the research, it was investigated how mean alpha was in different fields of study and whether the mean difference in reliability estimation was significant or not for these fields for both OHS and OHS-S. In OHS-S, the highest mean alpha was found in the field of social sciences, while the lowest mean alpha was found in the sport sciences. Also, the highest mean alpha was found in the field of psychology/health sciences, while the lowest mean alpha was found in the sport sciences for OHS. In line with all results, no major changes were observed in the reliability coefficients in all fields for all OHS forms according to REM. But in FEM analysis, the mean difference in reliability estimation was significant for OHS-S. When it was examined how much true variance was explained with field of study, it was seen that explained variance was low and not important for OHS-S. Although there was a significant effect in FEM, this group difference could be overcome by REM analysis. Therefore, researchers and practitioners may be advised to use REM analysis for such group differences. As a result, it was seen that reliability wasn't significantly different for field of study for the OHS-S when REM was used. Based on this, it can be said that field of study is generally not effective in reliability estimation for all OHS forms. When the researches in the literature were investigated, it was seen that Vicent et al. (2019) analyzed the effect of study focus on reliability estimation by classifying study focus as applied and psychometric. And they concluded that the effect of study focus on reliability estimation for CAPS sub-dimensions was significant. However, they found that the variance in which the variability in reliability was explained by the study focus was low, and in meta-regression analysis, they found that it was the variable that least explained variance. It can be that the reason why this study is different from Vicent et al.'s (2019) research is that the study focus and field of study concepts are distinct each other and the classification is made differently. In such a case, the different measurement tools may have affected the differentiation of the results. Another study in the literature classified the study type as medical and nonmedical and research design as psychometric/others and experimental/others (Barnes et al., 2002). But they stated that they didn't examine the relationship between reliability and study type; they examined how much reliability was reported in journals in different contexts. Also, they founded that there were very low correlations between internal consistency coefficient and the contexts of psychometric/others or experimental/others. The low correlations found are similar to this study, but although research design and fields of study are similar, they are not the same. Consequently, reliability coefficients didn't differ significantly in different fields of study according to REM and were acceptable for all of them. Therefore, it is thought that the OHS's forms can be used in different fields.

The reliability estimates of OHS and OHS-S showed acceptable level in the present study. However, as mentioned above, as each measurement depends on the different conditions of the sample or settings, the results in this study are specific to these conditions. Therefore, it is necessary to calculate reliability based on their own data, besides the RG studies (Capraro & Capraro, 2002). To summarize in general, administering OHS instead of OHS-S may be more suitable for reliability for Turkish sample. Also, scale forms can be used for both student sample and non-student sample and can be used for each field of study. But, in generally, it is suggested that REM analysis should be performed for these variables since some group difference can be overcome when REM is used. Finally, in deciding which form of OHS to use, this situation may vary with the variance explained by the scale, the properties of administration group, administration conditions, etc.

The limitations of this study are transforming Cronbach's Alpha coefficients into Fisher Z scores, examining the variables of scale type, sample type, and field of study as sources of measurement error of reliability, and performing the analysis in the CMA program. Future investigations can examine and compare the reliability estimates which use other reliability estimators like Hakstian-Whalen (1976) and Bonett (2002) transform, etc. The RG studies can be made for other reliability estimates which address different sources of measurement errors. As the different sources of variability, study language, sample size, year of study, race, gender, age, marital status, mean and standard deviation of the measurements obtained from the scale, reliability type, research design, different sample type, etc. can be selected.

REFERENCES

Aguayo, R., Vargas, C., Emilia, I., & Lozano, L. M. (2011). A meta-analytic reliability generalization study of the Maslach Burnout Inventory. *International Journal of Clinical and Health Psychology*, *11*(2), 343-361. Retrieved from http://www.redalyc.org/articulo.oa?id=33716996009

Anastasi, A. (1982) *Psychological testing* (5th ed.). New York: Macmillan.

Argyle, M., Martin, M., & Crossland, J. (1989). Happiness as a function of personality and social encounters. In J. P. Forgas & J. M. Innes (Eds.), *Recent advances in social psychology: An international perspective* (pp. 189- 203). Amsterdam: North Holland, Elsevier Science.

Argyle, M., Martin, M., & Lu, L. (1995). Testing for stress and happiness: The role of social and cognitive factors. In C. D. Spielberger, I. G. Sarason, J. M. T. Brebner, E. Greenglass, P. Laungani, & A. M. O'Roark (Eds.), *Series in stress and emotion: Anxiety, anger, and curiosity,* (Vol. 15). *Stress and emotion: Anxiety, anger, and curiosity* (p. 173-187). Washington, DC: Taylor & Francis.

Armor, D. J. (1973). Theta reliability and factor scaling. *Sociological Methodology, 5*(1973-1974), 17-50. doi: 10.2307/270831

Barnes, L. L., Harp, D., & Jung, W. S. (2002). Reliability generalization of scores on the Spielberger state-trait anxiety inventory. *Educational and Psychological Measurement*, *62*(4), 603-618. doi: 10.1177/0013164402062004005

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*(4), 1088-1101. doi: 10.2307/2533446

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

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics*, *27*(4), 335-340. doi: 10.3102/10769986027004335

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2013). *Introduction to meta-analysis*. UK: John Wiley & Sons.

Bornmann, L., Mutz, R., & Daniel, H. D. (2010). A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PloS One*, *5*(12), 1-10. doi: 10.1371/journal.pone.0014331

Capraro, R. M., & Capraro, M. M. (2002). Myers-briggs type indicator score reliability across: Studies a meta-analytic reliability generalization study. *Educational and Psychological Measurement,62*(4), 590-602. doi: 10.1177/0013164402062004004

Caruso, J. C. (2000). Reliability generalization of the NEO personality scales. *Educational and Psychological Measurement*, *60*(2), 236-254. doi: 10.1177/00131640021970484

Caruso, J. C., Witkiewitz, K., Belcourt-Dittloff, A., & Gottlieb, J. D. (2001). Reliability of scores from the Eysenck Personality Questionnaire: A reliability generalization study. *Educational and Psychological Measurement*, *61*(4), 675-689. doi: 10.1177/00131640121971437

Compton, W. C., & Hoffman, E. (2019). *Positive psychology: The science of happiness and flourishing*. Thousands Oak, CA: SAGE Publications.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Ohio: Cengage Learning.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. doi: 10.1007/bf02310555

Cronbach, L. J., Schönemann, P., & McKie, D. (1965). Alpha coefficients for stratified-parallel tests. *Educational and Psychological Measurement, 25*(2), 291-312. doi: 10.1177/001316446502500201

Demir, Ü. (2020). Aile özellikleri ve mutluluk: Çanakkale'de lise öğrencileri üzerine bir araştırma. *Kastamonu Eğitim Dergisi*, *28*(3), 1296-1306. Retrieved from https://dergipark.org.tr/en/download/article-file/1109602

Doğan, T., & Çötok, N. A. (2011). Oxford mutluluk ölçeği kısa formunun Türkçe uyarlaması: Geçerlik ve güvenirlik çalışması. *Türk Psikolojik Danışma ve Rehberlik Dergisi*, *4*(36), 165-172. Retrieved from http://www.turkpdrdergisi.com/index.php/pdr/article/view/100/101

Doğan, T., & Sapmaz, F. (2012). Oxford mutluluk ölçeği Türkçe formunun psikometrik özelliklerinin üniversite öğrencilerinde incelenmesi. *Düşünen Adam Psikiyatri ve Nörolojik Bilimler Dergisi*, *25*(4), 297-304. Retrieved from https://dusunenadamdergisi.org/storage/upload/pdfs/1586247664-tr.pdf

Duval, S., & Tweedie, R. (2000a). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*(449), 89-98. doi: 10.1080/01621459.2000.10473905

Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455-463. doi: 10.1111/j.0006-341x.2000.00455.x

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj*, *315*(7109), 629-634. doi: 10.1136/bmj.315.7109.629

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *The American Council on Education/Macmillan series on higher education. Educational measurement* (pp. 105-146). New York: Macmillan Publishing Co., Inc.; American Council on Education.

Field, A. P. (2003). The problems in using fixed-effects models of meta-analysis on real-world data. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, *2*(2), 105-124. doi: 10.1207/S15328031US0202_02

Francis, L. J., & Crea, G. (2018). Happiness matters: Exploring the linkages between personality, personal happiness, and work-related psychological health among priests and sisters in Italy. *Pastoral Psychology*, *67*(1), 17-32. doi: 10.1007/s11089-017-0791-z

Francis, L. J., & Katz, Y. J. (2000). Internal consistency reliability and validity of the Hebrew translation of the Oxford Happiness Inventory. *Psychological Reports*, *87*(1), 193-196. doi: 10.2466/pr0.2000.87.1.193

Francis, L. J., Brown, L. B., Lester, D., & Philipchalk, R. (1998). Happiness as stable extraversion: A cross-cultural examination of the reliability and validity of the Oxford Happiness Inventory among students in the UK, USA, Australia, and Canada. *Personality and Individual Differences*, *24*(2), 167-171. doi: 10.1016/S0191-8869(97)00170-0

Francis, L. J., Ok, Ü., & Robbins, M. (2017). Religion and happiness: A study among university students in Turkey. *Journal of religion and health*, *56*(4), 1335-1347. Retrieved from http://glyndwr.repositorytest.guildhe.ac.uk:8080/9182/1/Robbins_Religion_and_happiness.pdf

Gilmer, J. S., & Feldt, L. S. (1983). Reliability estimation for a test with parts of unknown lengths. *Psychometrika*, *48*(1), 99-111. doi: 10.1007/BF02314679

Graham, J. M., Diebels, K. J., & Barnow, Z. B. (2011). The reliability of relationship satisfaction: A reliability generalization meta-analysis. *Journal of Family Psychology*, *25*(1), 39-48. retrieved from https://pdfs.semanticscholar.org/4bad/75bc14e345e7dabfc01888b50b5ef502305d.pdf

Graham, J. M., Liu, Y. J., & Jeziorski, J. L. (2006). The dyadic adjustment scale: A reliability generalization meta-analysis. *Journal of Marriage and Family*, *68*(3), 701-717. doi:10.1111/j.1741-3737.2006.00284.x

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255-282. doi: 10.1007/bf02288892

Hakstian, A. R., & Whalen, T. E. (1976). A k-sample significance test for independent alpha coefficients. *Psychometrika*, *41*(2), 219-231. doi: 10.1007/BF02291840

Hanson, W. E., Curry, K. T., & Bandalos, D. L. (2002). Reliability generalization of working alliance inventory scale scores. *Educational and Psychological Measurement*, *62*(4), 659-673. doi: 10.1177/0013164402062004008

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77-89. doi: 10.1080/19312450709336664

Heise, D. R., & Bohrnstedt, G. W. (1970). Validity, invalidity, and reliability. *Sociological Methodology*, *2*(1970), 104-129. doi: 10.2307/270785

Henson, R. K., Kogan, L. R., & Vacha-Haase, T. (2001). A reliability generalization study of the teacher efficacy scale and related instruments. *Educational and Psychological Measurement*, *61*(3), 404-420. doi: 10.1177/00131640121971284

Hess, T. J., McNab, A. L., & Basoglu, K. A. (2014). Reliability generalization of perceived ease of use, perceived usefulness, and behavioral intentions. *Mis Quarterly*, *38*(1), 1-28. Retrieved from https://www.jstor.org/stable/26554866

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

Hills, P., & Argyle, M. (2002). The Oxford Happiness Questionnaire: A compact scale for the measurement of psychological well-being. *Personality and Individual Differences*, *33*(7), 1073-1082. doi: 10.1016/S0191-8869(01)00213-6

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I² index? *Psychological Methods*, *11*(2), 193-206. doi: 10.1037/1082-989X.11.2.193

İlhan, M., & Güler, N. (2017). Likert tipi ölçeklerde olumsuz madde ve kategori sayısı sorunu: Rasch modeli ile bir inceleme. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *8*(3), 321-343. doi: 10.21031/epod.321057

Kline, T. J. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks, CA: Sage Publications.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage.

Kristof, W. (1963). The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. *Psychometrika, 28*(3), 221-238. doi: 10.1007/bf02289571

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151-160. doi: 10.1007/bf02288391

Li, A., & Bagger, J. (2007). The Balanced inventory of desirable responding (BIDR): A reliability generalization study. *Educational and psychological measurement, 67*(3), 525-544. doi: 10.1177/0013164406292087

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Journal of Clinical Epidemiology, 62*(10), 1-34. doi: 10.1016/j.jclinepi.2009.06.006

Lin, C. Y., Imani, V., Griffiths, M. D., & Pakpour, A. H. (2020). Psychometric properties of the Persian Generalized Trust Scale: Confirmatory factor analysis and Rasch models and relationship with quality of life, happiness, and depression. *International Journal of Mental Health and Addiction*. doi: 10.1007/s11469-020-00278-0

Lu, L., & Shih, J. B. (1997). Personality and happiness: Is mental health a mediator? *Personality and Individual Differences*, *22*(2), 249-256. doi: 10.1016/S0191-8869(96)00187-0

McDonald, R. P. (1985). *Factor analysis and related methods*. London: Lawrence Erlbaum Associates Publishers.

McDonald, R. P. (1999). *Test theory: A unified treatment.* London: Lawrence Erlbaum Associates Publishers.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009) Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med, 6*(7), 1-6. doi: 10.1371/journal.pmed.1000097

Nilsson, J. E., Schmidt, C. K., & Meek, W. D. (2002). Reliability generalization: An examination of the career decision-making self-efficacy scale. *Educational and Psychological Measurement*, *62*(4), 647-658. doi: 10.1177/0013164402062004007

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

Okur, S., & Totan, T. (2019). Psikolojik iyi oluşu değerlendiren Bradburn duygulanım dengesi ölçeğinin Türkçede incelenmesi. *Adnan Menderes Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *6*(2), 1-12. Retrieved from https://dergipark.org.tr/en/download/article-file/935147

Raju, N. S. (1977). A generalization of coefficient alpha. *Psychometrika*, *42*(4), 549-565. doi: 10.1007/bf02295978

Revelle, W. (1979). Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behavioral Research, 14*(1), 57-74. doi: 10.1207/s15327906mbr1401_4

Romano, J. L., Kromrey, J. D., & Hibbard, S. T. (2010). A Monte Carlo study of eight confidence interval methods for coefficient alpha. *Educational and Psychological Measurement*, *70*(3), 376-393. doi: 10.1177/0013164409355690

Sánchez-Meca, J., López-López, J. A., & López-Pina, J. A. (2013). Some recommended statistical analytic practices when reliability generalization studies are conducted. *British Journal of Mathematical and Statistical Psychology*, *66*(3), 402-425. doi: 10.1111/j.2044-8317.2012.02057.x

Shields, A. L., & Caruso, J. C. (2003). Reliability generalization of the alcohol use disorders identification test. *Educational and Psychological Measurement*, *63*(3), 404-413. doi: 10.1177/0013164403063003004

Shields, A. L., & Caruso, J. C. (2004). A reliability induction and reliability generalization study of the CAGE questionnaire. *Educational and Psychological Measurement*, *64*(2), 254-270. doi: 10.1177/0013164403261814

Taşdibi-Ünlü, F. (2019). *Üniversite öğrencilerinde yaşam değeri, yaşamın anlamı ve sosyal iyi olmanın mutluluğu yordamadaki rolü* (Yüksek lisans tezi). Muğla Üniversitesi Eğitim Bilimleri Enstitüsü, Muğla. http://tez2.yok.gov.tr/

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

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement*, *60*(2), 174-195. doi: 10.1177/0013164400602002

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

Vacha-Haase, T., & Thompson, B. (2011). Score reliability: A retrospective look back at 12 years of reliability generalization studies. *Measurement and Evaluation in Counseling and Development*, *44*(3), 159-168. doi: 10.1177/0748175611409845

Vacha-Haase, T., Henson, R. K., & Caruso, J. C. (2002). Reliability generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement*, *62*(4), 562-569. doi: 10.1177/0013164402062004002

Vicent, M., Rubio-Aparicio, M., Sánchez-Meca, J., & Gonzálvez, C. (2019). A reliability generalization meta-analysis of the child and adolescent perfectionism scale. *Journal of Affective Disorders*, *245*(2019), 533-544. doi: 10.1016/j.jad.2018.11.049

Viswesvaran, C., & Ones, D. S. (2000). Measurement error in "Big Five Factors" personality assessment: Reliability generalization across studies and measures. *Educational and Psychological Measurement*, *60*(2), 224-235. doi: 10.1177/00131640021970475

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

Wheeler, D. L., Vassar, M., Worley, J. A., & Barnes, L. L. (2011). A reliability generalization meta-analysis of coefficient alpha for the Maslach Burnout Inventory. *Educational and Psychological Measurement, 71*(1), 231-244. doi: 10.1177/0013164410391579

Yıldırım, O., & Sezer, Ö. (2020). The relationship between nomophobia and trait anxiety, basic psychological needs, happiness in adolescents. *Journal of Human Sciences, 17*(2), 535-547. Retrieved from https://j-humansciences.com/ojs/index.php/IJHS/article/view/5917/3375

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

Appendix A. Search Databases

- Google Scholar,
- YOK (Higher Education Institution in Turkey) national thesis/dissertation center,
- Gazi University Central Library
  - EBSCOhost
    - Academic Research Complete,
    - Applied Science & Business Periodicals Retrospective,
    - Applied Science & Technology Index Retrospective,
    - Art Index Retrospective,
    - BIR Entertainment,
    - Book Index with Reviews,
    - Business Source Complete,
    - CINAHL Plus,
    - Dentistry & Oral Sciences Source,
    - ebook Collection,
    - Education Index Retrospective,
    - E-Journals, ERIC,
    - European Views of the Americas,
    - GreenFILE,
    - Humanities & Social Sciences Index Retrospective,
    - Library-Information Science & Technology Abstracts,
    - MasterFILE Complete,
    - MasterFILE Reference ebook Collection,
    - MathSciNet,
    - MEDLINE,
    - Newspaper Source Plus,
    - Newswires,
    - OpenDissertations,
    - Regional Business News,
    - Social Sciences Index Retrospective,
    - SPORTDiscus,
    - Teacher Reference Center,
    - ULAKBIM Turkish National Databases,
    - Web News
- Aydin Adnan Menderes University Library
  - ASCE,
  - BMJ,
  - CabDirect,
  - Cambridge University Press,
  - Clinical Key,
  - DergiPark,
  - Cochrane Library,
  - DOAJ,
  - Emerald Premier,
  - GALE-Archives Unbound,
  - IEEE/IEE Electronic Library,
  - IGI Global,
  - JOVE,
  - JSTOR
  - Archive Journal Content,
  - Nature Journals All/Academic Journals,
  - OECD,
  - Oxford Journals Online,
  - OVID Journals,
  - Philosophy Documentation Center,
  - SAGE,
  - Science Direct,
  - Springer Link/Palgrave Macmillan Journals,
  - Taylor & Francis,
  - UptoDate,
  - Wiley Online Library

Appendix B. Studies in Meta-Analysis

Adam-Karduz, F. F., & Saricam, H. (2018). The relationships between positivity, forgiveness, happiness and revenge. *Romanian Journal for Multidimensional Education/Revista Romaneasca pentru Educatie Multidimensionala, 10*(4), 1-22. doi: 10.18662/rrem/68

Akbaba, A. Y., & Eldeleklioğlu, J. (2019). Adaptation of positive mental health scale into Turkish: A validity and reliability study. *Journal of Family Counseling and Education*, *4*(1), 44-54. doi: 10.32568/jfce.569976

Akbaba, T. P., & Erol, D. (2019). Üniversite öğrencilerinde romantik ilişkilerde akılcı olmayan inançlar ile mutluluk arasındaki ilişkinin incelenmesi. *International Journal of Current Approaches in Language, Education and Social Sciences*, *1*(1), 32-44. Retrieved from https://dergipark.org.tr/en/download/article-file/771960

Akman, E. (2018). *Sağlık profesyonellerinde akış deneyiminin öznel iyi oluş üzerine etkisi: Bir kamu ve özel hastane örneği* (Master's thesis). Marmara University, Sağlık Bilimleri Enstitüsü, İstanbul. Retrieved from https://tez2.yok.gov.tr/

Aktaş, E., Şahin, N., & Gürbüz, E. C. (2018). Mutluluk ve tüketim arasındaki ilişki: Mersin Üniversitesi İİBF öğrencileri örneği. *BENGİ Dünya Yörük-Türkmen Araştırmaları Dergisi*, *2018*(2), 86-104. Retrieved from https://dergipark.org.tr/en/pub/bengi/issue/52630/693084

Aktaş, E., Şahin, N., & Gürbüz, E. C. (2020). Tüketimin mutluluk üzerindeki etkisi: Çukurova Bölgesi örneği. *Bulletin of Economic Theory and Analysis*, *5*(1), 21-40. doi: 10.25229/beta.563857

Aydın, M. (2018). Genç yetişkinlerde mutluluğun özgünlük ve kişisel erdemler açısından incelenmesi. *Anemon Muş Alparslan Üniversitesi Sosyal Bilimler Dergisi*, *6*(6), 1023-1030. doi: 10.18506/anemon.424118

Aydın, M. (2019). Genç yetişkinlerde mutluluk, maneviyat ve kanaat. *The Journal of Social Science*, *3*(6), 439-448. doi: 10.30520/tjsosci.571198

Aydın, M., & Egemberdiyeva, A. (2018). Üniversite öğrencilerinin psikolojik sağlamlık düzeylerinin incelenmesi. *Türkiye Eğitim Dergisi*, *3*(1), 37-53. Retrieved from https://dergipark.org.tr/tr/download/article-file/496083

Aydın, M., & Vural, G. Z. (2018). Üniversite öğrencilerinin beden imgelerinin yaşam niteliklerine etkisi. *Eğitim Kuram ve Uygulama Araştırmaları Dergisi*, *4*(3), 111-121. Retrieved from http://ekuad.com/articles/universite-ogrencilerinin-beden-imgelerinin-yasam-niteliklerine-etkisi.pdf

Aydoğan, D., Özbay, Y., & Büyüköztürk, Ş. (2017). Özgünlük Ölçeği'nin uyarlanması ve özgünlük ile mutluluk arasındaki ilişkide maneviyatın aracı rolü. *The Journal of Happiness & Well-Being*, *5*(1), 38-59. Retrieved from https://www.researchgate.net/profile/Didem_Aydogan2/publication/316034028

Basaran, Z., Çalışkan, F., Çolak, S., & Erdal, R. (2018, Nisan). *Sportif rekreasyon etkinliklerinin yaşlıların mutluluk ve mental iyi oluş düzeylerine etkisi*. Congress Papers of the Association of Sports Sciences, Alanya, Antalya. Retrieved from http://eds.a.ebscohost.com/eds/pdfviewer/pdfviewer?vid=4&sid=43845fc5-1a53-462f-8805-967ec779916f%40sdc-v-sessmgr02

Başusta, B. (2019). *Ergen mutluluğunun ebeveynlerin kişilik özellikleri ve evlilik çatışması açısından incelenmesi.* (Master's thesis). Fatih Sultan Mehmet Vakif University, Lisansüstü Eğitim Enstitüsü, İstanbul. Retrieved from https://tez2.yok.gov.tr/

Bekil, M. (2019). *Öğretmenlerde mutluluğun yordayıcıları olarak sosyal bağlılık, özgecilik ve sosyal empati.* (Master's thesis). Mugla Sitki Kocman University, Eğitim Bilimleri Enstitüsü, Mugla. Retrieved from https://tez2.yok.gov.tr/

Biçen, G. (2019). *İşgören mutluluk düzeylerinin iş tatmini ve iş performansı üzerine etkileri: Konaklama işletmelerinde bir inceleme* (Master's Thesis). Haci Bayram Veli University, Lisansüstü Eğitim Enstitüsü, Ankara. Retrieved from http://tez2.yok.gov.tr/

Bilge, H. (2018). Belediye çalışanlarının mutlulukları: Manisa büyükşehir belediyesi örneği. *Dokuz Eylül Üniversitesi İktisadi İdari Bilimler Fakültesi Dergisi*, *33*(2), 657-678. doi: 10.24988/deuiibf.2018332808

Cabbaroğlu, M. (2019). *Sportif rekreasyon etkinliği olarak zumba ve pilatesin yaşam doyumu ve mutluluk üzerine etkisi (Mugla İli örneği)* (Master's Thesis). Muğla Sitki Kocman University, Sosyal Bilimler Enstitüsü, Mugla. Retrieved from http://tez2.yok.gov.tr/

Cihangir-Çankaya, Z., & Denizli, S. (2020). An explanation of happiness with secure attachment, basic psychological needs and hope: the case of Turkish university students. *European Journal of Educational Research, 9*(1), 433-444. doi: 10.12973/eu-jer.9.1.433

Cihangir-Çankaya, Z., & Meydan, B. (2018). Ergenlik döneminde mutluluk ve umut. *Electronic Journal of Social Sciences, 17*(65), 207-222. doi: 10.17755/esosder.316977

Cindiloğlu-Demirer, M. (2019). Kişi örgüt uyumunun iş performansı üzerine etkisi: Mutluluğun aracılık rolü. *Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi, 33*(1), 283-301. Retrieved from https://hdl.handle.net/11491/5438

Cop, R., Topuz, Y. V., Eru, O., & Yüzüak, A. (2019). Effects of relational experience and subjective well being on customer satisfaction. *Third Sector Social Economic Review, 54*(1), 502-513. doi: 10.15659/3.sektor-sosyal-ekonomi.19.03.1102

Çevik, G. B., & Yıldız, M. A. (2016). Pedagojik formasyon öğrencilerinde umutsuzluk ile mutluluk arasındaki ilişkide benlik saygısının aracılık rolü. *Dicle Üniversitesi, Eğitim Fakültesi Dergisi*, *27*(2016), 96-107.

Demirbatır, R. E. (2015). Relationships between psychological well-being, happiness, and educational satisfaction in a group of university music students. *Educational Research and Reviews*, *10*(15), 2198-2206. doi: 10.5897/ERR2015.2375

Demirtaş, A. S. (2018). Duygu düzenleme stratejileri ve benlik saygısının mutluluğu yordayıcılığı. *Electronic Turkish Studies*, *13*(11), 487-503. doi: 10.7827/TurkishStudies.13465

Demirtaş, A. S., & Baytemir, K. (2019). Warwick-Edinburgh mental iyi oluş ölçeği kisa formu'nun Türkçe'ye uyarlanması: Geçerlik ve güvenirlik çalışması. *Electronic Journal of Social Sciences*, *18*(70), 654-666. doi: 10.17755/esosder.432708

Dikici, İ. (2020). *Serbest zamanlarını gençlik merkezlerinde değerlendiren üniversite öğrencilerinin serbest zaman doyum, yaşam doyum ve mutluluk düzeylerinin incelenmesi* (Master's thesis). Mugla Sitki Kocman University, Sosyal Bilimler Enstitüsü, Mugla. Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi

Doğan, M. N. (2018). *Hukuk fakültesi öğrencilerinin rekreasyonel aktivitelerden elde ettikleri faydaların ve mutluluk düzeylerinin incelenmesi* (Master's thesis). Gazi University, Sağlık Bilimleri Enstitüsü, Ankara. Retrieved from http://tez2.yok.gov.tr/

Doğan, T. (2015). Kısa psikolojik sağlamlık ölçeği'nin Türkçe uyarlaması: Geçerlik ve güvenirlik çalışması. *The Journal of Happiness & Well-Being*, *3*(1), 93-102. Retrieved from https://www.tayfundogan.net/wp-content/uploads/2016/09/K%C4%B1saPsikolojikSaglamlikOlcegi.pdf

Doğan, T., & Çötok, N. A. (2016). Oxford mutluluk ölçeği kısa formunun Türkçe uyarlaması: Geçerlik ve güvenirlik çalışması. *Turkish Psychological Counseling and Guidance Journal*, *4*(36), 165-172. Retrieved from http://www.turkpdrdergisi.com/index.php/pdr/article/view/100

Doğan, T., & Sapmaz, F. (2012). Oxford mutluluk ölçeği Türkçe formunun psikometrik özelliklerinin üniversite öğrencilerinde incelenmesi. *Düşünen Adam Psikiyatri ve Nörolojik Bilimler Dergisi*, *25*, 297-304. doi: 10.5350/DAJPN2012250401

Doğaner, S. (2017). *Düzenli egzersiz programının bireylerin stres, mutluluk ve serbest zaman doyum düzeylerine etkisi* (Doctoral dissertation). Ankara University, Sağlık Bilimleri Enstitüsü, Ankara. Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi

Doğaner, S., & Balcı, V. (2018). Effect of regular physical activity on individuals' stress, happiness and leisure satisfaction levels. *Spormetre, 16*(3), 132-148. doi: 10.1501/Sporm_0000000382

Düzen, A. Ç., & Özçelik, İ. Y. (2019, Kasım). *Investigation of psychological resilience and happiness levels of active and non-active high schoolers*. Congress Papers of the Association of Sports Sciences, Antalya. Retrieved from http://web.a.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=2&sid=5ab81a1d-5df4-47d9-866a-74eb94150602%40sessionmgr4007

Ekinci, N., & Hamarta, E. (2020). Meslek yüksekokulu öğrencilerinin azim ile mutluluk düzeylerinin incelenmesi. *OPUS Uluslararası Toplum Araştırmaları Dergisi*, *15*(21), 125-141. doi: 10.26466/opus.569805

Ergin, R., Ermiş, E., Erilli, A., & Ağaoğlu, S. A. (2019, Kasım). *Badminton sporcularinin mutluluk ve sosyalleşme düzeylerinin incelenmesi*. Congress Papers of the Association of Sports Sciences, Antalya. Retrieved from http://eds.b.ebscohost.com/eds/pdfviewer/pdfviewer?vid=1&sid=347184b0-a666-4a52-813a-9c4ca9385dfd%40sessionmgr103

Ezer, H. İ. (2017). *Ergenlik döneminde yaygın kullanılan savunma mekanizmaları ile psikolojik sağlamlık ve mutluluk düzeyi arasındaki ilişki: Hatay il merkezi örneği* (Master's thesis). Cag University, Sosyal Bilimler Enstitüsü, Mersin. Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi

Francis, L. J., Ok, Ü., & Robbins, M. (2017). Religion and happiness: A study among university students in Turkey. *Journal of Religion and Health, 56*(4), 1335-1347. doi: 10.1007/s10943-016-0189-8

Gönen, M., (2019). *Antrenör-sporcu ilişkisinin sporcuların durumluk kaygı, öfke ve öznel iyi oluş düzeylerine etkisi: taekwondo ve korumalı futbol örneği*. (Doctoral dissertation). Gazi University, Sağlık Bilimleri Enstitüsü, Ankara. Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi

Güdü-Demirbulat, Ö., Saatçi, G., & Avcıkurt, C. (2019). İşkoliklik ve bireysel mutluluk arasındaki ilişkinin turizm akademisyenleri açısından incelenmesi. *MANAS Sosyal Araştırmalar Dergisi, 8*(4), 3500-3315. doi: 10.33206/mjss.460701

Güllüce, A. Ç., Kaygın, E., & Borekci, N. E. (2019). Üniversite öğrencilerinin nomofobi düzeyi ile öznel iyi olma durumları arasındaki ilişkinin belirlenmesi: Ardahan örneği. *Hacettepe Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, *37*(4), 651-673. doi: 10.17065/huniibf.482061

Gürdoğan, E. P., & Uslusoy, E. C. (2019). The relationship between quality of work life and happiness in nurses: A sample from Turkey. *International Journal of Caring Sciences*, *12*(3), 1364-1371. Retrieved from http://www.internationaljournalofcaringsciences.org/docs/7_gurdogan_original_12_3[20352].pdf

Işık, Z., Çetinkaya, N., & Işık, M. F. (2017). Mutluluğun iş tatmini üzerindeki rolü: Erzurum ili Palandöken Kış Turizm Merkezinde yer alan konaklama işletmelerindeki kadın çalışanlar üzerine bir uygulama. *Journal of Graduate School of Social Sciences*, *21*(2), 457-471. Retrieved from https://dergipark.org.tr/en/download/article-file/474092

İhtiyaroğlu, N. (2018). Analyzing the Relationship between happiness, teachers' level of satisfaction with life and classroom management profiles. *Universal Journal of Educational Research*, *6*(10), 2227-2237. Retrieved from http://files.eric.ed.gov/fulltext/EJ1192735.pdf

İlhan, M., & Güler, N. (2017). Likert tipi ölçeklerde olumsuz madde ve kategori sayısı sorunu: Rasch modeli ile bir inceleme. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *8*(3), 321-343. doi: 10.21031/epod.321057

İsmail, M. (2018). *Mutluluğun yordayıcısı olarak cinsel benlik şeması ve sosyal görünüş kaygısı* (Master's thesis). İstanbul Ticaret University, Sosyal Bilimler Enstitüsü, İstanbul. Retrieved from http://tez2.yok.gov.tr/

Kanten, P., Kanten, S., & Dündar, G. (2016). Ücret tatmininin ve işin özelliklerinin işe gömülmüşlük üzerindeki etkisinde mutluluğun rolü. *İşletme Araştırmaları Dergisi, 8*(3) 64-88.

Karahan, G. (2018). *Örgütsel iletişimde işe ilişkin duyuşsal iyilik algısı ile öznel mutluluk arasındaki ilişkinin incelenmesi: Türkiye'de havayolu şirketlerinde kabin memurları üzerine bir alan araştırması* (Doctoral dissertation). Maltepe University, İstanbul. Retrieved from http://tez2.yok.gov.tr/

Kavaklı, M., Kozan, H. İ. Ö., Kesici, Ş., & Ak, M. (2019). How can we feel happy? The examination of relationships among happiness, mindfulness and forgiveness. *Research on Education and Psychology*, *3*(2), 198-208. Retrieved from https://www.researchgate.net/publication/338103622_How_Can_We_Feel_Happy_The_Examination_of_Relationships_Among_Happiness_Mindfulness_and_Forgiveness

Keldal, G. (2015). Warwick-Edinburgh mental iyi oluş ölçeği'nin Türkçe formu: Geçerlik ve güvenirlik çalışması. *The Journal of Happiness & Well-Being*, *3*(1), 103-115. Retrieved from https://toad.halileksi.net/sites/default/files/pdf/warwick-edinburgh-mental-iyi-olus-olcegi-toad.pdf

Keleş, A. (2019). *Din görevlilerinde dindarlık ve mutluluk arasındaki ilişki (Sivas örneği)* (Master's thesis). Sivas Cumhuriyet University, Sosyal Bilimler Enstitüsü, Sivas. Retrieved from http://tez2.yok.gov.tr/

Kırık, A. M., & Sönmez, M. (2017). İletişim ve mutluluk ilişkisinin incelenmesi. *İnif e-Dergi*, *2*(1), 15-26. Retrieved from https://dergipark.org.tr/en/pub/inifedergi/issue/27641/292862

Konan, N., & Taşdemir, A. (2019). Öğretmenlerin örgütsel ikiyüzlülük algıları ile mutluluk düzeyleri algıları arasındaki ilişki. *Scientific Educational Studies*, *3*(2), 132-152. doi: 10.31798/ses.655939

Köse, A., & Çevik, A. (2018). Happiness as a predictor of attitude towards teaching profession: Pedagogical formation case. *Bartın Üniversitesi Eğitim Fakültesi Dergisi*, *7*(3), 853-873. doi: 10.14686/buefad.393207

Meral, B. F. (2014). Kişisel iyi oluş indeksi-yetişkin Türkçe formunun psikometrik özellikleri. *The Journal of Happiness and Well-Being*, *2*(2), 119-131. Retrieved from https://www.journalofhappiness.net/frontend/articles/pdf/v02i02/3.pdf

Namdar, A. (2018). *Bir grup öğrencide umut, kaygı ve mutluluk arasındaki ilişki* (Master's thesis). Uskudar University, Sosyal Bilimler Enstitüsü, İstanbul. Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi

Okur, S., & Totan, T. (2019). Psikolojik iyi oluşu değerlendiren Bradburn duygulanım dengesi ölçeğinin Türkçede incelenmesi. *Adnan Menderes Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *6*(2), 1-12. Retrieved from https://dergipark.org.tr/en/download/article-file/935147

Öner, C. (2019). *İş yaşamında mutluluk, iş yaşam dengesi ve işe adanma ilişkisi* (Master's thesis). Bahcesehir University, Sosyal Bilimler Enstitüsü, İstanbul. Retrieved from http://tez2.yok.gov.tr/

Öz, T. (2019). *Çiftlerde depresyon, mutluluk ve psikolojik iyi oluş ile evlilik doyumu arası ilişkilerde evlilik süresinin aracı rolünün incelenmesi* (Master's thesis). Fatih Sultan Mehmet Vakif University, Lisansüstü Eğitim Enstitüsü, İstanbul. Retrieved from http://tez2.yok.gov.tr/

Özberk, F. (2018). *Annelerin eğitimleri çalışıp çalışmama durumuna göre lise öğrencilerinin duygusal zekâ mutluluk ve sosyal kaygı düzeylerinin karşılaştırılması* (Master's thesis). Istanbul Arel University, Sosyal Bilimler Enstitüsü, İstanbul. Retrieved from http://tez2.yok.gov.tr/

Özçakır, A., Doğan, F. O., Çakır, Y. T., Bayram, N., & Bilgel, N. (2014). Subjective well-being among primary health care patients. *PloS one*, *9*(12), 1-15. doi: 10.1371/journal.pone.0114496

Özdemir, Ş. (2019). *Büyük ölçekli fitness merkezi kullanıcılarının tatmininin, ilgilenim ve mutluluk düzeylerine etkisinin belirlenmesi* (Master's thesis). Istanbul Okan University, Sosyal Bilimler Enstitüsü, İstanbul. Retrieved from http://tez2.yok.gov.tr/

Özgün, A., Yaşartürk, F., Ayhan, B., & Bozkuş, T. (2017). Hentbolcuların spora özgü başarı motivasyonu ve mutluluk düzeyleri arasındaki ilişkinin incelenmesi. *Uluslararası Kültürel ve Sosyal Araştırmalar Dergisi (UKSAD), 3*(Special Issue 2), 83-94. Retrieved from https://dergipark.org.tr/en/pub/intjcss/issue/33182/369309

Özkan, A. (2019). *Üniversite öğrencilerinin sosyal uyum düzeyleri ile anlamlı yaşam çabası ve mutluluk arasındaki ilişkinin incelenmesi.* (Master's thesis). Ufuk University, Sosyal Bilimler Enstitüsü, Ankara. Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi

Öztaş, İ. (2018). *Farklı kurumlarda çalışan memurların serbest zaman doyum ve mutluluk düzeylerinin belirlenmesi* (Kırıkkale ili örneği) (Master's thesis). Agri Ibrahim Cecen University, Sosyal Bilimler Enstitüsü, Ağrı. Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi

Öztemel, K., & Yıldız-Akyol, E. (2019). The predictive role of happiness, social support, and future time orientation in career adaptability. *Journal of Career Development, 20*(1), 1-14. doi: 10.1177/0894845319840437

Öztürk, L., Meral, İ. G., & Yılmaz, S. S. (2017). Lisans öğrencilerinin mutluluk ve dindarlık ilişkisi: Kırıkkale Üniversitesi örneği. *Akademik Yaklaşımlar Dergisi*, *8*(1), 23-39. Retrieved from https://dergipark.org.tr/tr/download/article-file/316231

Peker, A., Eroğlu, Y., & Özcan, N. (2015). Özel gereksinimli çocuğa sahip anneler ile tipik gelişim gösteren çocuğa sahip annelerin psikolojik sağlamlık, iyilik hali ve mutluluk düzeylerinin incelenmesi. *Sakarya University Journal of Education, 5*(3), 142-150. doi: Retrieved from https://dergipark.org.tr/tr/download/article-file/192381

Sapmaz, F., & Doğan, T. (2012). Mutluluk ve yaşam doyumunun yordayıcısı olarak iyimserlik. *Mersin Üniversitesi Eğitim Fakültesi Dergisi,* 8(3), 63-69. Retrieved from https://www.academia.edu/5249105/Mutluluk_ve_Ya%C5%9Fam_Doyumunun_Yorday%C4%B1c%C4%B1s%C4%B1_Olarak_%C4%B0yimserlik_

Sarıçam, H. (2014). Belirsizliğe tahammülsüzlüğün mutluluğa etkisi. *Sosyal Bilimler Dergisi, 4*(8), 1-12. Retrieved from https://dergipark.org.tr/en/download/article-file/717400

Sarıçam, H. (2016). Examining the relationship between self-rumination and happiness: The mediating and moderating role of subjective vitality. *Universitas Psychologica, 15*(2), 383-396. doi: 10.11144/Javeriana.upsy15-2.errh

Sariçam, H. (2015). Metacognition and happiness: The mediating role of perceived stress. *Studia Psychologica, 57*(4), 271-283. Retrieved from https://www.researchgate.net/profile/Hakan_Saricam2/publication/286449970_

Sezer, Ö., & Sumbas, E. (2018). Lise öğrencilerinin yaz tatillerini kullanım biçimlerinin bazı değişkenler açısında incelenmesi. *İnönü University International Journal of Social Sciences (INIJOSS), 7*(1), 173-189. Retrieved from http://dergipark.gov.tr/download/article-file/503986

Taşdibi-Ünlü, F. (2019). *Üniversite öğrencilerinde yaşam değeri, yaşamın anlamı ve sosyal iyi olmanın mutluluğu yordamadaki rolü* (Master's thesis). Mugla Sitki Kocman University, Eğitim Bilimleri Enstitüsü, Muğla. Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi

Tatar, A. (2020). Madde yanıt kuramıyla A İş Stresi Ölçeği-20'nin geliştirilmesi: Güvenilirlik ve geçerlilik çalışması. *Anadolu Psikiyatri Dergisi, 21*(Ek Sayı 1), 14-22. doi: 10.5455/apd.77173

Teke, M. (2020). *Yoksul kadınlarda mutluluk anlayışı* (Master's thesis). Sivas Cumhuriyet University, Eğitim Bilimleri Enstitüsü, Sivas. Retrieved from http://tez2.yok.gov.tr/

Totan, T., Ercan, B., & Öztürk, E. (2019). Mutluluk ve benlik saygısının yalnızlıkla internet bağımlılığına etkilerinin incelenmesi. *EDU7, 8*(10), 20-35. Retrieved from https://dergipark.org.tr/en/download/article-file/913720

Traş, Z., Öztemel, K., & Koçak, M. (2019). Öğretmen adaylarının mutluluk düzeylerinin bazı öznel niteliklerine göre incelenmesi. *Necmettin Erbakan Üniversitesi Ereğli Eğitim Fakültesi Dergisi, 1*(1), 47-56. Retrieved from https://dergipark.org.tr/en/pub/neueefd/issue/45274/547448

Traş, Z., Öztemel, K., & Koçak, M. (2020). Üniversite öğrencilerinin mutluluk, yalnızlık ve sabır düzeyleri arasındaki ilişkinin incelenmesi. *OPUS Uluslararası Toplum Araştırmaları Dergisi*, *15*(22), 878-894. doi: 10.26466/opus.575329

Türk, R., Akkuş, Y., & Sönmez, T. (2019). Relationship between self-care ability and happiness in elderly individuals. *Cukurova Medical Journal*, *44*(Suppl 1), 366-374. doi: 10.17826/cumj.560455

Uz-Baş, A., & Soylu, Y. (2018). Pozitif duyguların psikolojik danışman adaylarının entelektüel becerileri ve mutluluk düzeylerine etkisi. *Journal of Higher Education & Science/Yüksekögretim ve Bilim Dergisi*, *8*(2), 264-270. doi: 10.5961/jhes.2018.269

Yağmur, Y., Oltuluoğlu, H., & Ergin, İ. O. (2019). İntrauterin dönemde fetal cinsiyetin annelerin mutluluk düzeyine etkisi. *ACU Sağlık Bil Derg, 10*(1), 89-93. doi: 10.31067/0.2018.98

Yalçın, R. Ü., & Gençdoğan, B. (2019). Mutluluk saldırganlığın yordayıcısı mıdır? Üniversite öğrencileri ile bir yapısal eşitlik modeli çalışması. *Cumhuriyet International Journal of Education, 8*(3), 593-608. Retrieved from https://dergipark.org.tr/tr/pub/cije/issue/48894/510491

Yaprak, P., Güçlü, M., & Ayyıldız Durhan, T. (2018). The happiness, hardiness, and humor styles of students with a bachelor's degree in sport sciences. *Behavioral Sciences, 8*(9), 1-21. doi: 10.3390/bs8090082

Yavaşoğlu, E. (2019). *Evli bireylerin mutluluk düzeyleri ile özgünlük ve değerler arasındaki yordayıcı ilişkiler* (Master's thesis). Istanbul Sabahattin Zaim University, Sosyal Bilimler Enstitüsü, İstanbul. Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi

Yazgeç, G. (2019). *Doğa ve macera rekreasyonu etkinliklere katılan bireylerin serbest zaman doyum ve mutluluk düzeylerinin incelenmesi: Fethiye destinasyonu örneği* (Master's thesis). Manisa Celal Bayar University, Sosyal Bilimler Enstitüsü, Manisa. Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi

Yeşiltaş, A., Şahin, S., & Serezli, G. (2019). Çalışan mutluluğunun ve işe bağlılığın örgüt performansına etkisi. *Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 11*(30), 664-673. doi: 10.20875/makusobed.555411

Yıldırım, O. (2019). *Ergenlerde akıllı telefondan yoksun kalma korkusu (Nomofobi) ile sosyodemografik değişkenler, temel psikolojik ihtiyaçlar, sürekli kaygı ve mutluluk arasındaki ilişkinin incelenmesi* (Master's thesis). Inonu University, Eğitim Bilimleri Enstitüsü, Malatya. Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi

Yıldız, Y., & Ekici, S. (2017). Sporun üniversite öğrencileri üzerinde mutluluk ve sosylalleşme düzeylerine etkisinin incelenmesi. *Niğde Üniversitesi Beden Eğitimi ve Spor Bilimleri Dergisi, 11*(2), 181-187. Retrieved from https://www.researchgate.net/profile/Yasin_Yildiz6/publication/320755977_Investigation_Of_The_Effect_Of_Sports_On_The_Level_Of_Happiness_And_Socialization_Of_University_Students/links/5a6e1f75458515d407584ce6/Investigation-Of-The-Effect-Of-Sports-On-The-Level-Of-Happiness-And-Socialization-Of-University-Students.pdf

Yılmaz, E. (2018). *Davranışsal ekonomide gelir ve mutluluk ilişkisi ve bir uygulama: Samsun ili örneği* (Master's thesis). KTO Karatay University, Sosyal Bilimler Enstitüsü, Konya. Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi

Yurcu, G., & Atay, H. (2015). Çalışanların öznel iyi oluşunu etkileyen demografik faktörlerin incelenmesi: Antalya ili konaklama işletmeleri örneği. *Manas Sosyal Araştırmalar Dergisi, 4*(2), 17-34. Retrieved from https://dergipark.org.tr/en/pub/mjss/issue/40495/485114

Yüksekbilgili, Z., & Akduman, G. (2016). Bireysel mutluluk ve işkoliklik ilişkisi. *Kocaeli Üniversitesi Sosyal Bilimler Dergisi,* (31), 95-112. Retrieved from https://dergipark.org.tr/tr/pub/kosbed/issue/25688/271103

Zorba, E., Pala, A., & Göksel, A. G. (2016). Examining the relation between emotional ıntelligence and happiness status of wellness trainers. *Journal of Education and Learning, 5*(3), 159-165. doi: 10.5539/jel.v5n3p159

Oxford Mutluluk Ölçeğinin Meta-Analizle Türk Örnekleminde Güvenirlik Genellemesi

Giriş

Ölçme ve değerlendirmenin eğitim ve psikolojideki yeri ve önemi oldukça büyüktür. Buna göre, bu disiplinler arası ilişki ve dinamiğin sağlanması önemlidir. Ölçme ve değerlendirmenin en önemli iki kavramı ise *güvenirlik* ve *geçerlik*tir. Ölçme ve değerlendirmenin temel kuramlarından birisi olan klasik test kuramına göre güvenirlik, ölçüm puanlarının tesadüfi hatalardan arınmışlık derecesi olarak tanımlanmaktadır. Ayrıca, güvenirlik, aynı veya paralel testi alan bireylerin aldığı puanlar arasındaki tutarlılık anlamına da gelmektedir (Anastasi, 1982). Güvenirliği yorumlayabilmek için geliştirilen formüller bulunmaktadır. Bu noktada, *güvenirlik indeksi* ve *güvenirlik katsayısı* kavramları arasındaki farkı belirtmek yararlı olacaktır. Güvenirlik indeksi; gözlenen puanlar ile gerçek puanlar arasındaki ilişkiye odaklanırken, güvenirlik katsayısı; paralel formlardan alınan puanlar arasındaki ilişkiyi ifade eder. İki kavram arasındaki matematiksel ilişkiye dayanarak, güvenirlik katsayısının gerçek puan varyansının gözlenen puan varyansına oranı olduğu söylenebilir (Crocker & Algina, 2008). Güvenirlik katsayısının hesaplanmasında araştırmacılar tarafından önerilen farklı formüller bulunmaktadır. Bu katsayılardan bazıları tek bir testin uygulamasını gerektirirken, bazıları birden fazla testin uygulamasını gerektirmektedir. Tek bir test uygulaması ile o teste ilişkin iç tutarlılık anlamındaki güvenirliğin elde edilmesinde ve yorumlanmasında, alan yazında en sık kullanılan güvenirlik katsayısı Cronbach Alfa'dır. Diğer güvenirlik katsayıları gibi, Cronbach Alfa katsayısı da aynı ölçme aracının kullanıldığı farklı çalışmalarda, çalışmadan çalışmaya farklılık göstermektedir. Örneğin Oxford Mutluluk Ölçeği'nin kullanıldığı ve örneklemin üniversite öğrencilerinden seçildiği çalışmalardan birinde Cronbach Alfa katsayısı .29 iken (Taşdibi-Ünlü, 2019), başka bir çalışmada .97'dir (İlhan & Güler, 2017). Bir ölçme aracının güvenirlik katsayısı bu örnekte olduğu gibi, farklı çalışmalardaki örneklem özelliklerine, örneklem büyüklüğüne, uygulama koşullarına, uygulama süresine vb. bağlı olarak değişebilmektedir. Güvenirlik katsayılarında belirtilen çalışma özelliklerine bağlı olarak oluşan bu farklılıklar, güvenirliğin genellemesini gerektirmiştir. Meta-analize dayalı olarak yapılan ilk güvenirlik genellemesi çalışması Vacha-Haase (1998) tarafından yapılmıştır. Vacha-Haase'e göre, güvenirlik genellemesi çalışmaları ile farklı ölçüm ve çalışmalardaki güvenirlik katsayılarının kaynakları ve değişkenlik derecesi incelenebilir. Başka bir deyişle, güvenirlik genellemesi çalışması, güvenirlik katsayısının çalışmalar arasında farklılık gösterip göstermediğini inceleme olanağı tanımaktadır.

Alan yazında bu konuda yer alan çalışmalar incelendiğinde, genel etki büyüklüğünün kestirilmesinin yanı sıra güvenirlik katsayısının; test uzunluğu (veya madde sayısı), örneklem büyüklüğü, örneklem türü, cinsiyet, yaş, ırk, güvenirlik katsayısı türü vb. değişkenlere göre farklılık gösterip göstermediği incelenen çalışmalar da bulunmaktadır (Caruso, 2000; Caruso, Witkiewitz, Belcourt-Dittloff, & Gottlieb, 2001; Graham, Diebels, & Barnow, 2011; Hanson, Curry, & Bandalos, 2002; Wallace & Wheeler, 2002). Ülkemizde alanyazın incelendiğinde ise eğitim alanında meta-analiz çalışmalarının olması ile birlikte güvenirlik katsayılarının meta-analiz yöntemi ile genelleştirildiği çalışmalara neredeyse hiç rastlanmamıştır. Bu çalışma kapsamında yabancı literatürdeki çalışmalar incelenerek, bu çalışmalara paralel bir şekilde, Oxford mutluluk ölçeğinin (OMÖ) kısa ve uzun formlarının meta-analiz yöntemi ile güvenirlik genellemesinin incelenmesi amaçlanmıştır. Bu amaçla OMÖ kısa ve uzun formunun kullanıldığı çalışmaların meta-analizi ile Cronbach Alfa katsayılarının genel etki büyüklüğü kestirilmiştir. Bununla birlikte ölçek türü, çalışma alanı ve örneklem türü moderatör değişkenlerinin genel güvenirlik kestirimini nasıl etkilediği araştırılmıştır.

Çalışmada OMÖ'nün seçilmesinin nedeni, pozitif psikoloji alanındaki çalışmaların son yıllarda artması ve mutluluğun pozitif psikoloji alanında en sık araştırılan kavramlardan biri olmasıdır (Compton & Hoffman, 2019). Ayrıca, hem Türk hem de yabancı literatür incelendiğinde, mutluluğun ölçülmesinde OMÖ'nün sıklıkla kullanıldığı görülmektedir (Demir, 2020; Francis & Crea, 2018; Lin, Imani, Griffiths, & Pakpour, 2020; Okur & Totan, 2019; Yıldırım & Sezer, 2020). Vacha-Haase, Henson ve Caruso (2002), güvenirlik genelleme çalışmalarının test uygulayıcılar ve araştırmacılar için önemli bilgi kaynağı olarak katkıda bulunabileceğini ifade etmişlerdir. Tüm bunlara paralel olarak bu çalışmayı ülkemiz alan yazınına ve eğitim ve psikoloji alanına kazandırmanın önemli olduğu düşünülmektedir.

Yöntem

OMÖ kısa ve uzun formunun genel güvenirliğine ilişkin bilgi edinmeyi hedefleyen bu araştırmada meta-analiz Liberati ve diğerleri (2009) tarafından geliştirilen PRISMA yönergelerine uygun olarak yapılmıştır. Bu kapsamda, iki araştırmacı birbirinden bağımsız bir şekilde, Google Akademik, YÖK ulusal tez/tez merkezi, Gazi Üniversitesi Merkez Kütüphanesi ve Aydın Adnan Menderes Üniversitesi Kütüphanesi veri tabanlarında, 2011-2020 yılları arasında yayınlanan, Oxford mutluluk ölçeğinin uzun ya da kısa formunu kullanan çalışmaları taramıştır. Belirtilen veri tabanları "Oxford mutluluk" ve "Oxford happiness" anahtar kelimeleri ile taranarak, toplam 6906 çalışma; başlık ve özetlerine göre incelenmiştir. Ardından çift kodlama yapılan çalışmalar çıkarılmış ve tam metinler incelenmiştir. Daha sonra çalışmaya dahil etme kriterleri belirlenmiştir. Belirlenen ölçütler; i) belirlenen veri tabanlarında yayımlanmış olmak, ii) Cronbach Alfa katsayısının raporlanmış veya hesaplanabilir olması, iii) Örneklem grubunda, örneklem büyüklüğüne, ölçek formu ya da madde sayısına çalışmada yer verilmiş olması, iv) örneklem grubunun Türk bireylerden oluşmuş olması ve v) çalışma dilinin İngilizce ya da Türkçe olması şeklindedir. Son aşamada bu ölçütlere uygunluk kontrol edilerek 94 çalışma ve bu çalışmalarda da 104 Cronbach Alfa katsayısı olduğu tespit edilmiştir. Ancak bazı çalışmalarda madde atıldığı ya da tamamının kullanılmadığı gözlemlenmiş ve orijinal ölçek formlarındaki 7 ve 29 maddeden farklı sayıda madde sayısı belirten çalışmalar araştırmadan çıkarılmıştır. Sonuç olarak 27'si tez 65'sı makale olmak üzere toplam 92 çalışma meta-analize dâhil olmuştur. Ayrıca, birden fazla güvenirlik katsayısı içeren çalışmalar ayrı ayrı kodlandığı için toplam 95 Cronbach Alfa katsayısının meta-analizi gerçekleştirilmiştir. Güvenirlik katsayılarının dağılımını normalleştirmek için meta-analizden önce Alfa katsayılarına Fisher Z dönüşümü uygulanmıştır. Meta-analize dâhil olan çalışmalar farklı alanlarda, farklı yıllarda olduğu ya da farklı örneklemleri içerdiği için etki büyüklüklerinin çalışmadan çalışmaya farklı olabileceği düşünülerek heterojenlik τ^2 ve Q istatistiğinin bir fonksiyonu olan I^2 istatistikleri ile incelenerek Rastgele etki modeli (REM) tercih edilmiştir. REM altında çalışmalar arası varyansın kestiriminde ise DerSimonian-Laird yöntemi kullanılmıştır.

Meta-analize dahil edilme kriterlerine göre seçilen çalışmaların kodlanması aşamasında belirtilen çalışma özellikleri ele alınmıştır: (i) çalışma adı, (ii) yazar(lar)ın adı, (iii) çalışmanın yayınlandığı yıl, (iv) çalışmanın yayın dili, (v) çalışma türü (makale/tez), (vi) ölçek türü (kısa form/uzun form) (vii) güvenirlik katsayısı (viii) güvenirlik türü, (ix) örneklem büyüklüğü, (x) ölçekteki madde sayısı, (xi) çalışma alanı ve (xii) katılımcı özellikleri. Belirtilen özelliklere göre çalışmalar iki araştırmacı tarafından kodlanmıştır ve kodlayıcılar arası uyum yüzdesi %94.80, Krippendorff Alfa katsayısı ise .94 bulunmuştur. Bu katsayı .80'den yüksek bulunduğu için kodlayıcılar arası güvenirliğin yeterli düzeyde olduğu söylenebilir (Krippendorff, 2004). Yayın yanlılığının incelenmesinde huni diyagramı asimetrisi için Egger'in regresyon testi, Begg ve Mazumdar'ın sıra korelasyon testi ve Duval ve Tweedie'nin kırpma ve doldurma testi kullanılmıştır. Ayrıca fail-safe N yönteminden de yararlanılmıştır.

Araştırma kapsamında moderatör değişken olarak ölçek türü (kısa form/uzun form), örneklem türü (öğrenci/öğrenci değil) ve çalışma alanı (sosyal bilimler, psikoloji/sağlık bilimleri ve spor bilimleri) ele alınmıştır. Moderatör değişkenlere göre çalışmalar incelendiğinde, ölçek türü bağlamında 56 çalışma kısa formu, 36 çalışmanın uzun formu kullanmıştır. Örneklem türü bağlamında ise, 50 çalışmanın öğrenci örneklemi için veri topladığı, 42 çalışmanın ise öğrenci olmayan örneklemlerden veri topladığı görülmüştür. Son olarak çalışma alanları incelendiğinde ise sosyal bilimler alanında 63; spor bilimleri alanında 11; psikoloji/sağlık bilimleri kategorisinde ise 18 çalışma olduğu belirlenmiştir. Belirtilen bu moderatör değişkenlerin güvenirlik kestiriminin değişkenliği üzerindeki etkisi, Analog ANOVA ile incelenmiştir.

Sonuç ve Tartışma

Araştırma sonucunda heterojenlik incelendiğinde, $\tau^2_{(\text{DerSimonian-Laird})} = 0.08$, $Q_{\text{total}}(94) = 2639.66$, $p < .01$ ve $I^2 = \%96.44$ olduğu görülmüştür. Q istatistiğinin manidar olması ve I^2'nin %75'ten yüksek olması güvenirlik katsayılarının büyük oranda farklılaştığını ve heterojen dağıldıklarını göstermektedir. Yanlılığı yorumlama amacıyla, huni diyagramı için Egger'in regresyon testinin (sabit = -2.94, $p = .09$) ve Begg ve Mazumdar'ın sıra korelasyon testinin (tau = -0.06, $p = .41$) manidar olmadığı bulunmuştur, bu nedenle huni diyagramının asimetrik olmadığı söylenebilir. Ayrıca, Duval ve Tweedie kırpma ve doldurma testine göre, gözlenen ve gerçek etki büyüklükleri arasında bir farklılık olmadığı görülmüştür. Huni diyagramının asimetrisine ilişkin bu testler sonucunda diyagramın simetrik bulunması yayın yanlılığının olmadığının bir göstergesidir. Ek olarak fail safe N sonuçları incelendiğinde N 4975 ($p < .01$) olarak hesaplanmıştır. Bu sonuç doğrultusunda yayın yanlılığının söz konusu olmadığı söylenebilir. Verilerin heterojen olarak dağılması nedeniyle REM'e dayalı olarak yapılan güvenirlik genellemesine ve moderatör değişkenlerinin güvenirlik katsayısına etkisine bu başlıkta yer verilmiştir.

Araştırma sonucunda tüm çalışmalarda yer alan Cronbach Alfa katsayılarının genel etki büyüklüğü .81 bulunmuştur. Bu katsayının alt ve üst limiti %95 güven aralığında .78-.82 olarak bulunmuştur. Buna dayalı olarak OMÖ ile elde edilen ölçümlerin güvenirliğinin Türk örneklemi için genel itibariyle yeterli düzeyde olduğu söylenebilir. Diğer yandan OMÖ'nün uzun ve kısa formuna ilişkin sonuçlar incelendiğinde, uzun forma ilişkin ortalama α'nın .87 olduğu, kısa forma ilişkin ortalama α'nın ise .76 olduğu görülmüştür. İki ölçek formunun ortalama α'ları arasındaki fark Analog ANOVA ile incelendiğinde farkın $p < .05$ düzeyinde manidar olduğu gözlemlenmiştir. Bu farklılık uzun form lehinedir, buna dayalı olarak test uzunluğu arttıkça güvenirlik katsayısının arttığı söylenebilir ve literatürde de benzer sonuçlara rastlanmıştır (Henson, Kogan, & Vacha-Haase, 2001; Nilsson, Schmidt, & Meek, 2002). Bu sonuçlara dayalı olarak, ortalama α'ların her iki ölçek formunda da kabul edilebilir düzeyde bulunması nedeniyle Türkiye örnekleminde kullanılabilir olduğu düşünülmektedir. Ancak formların ortalama α'ları arasındaki manidar farklılığa dayalı olarak kısa form yerine uzun form kullanımı daha uygun olabilir. Bu durum ölçeğin açıkladığı varyansa, uygulama grubunun özelliklerine, uygulama koşullarına vb. göre değişebilir. Uzun form ve kısa formun genel etki büyüklüğü arasında manidar bir farklılık olması nedeniyle örneklem büyüklüğünün ve çalışma alanının etkisi iki ölçek formu için ayrı ayrı incelenmiştir.

Moderatör değişkenlerden örneklem türü OMÖ kısa formu için incelendiğinde, örneklemi öğrencilerden oluşan ve öğrencilerden oluşmayan çalışmalar için ortalama α değerleri sırasıyla .75 ve .77'dir. İki örneklem türünün ortalama α'ları arasındaki fark Analog ANOVA ile incelendiğinde; sabit etkiler modeline (SEM) göre manidar farklılık ($p < .05$) bulunmuştur. Grup içi heterojenliğin de yüksek olduğu kategorilerde REM'e göre ise manidar farklılık ($p < .05$) olmadığı gözlenmiştir. OMÖ uzun formu için sonuçlar incelendiğinde ise, örneklemi öğrencilerden oluşan ve öğrencilerden oluşmayan çalışmalar için ortalama α değerleri sırasıyla .89 ve .85 bulunmuş ve bu iki örneklem türünün ortalama α'ları arasındaki fark ise her iki modele göre $p < .05$ düzeyinde manidar bulunmuştur. OMÖ uzun formu için örneklem türü, ortalama α'nın değişkenliğini manidar olarak etkilese de örneklem türünün açıkladığı varyans %11.40 olup ortalama α'nın değişkenliğini düşük bir oranda açıkladığı söylenebilir. Bu doğrultuda örneklem türünün güvenirlik katsayısı kestirimi üzerinde etkisi ölçeğin kısa ve uzun formu için farklıdır. Graham ve diğerleri (2011), Thompson ve Cook (2002) ve Wallace ve Weller (2002) de OMÖ kısa formunda olduğu gibi örneklem türünün genel güvenirlik kestirimine bir etkisi olmadığı sonucuna varmışlardır. Buna karşın, OMÖ uzun formundaki sonuçlara benzer şekilde Caruso ve diğerleri (2001), Vacha-Haase (1998) ve Yin ve Fan (2000) örneklem türünün genel güvenirlik kestirimini etkilediği sonucuna varmışlardır. Bu sonuçlara dayalı olarak, OMÖ uzun formunun öğrenci örnekleminde daha güvenilir sonuçlar verdiği görülse de açıkladığı varyans oranı düşüktür, böylece iki ölçek formunun da Türkiye örneklemi için hem öğrenci hem de öğrenci olmayan örneklemlerde kullanılabilir olduğu söylenebilir.

Son olarak moderatör değişkenlerden çalışma alanı OMÖ kısa formu için incelendiğinde, ortalama α'nın sosyal bilimler alanı için .77, psikoloji/sağlık bilimleri alanı için .74, spor bilimleri için .72

olduğu görülmüştür. Bu alanlara ait ortalama α'lar arasındaki fark kısa form için Analog ANOVA ile incelendiğinde SEM'e göre manidar farklılık (p < .05) bulunsa da çalışma alanının gerçek varyansı açıklama oranı %1'den düşüktür. Rastgele etkiler modeli dikkate alındığında p < .05 düzeyinde manidarlık olmadığı gözlemlenmiştir. OMÖ uzun formu için sonuçlar incelendiğinde ise, ortalama α, sosyal bilimler alanı için .87, psikoloji/sağlık bilimleri alanı için .89, spor bilimleri için .86'tir. OMÖ uzun formu için bu alanlara ait ortalama α'lar arasındaki fark ise her iki modele göre p < .05 düzeyinde manidar bulunmamıştır. Bu doğrultuda her iki form türü için de ortalama güvenirlik katsayılarının yeterli düzeyde olması ve alanlar arasında manidar bir farklılık olmaması nedeniyle OMÖ formlarının Türkiye örnekleminde bu alanlar için uygun olduğu söylenebilir.

Sonuç olarak, ortalama α'ların tüm moderatör değişken kategorilerinde kabul edilebilir düzeyde olması nedeniyle OMÖ'nün uygulayıcılar ve araştırmacılar açısından kullanılabilir olduğu belirtilebilir. Bu değişkenlere ait grup farklılıklarının üstesinden gelebilmek için REM ile analiz yapılması önerilebilir. Ölçek formlarından uzun formun ortalama güvenirlik katsayısının kısa forma göre manidar bir şekilde yüksek olması nedeniyle kısa form yerine uzun formun kullanılması ölçümlerin güvenirliği bağlamında tercih edilebilir. Ancak bu durum araştırmanın amacı, ölçeklerin açıkladığı varyans ve uygulama koşullarına bağlı olarak uygulayıcı ve araştırmacılar açısından değişkenlik gösterebilir.

Bu araştırmada OMÖ'nün meta-analitik güvenirlik genellemesi yapılmış ve moderatör değişkenler ölçek türü, örneklem türü ve çalışma alanı olarak belirlenmiştir. Diğer araştırmacılar farklı ölçekler için güvenirlik genellemesi çalışması yürütebilir ya da amaçları doğrultusunda farklı moderatör değişkenleri ele alabilirler. Farklı moderatör değişken olarak çalışma dili, örneklem büyüklüğü, çalışma yılı, ırk, cinsiyet, yaş, ölçekten elde edilen ölçümlerin ortalaması ve standart sapması, güvenirlik türü, araştırma deseni, farklı örneklem türü vb. seçilebilir. Ayrıca bu çalışmada Türkiye örneklemi ele alınmış olup başka araştırmacılar farklı ülkelerin örneklemlerini ya da dünya genelinde yayınlanan ölçeklerle ilgili tüm çalışmaları ele alabilirler.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(4); 405-429

Drawing a Sample with Desired Properties from Population in R Package *"drawsample"*

Kübra ATALAY KABASAKAL * Tuba GÜNDÜZ **

Abstract
The aim of this study is to develop an R package called *drawsample*, which will be used to draw samples with the desired properties from a real data set. In accordance with the aim of the study, a sample with the desired properties can be drawn by purposive sampling with determining several conditions, such as deviation from normality (skewness and kurtosis) and sample size. Different applications of the package *drawsample* are illustrated using real data from the "Science and Technology(Score_1)" and "Social Studies (Score_2)" subtests of 6th Grade Public Boarding and Scholarship Examinations (PBSE). As the importance given to research with real data has increased in recent years, a good approach would be to draw a sample of the population. With this package, it is expected that researchers will draw samples as close as possible to the desired properties from the population or a large sample. It is thought that using the drawn samples obtained from real data with package *drawsample* will provide an alternative to simulation studies as well as a complement for these studies.

Key Words: R package "drawsample", distribution, real data, simulation.

INTRODUCTION

In the field of measurement and evaluation in education and psychology, the distribution of scores has an important role in the description of the groups. In addition to the description of groups, testing for normality to conduct many procedures of statistical inference, which are based on the assumption of normality, is crucial. However, as Erceg-Hurn and Mirosevich (2008) pointed out, the assumption of normality is rarely met when analyzing real data. Therefore, in applications, non-normal distributions are more common than normal distributions (Blanca, Arnau, López-Montiel, Bono, & Bendayan, 2013; Geary, 1947; Micceri, 1989; Olivier & Norberg, 2010; Pearson, 1932). Due to the failure of the normality assumption, violation of normality, and distribution types have been the focus of many researchers working on important issues such as test equating, computer adaptive testing, differential item functioning, classification, and latent score estimation (Custer, Omar, &Pomplun, 2006; Finney & DiStefano, 2006; Gotzmann, 2011; Kieftenbeld & Natesan, 2012; Kirisci, Hsu, & Yu, 2001; Kolen, 1985; Kogar, 2018; Seong, 1990; Uysal, 2014; Yıldırım, 2015).

In the process of collecting data in a study, researchers may obtain different types of distributions. For example, most of the time, mathematics achievement scores differ from a normal distribution (skewed to the right) in selection exams (Ministry of National Education-MoNE, 2020; Student Selection and Placement Center- SSCP, 2019). If a researcher plans to conduct a study to investigate relations to antecedent and subsequent factors with mathematics scores obtained by a selection exam, and the statistical analysis intended to be used requires normality assumption, the researcher would not make use of the data because the results would be suspenseful. Since a sample selected from this data would also be skewed to the right, drawing a sample from this population will not solve the problem either. Otherwise, the scenario may be the opposite. For example, the aim of researchers may be to test the violations of the normality assumption in a psychometric analysis, and the data they collected may show normal distribution.

* Assist. Prof. Dr., Hacettepe University, Department of Educational Sciences, Ankara-Turkey, katalay@hacettepe.edu.tr, ORCID ID: 0000-0002-3580-5568
** Res. Assist., Gazi University, Gazi Faculty of Education, Ankara-Turkey, tubagunduz@gazi.edu.tr, ORCID ID: 0000-0002-0921-9290

To cite this article:
Atalay Kabasakal, K. & Gündüz, T. (2020). Drawing a Sample with Desired Properties from Population in R Package "drawsample". *Journal of Measurement and Evaluation in Education and Psychology, 11*(4), 405-429. doi: 10.21031/epod.790449

Received: 04.09.2020
Accepted: 22.12.2020

In order to conduct studies with different distribution types, generated data are used in simulation studies (Abdel-Fattah, 1994; Bıkmaz-Bilgen & Doğan, 2017; Dolma, 2009; Kaya, Leite, & Miller, 2015; Urry, 1974; Yıldırım Uysal-Saraç, & Büyüköztürk, 2018; Yoes, 1993). There are many software packages used to generate data with different distribution types such as normal, uniform, and skewed distributions. Bahry (2012), using a beta distribution, generated samples with three distribution types (extreme and moderate skewness and a baseline condition) and seven sample sizes (from n = 100 to n = 3,000) by using WinGen 3.1 (Han, 2007). As an alternative to WinGen 3.1, SAS software (SAS Institute, 2009) can also be used to obtain different types of distribution. Gotzmann (2011) simulated normal and negatively skewed population distributions of ability parameters (N = 2,000,000). In his study, the population distributions of thetas were generated using the Normal Distribution function in SAS, and the negatively skewed distributions were created using the RAND Beta Distribution function in SAS (SAS Institute, 2009). Of these data, the ability parameter was determined to be appropriate for the purpose of his study, and random samples of different sample sizes (1,500 and 3,000) were selected. The use of beta distributions makes it easy to simulate skewed score distributions (Han & Hambleton, 2007). The components of beta distribution are parameters α and β. Some researchers draw a sample from the simulation data based on the desired properties. For this purpose, Fleishman’s (1978) power method is suitable to draw a sample with skewed or platykurtic/flat distribution from the original data set (Blanca, Alarcón, Arnau, Bono, & Bendayan, 2017; Kieftenbeld & Natesan, 2012; Sen, Cohen, & Kim, 2014; Stone, 1992).

Simulation methods are flexible and can be applied to a number of problems to obtain quantitative answers to questions that may not be possible to derive (Hallgren, 2013). Although simulation is a powerful technique, it has some limitations, which include difficulty in generalizing the results, organizing the results, and applying the results to real data (Wicklin, 2013). Simulation data provide a perfect fit that cannot be reached in real data. As Hallgren (2013) pointed out, real-world datasets are likely to be more “dirty” than the “clean” datasets that are generated in simulation studies, which are often generated under idealistic conditions which can be referred to as a perfect fit. Sireci (1991) stated that when real test data were not used, it was difficult to know whether the simulated data accurately reflected the characteristics of small sample data encountered in practice, and its validity could not be tested. Therefore, the use of real data has increased the importance of the studies conducted lately. In addition, some prestigious journals such as Educational Measurement: Issues and Practice (EM: IP) and the Journal of Educational and Behavioral Statistics (JEBS) have stated that simulation-based studies are from “examples of inappropriate manuscript topics” or considered to “have low priority,” although most of the articles in these journals so far are simulation studies (American Educational Research Association, 2020; John Wiley & Sons Inc., 2019).

In empirical research, the process of data collection is challenging. The sample may not be representative of the population distribution; alternately, it may not be normally distributed, or it may be unsuitable for the desired distribution. To meet the assumption of normality in the literature, many studies in which the data set was manipulated have been found. For example, Gelbal (1994), in accordance with the purpose of his research, examined test scores, which included approximately two thousand fifth grade students who took both the Turkish language test and Math test. In order to get the desired distributions, approximately five hundred students from each test were removed. Doğan and Tezbaşaran (2003), in their study, selected participants with the required attributes to ensure the desired distribution. The researchers stated that random and purposive sampling techniques were used in the selection of the samples. For the purpose of their study, the students were drawn from a population consisting of students who had taken the Secondary Education Institutions Student Selection and Placement Examination in 2001. The samples were drawn randomly, right-skewed, left-skewed, flattened, and normal distribution, ranging in sample size from 2,353 to 29,244. In their study, in skewed samples, absolute values of skewness (±1.00) and kurtosis (1.37) were kept equal among samples to increase the accuracy for comparisons. Similar to the study of Doğan and Tezbaşaran (2003), Şahin ve Yıldırım (2018), obtaining the ability parameters, both right-skewed and left-skewed ability distributions were chosen from the real data. The real data were obtained from mathematics subtests of the Placement Test (SBS) applied in 2012. The selection of the right-skewed distributions was made randomly because it was originally a right-skewed data set (skewness value=1.05). For the left-skewed data sets, the intended sample distribution was achieved through purposive sampling, and the groups whose skewness value is approximately -1.00 were chosen for all samples.

In addition to the above, in the literature, many researchers have chosen to draw samples from the real data set (population) in accordance with the purpose of their studies (Courville, 2004; Doğan & Kılıç, 2018; Fan, 1998; Nartgün, 2002; Reyhanlıoğlu Keçeoğlu, 2018). In the process of sampling from the population, it is important for future studies to have a function that makes the sample selection easier and brings it closer to the desired properties. In fact, it is suggested that the study of different abilities with non-normal distributions or samples with different levels of ability is the result of some research in the literature (Çelikten & Çakan, 2019). When the studies are examined, it was concluded that there is a need for a tool to enable researchers to draw samples with the desired properties from a large data set.

Purpose of the Study

In this study, the package *drawsample*, which aims to draw a sample based on the information of total score or ability parameter in accordance with the desired sample size and deviation from normality (skewness and kurtosis), was developed.With this package, it is expected that researchers will draw samples as close as possible to the desired properties from the population or a large sample, and it is thought that it will pave the way for the studies to be conducted on different topics based on the distribution in the literature. With this function, it is possible for researchers to draw samples with desired properties from large data in order to conduct statistical analysis under different conditions.

Fleishman’s Power Method

In this section, Fleishman's (1978) power method, which is used to select the desired measures of deviation from normality (skewness and kurtosis), is explained briefly. Fleishman (1978) used a cubic transformation of a standard normal variable to create a distribution with pre-specified moments. Fleishman’s (1978) power method, $Y = a + bz + cz^2 + dz^3$, was used to generate a non-normal distribution, where Y is a non-normal deviate with specified skewness and kurtosis. The value of z is a standard normal deviate, and a, b, c, and d are constants for transforming the standard normal variable to a variable with known skewness and kurtosis. (Kirisci, 2001). These constants for the normal distribution are 0.0, 1.0, 0.0, and 0.0 $(a = c)$ respectively.

Fleishman (1978) tabulated these coefficient values for the selected skewness and kurtosis values. Writing the function in R, the values in this table were used to get the non-normal distributions. The values in this table can also be accessed using the *find_constants()* function in the “SimMultiCorrData” (Fialkowski, 2018) package in R. The *find_constants()* function is a function that calculates Fleishman's third or Headrick's (2002) fifth-order constants, converting a standard normal random variable into a continuous variable with a certain skewness and standardized kurtosis value. When the skewness value of the function is 0 and the standardized kurtosis value is 0, the usage example is given in Table 1.

Table 1. R Code to Find Fleishman's Third Order Constants

```
library(SimMultiCorrData)
find_constants(method = c("Fleishman"), skews = 0, skurts = 0)
## $constants
## c0 c1 c2 c3
##  0  1  0  0
##
## $valid
## [1] "TRUE"
```

Since the use of the function used in the example given in Table 1 extends the operation process, an R object named "constants_table" was created with the values obtained using this function.

Skewness and Kurtosis Statistics

The first four moments of the distribution are mean, variance, skewness, and kurtosis, respectively, which are the most important characteristic of frequency distributions (D'agostino, Belanger, & D'Agostino, 1990).

The following equations are for the third and fourth moments, skewness and kurtosis statistics, in Equation 1 and 2. These equations are used routinely; for example, SAS and SPSS give skewness and kurtosis statistics using them in their descriptive statistics output (D'agostino, Belanger, & D'Agostino 1990).

$$\text{skewness} = \frac{n\sum(X-\bar{X})^3}{(n-1)(n-2)S^3} \qquad (1)$$

$$\text{kurtosis} = \frac{n(n+1)\sum(X-\bar{X})^4}{(n-1)(n-2)(n-3)S^4} - \frac{3(n-1)^2}{(n-2)(n-3)} \qquad (2)$$

There are many R packages to calculate the skewness and kurtosis values. In this study, the *describe()* function in the *psych* package was used to calculate skewness and kurtosis values. Table 2 shows the example of calculating descriptive statistics of the vectors of "normal_dis" and "skew_dis" generated by *rnorm()* and *rbeta()* functions, respectively.

Table 2. R Code to Calculate Descriptive Statistics of a Vector

```
library(psych)
set.seed(41)
normal_dis <- rnorm(1000)
describe(normal_dis)
##    vars    n mean sd median trimmed  mad   min  max range  skew kurtosis   se
## X1    1 1000    0  1   0.03    0.01 1.01 -3.26 3.33  6.58 -0.01    -0.03 0.03
skew_dis <- rbeta(1000,2,5)
describe(skew_dis)
##    vars    n mean   sd median trimmed  mad min  max range skew kurtosis se
## X1    1 1000 0.28 0.16   0.26    0.27 0.17   0 0.83  0.83 0.56    -0.19  0
```

As shown in Table 2, the *describe()* function has 13 different outputs. From the output of this function, the skewness and kurtosis values can be extracted, as shown in Table 3.

Table 3. R Code to Extract Skewness and Kurtosis Values

```
describe(normal_dis)$skew
## [1] -0.006120114
describe(normal_dis)$kurtosis
## [1] -0.03008443
```

Drawing Samples

The most commonly used function for selecting samples in R is the *sample()* function in the base package. This function takes a sample of the specified size from a determined vector using either with or without replacement

In this study, *sample_n()* function which is a function of *dplyr* package (Wickham, François, Henry, & Müller; 2019) is used to select samples. The *sample_n()* function has similar arguments with the *sample()* function in the base package. The *sample()* function works with vectors, while the *sample_n()* function works with data sets. The *sample_n()* function has the “weight” argument instead of the “prob” argument in the *sample()* function. The value of the “weight” argument can be any column in the data set or data frame. In order to demonstrate the use of the *sample_n()* function, “example1” data set consisting of four variables with 100 observations was created. The variables in the data set “example1” are “id,” “gender,” “math_score” and “science_score.” In order to create a new data frame with students who have higher science scores, the “weight” argument was used with the value of this variable (science_score). Table 4 shows the example of using *sample_n()* function.

Table 4. An Example of Using *sample_n()* Function

```
library(dplyr)
set.seed(41)
example1 <- data.frame( id=paste("id",101:200,sep=""),
gender = sample(c("F","M"),replace=TRUE,100),
math_score = sample(0:100,100,replace=TRUE),
science_score =sample(0:100,100,replace=TRUE))
summary(example1)
##       id               gender            math_score    science_score
##  Length:100         Length:100         Min.   : 1.00   Min.   : 0.00
##  Class :character   Class :character   1st Qu.:27.75   1st Qu.:26.75
##  Mode  :character   Mode  :character   Median :50.50   Median :57.00
##                                        Mean   :49.87   Mean   :52.88
##                                        3rd Qu.:71.25   3rd Qu.:76.00
##                                        Max.   :99.00   Max.   :98.00
example2 <- sample_n(example1, 10, weight = science_score)
summary(example2)
##       id               gender            math_score    science_score
##  Length:10          Length:10          Min.   :15.00   Min.   :32.00
##  Class :character   Class :character   1st Qu.:22.75   1st Qu.:43.00
##  Mode  :character   Mode  :character   Median :52.50   Median :72.00
##                                        Mean   :51.40   Mean   :63.30
##                                        3rd Qu.:76.75   3rd Qu.:79.75
##                                        Max.   :91.00   Max.   :89.00
```

In Table 4, “example1” data set was created, and summary information about the data set was printed. While creating the “example2” data set, the students were weighted according to the “science score” variable, and the sampling was selected. When the summary information about “example2” data set is examined, it is seen that the minimum, quartiles, median, and median values of “science_score” are higher than “example1”.

In the *drawsample* package, the *draw_sample()* function has been improved to get a sample with the desired distribution properties and sample size in accordance with skewness and kurtosis. The code belonging to this function is explained below.

R CODE FOR *draw_sample()* FUNCTION

draw_sample() function with 6 arguments was written to draw a sample with the desired properties. The arguments of the function are given in Table 5.

Table 5. The arguments of *draw_sample()* function

Argument	*Value*
dist	data frame: consists of id and scores with no missing
n	numeric: desired sample size
skew	numeric: the skewness value
kurts	numeric: the kurtosis value
replacement	logical: sample with or without replacement? (default is FALSE).
output_name	character: a vector of two components. The first component is the name of the output file, user can change the second component.

When determining “skew” and “kurts” from the arguments in Table 5, the Fleishman Power Method Weights table must be consulted. Fleishman coefficients corresponding to some combinations, such as skewness value 1 and kurtosis value 0, are absent. The minimum and maximum values of the kurtosis coefficient corresponding to a determined skewness coefficient are presented in this table created by using the Flesihman’s (1978) Power Method Weights Table. For example, if the skewness coefficient is selected as 2, the kurtosis coefficient must be entered between 5 and 20. In other words, the minimum and maximum value of kurtosis values corresponding to each skewness coefficient that can be used are presented in Table 6.

Table 6. Minimum and Maximum Kurtosis Coefficient Corresponding to the Skewness Coefficient

Skewness	*Kurtosis (min)*	*Kurtosis (max)*	*Skewness*	*Kurtosis (min)*	*Kurtosis (max)*
0	-1.2	20	1.9	4.4	20
0.1	-1.2	20	2	5	20
0.2	-1.1	20	2.1	5.6	20
0.3	-1.1	20	2.2	6.3	20
0.4	-0.9	20	2.3	7.1	20
0.5	-0.8	20	2.4	7.8	20
0.6	-0.6	20	2.5	8.6	20
0.7	-0.4	20	2.6	9.5	20
0.8	-0.2	20	2.7	10.4	20
0.9	0.1	20	2.8	11.4	20
1	0.4	20	2.9	12.4	20
1.1	0.7	20	3	13.4	20
1.2	1	20	3.1	14.4	20
1.3	1.4	20	3.2	15.5	20
1.4	1.8	20	3.3	16.5	20
1.5	2.3	20	3.4	17.6	20
1.6	2.7	20	3.5	18.8	20
1.7	3.2	20	3.6	19.9	20
1.8	3.8	20			

R commands for *draw_sample()* function are given in Table 7. In this function, the value of the “dist” argument must be a data frame that has two columns. Note that the data includes student IDs in the first column and student total test scores or abilities (thetas) in the second column. For that purpose, with the command of names(dist), the columns of the imported object columns in the R environment are named “id” and “x” (Table 6, Line 8). Then, the x is extracted as the variable x" in Line 10, so "x" becomes a vector that can provide convenience. If "n" from the arguments of the function, the desired sample size, is larger than the length of the data, it gives the following error: “Cannot take a sample larger than the length of the data”. For example, although the sample size of the imported data is 1,000 and users desire to take sample size 2,000, the function gives the error and stops running (Lines 13 to 16).

The values in Fleishman's (1978) table are used to get a sample with the desired distribution properties. These values are found in the object called "constants_table" in the package. *SimMultiCorrData* (Fialkowski, 2018) package was used in the preparation of the table including these constants. In this object, there are b, c, and d constants belonging to a set of 5,292 lines consisting of kurtosis values corresponding to each skewness value and consisting of skewness values increasing by 0.1 units from 0 to 3.6. This table includes only 0 and positive skewness values and corresponding kurtosis and constants. Between lines 18-22, if the user enters the skew as a negative value, "Skew" and "c" columns of the table by multiplying -1 will be rearranged with *if* statement. If the skewness value given by the user is not included in the table or if the kurtosis value corresponding to that value is not found in the table even if the skewness value is found in the table, the function stops working with an error "No valid power method constants could be found for the specified values. Change the values" (Line 25 to 32). It is suggested to use Fleishman's (1978) table when making choices with regard to skewness and kurtosis values. Within the *repeat* loop, the reference distribution with the skewness and kurtosis values entered by the user between Line 38 and Line 53 is formed. According to the minimum and maximum values of the distribution formed in this loop and then included in the user's data set (Line 65), the rescaled "reference_v4" distribution forms the basis for the function's work. Before the *repeat* loop, an empty vector was created to form a distribution with the skewness and kurtosis values entered by the user. Firstly, an object with a normal distribution called "reference" with a mean of 0 and a standard deviation of 1 is formed in the loop (Line 41). Within the *repeat* loop, the "reference_v2" object is formed by multiplying the "reference" object by the b, c, and d coefficients in the table, respectively. When the skewness and kurtosis values of the "reference_v2" object are equal to the skewness and kurtosis values entered by the user, the loop is stopped, and the "reference_v2" object is assigned to "reference_v3" (line 50). If the calculated values are not equal to the values defined by the user, the "reference_v3" object is left empty, and the loop is repeated. With the *draw_sample()* function, it is aimed to form a similar distribution from the values in the user's data set based upon the "reference_v4" object formed in accordance with the values entered by the user. On lines 67-69, the outputs of the *hist (reference_v4)* function are used for this purpose. The starting and ending points of each bar of the histogram are assigned to "x_break" objects, the number of bars in the histogram to "n_break" objects, and the number of elements in each bar to "x_counts" objects.

The vector "x" is categorized by "x_break" and identified as "x_v1". The categorized object is added as a new column to the user's data set. The information about how many individuals are in each category is assigned to the "x_n" object. The specified operations are defined between 71-73. The information on how many individuals there are in each category is crucial in terms of determining whether the function will select the sample of the user's desired properties without resampling. When the number of individuals in each category in the data set is higher than in each category of the reference distribution, the function can be performed without resampling, with the default value of the *"*replacement" argument. This situation is checked between lines 73 and 79. If the number of the individuals in at least one category in the data set is less than the number of the individuals in the relevant category of the reference distribution, the function gives an error: "Cannot take a sample form that data without replacement. Please change replacement = TRUE." In this situation, the function can be used by changing the value of the *"*replacement" *argument*. The codes working up to line 83 have been written in order to prepare for drawing sample. The drawing sample process is carried out through the *for* loop between 89-105. For data manipulation in the loop, *filter()* and *sample_n()* functions in the package of *dplyr* (Wickham, François, Henry, & Müller; 2019) are used. The scores belonging to the individuals to be formed in the *for* loop were created in the "new_sample" and the empty matrices named "ID_list" for the identity information of the individuals on lines 83 and line 84. In both matrices formed, the number of lines was determined as the number of categories ("n_break") and the number of columns as the maximum number of individuals in these categories.

Table 7. R Commands for *draw_sample()* Function

```
draw_sample <-  function(dist,n,skew,kurts,
                         replacement =FALSE,
                         output_name = c("sample","default")){

  # rename the data
  skew <- round(skew,1)
  kurts <- round(kurts,1)
  names(dist) <- c("id","x")
  # extract x column
  x <- dist$x


  N <- length(x)
  if(n >= length(x)){
    stop("Cannot take a sample larger than the length of the data")
  }

  # arrange table for negative skewness
  if(skew<0){
    constants_table$c <- -1*constants_table$c
    constants_table$Skew <- -1*constants_table$Skew
  }


  if(skew %in% constants_table$Skew == FALSE){
    stop("No valid power method constants could be found for
          the specified values. Change the values")
  }else if (skew %in% constants_table$Skew == TRUE &
  kurts %in%  constants_table[constants_table$Skew==skew,]$Kurtosis == FALSE){
    stop("No valid power method constants could be found for
           the specified values.Change the values")
  }

  reference_v3 <- NULL

  # conduct Fleishman's power method for the specified
  # skewness and standardized kurtosis
  repeat{
    for( i in 1:dim(constants_table)[1]){
      #  random generation for the normal distribution
      reference <- stats::rnorm(n,0,1)
      constants <- constants_table[i,3:5]
      b <- constants$b
      c <- constants$c
      d <- constants$d
      reference_v2 <- -c + b*reference + c*(reference^2) + d*(reference^3)
      skew_value <- round(psych::describe(reference_v2)$skew,1)
      kurt_value <- round(psych::describe(reference_v2)$kurtosis,1)
      if(skew_value == skew & kurt_value == kurts){
        reference_v3 <- reference_v2
        break
      }
    }
    if(is.null(reference_v3) == FALSE)
      break
  }

  # Rescale the reference vector to have specified minimum and maximum
  scale_ref <- function(x, from, to) {
    x <- x - min(x)
    x <- x / max(x)
    x <- x * (to - from)
    x + from
  }
  reference_v4 <- scale_ref(reference_v3, from=min(x),to=max(x))

  x_counts <-  graphics::hist(reference_v4)$counts
  n_break <-   length(graphics::hist(reference_v4)$breaks) -1
  x_break <-   graphics::hist(reference_v4)$breaks

  x_v1 <- as.numeric(cut(x,x_break,include.lowest = TRUE))
  dist2 <- data.frame(dist,x_v1)
  x_n <- unname(table(x_v1))

  control <- sum(x_n>= x_counts)
  if(control!=length(x_counts)){
    if(replacement==FALSE){
      stop("Cannot take a sample form that data without replacement.
           Please change replacement=TRUE")
    }
  }

  new_sample <-  matrix(NA, nrow = n_break, ncol = max(x_counts))
  ID_list <-  matrix(NA, nrow = n_break, ncol = max(x_counts))

  new_sample_2 <- list()
  ID_list_2 <- list()

  for(i in 1:n_break){
    new_count <- 0
    j <- 0
    while(new_count < x_counts[i]){
      j <- j + 1
      IDx <- dplyr::filter(dist2,x_v1==i)
      IDx <- dplyr::sample_n(IDx,1)
      if(replacement==FALSE){
        dist2 <- dplyr::filter(dist2,id!=IDx$id)
      } else{ dist2 <- dist2}
      new_count <- new_count + 1
      new_sample[i,j] <- IDx$x
      ID_list[i,j]<-   IDx$id
    }
    new_sample_2[[i]] <-  stats::na.omit(new_sample[i,])
    ID_list_2[[i]] <-  stats::na.omit(ID_list[i,])
  }

  new_sample_3 <- unlist(new_sample_2)
  ID_list_3 <- unlist(ID_list_2)

  S1 <- data.frame(id=ID_list_3,x=new_sample_3)

  # Save the output
  if (output_name[2] == "default") {
    wd <- paste(getwd(), "/", sep = "")
  }else {wd <- output_name[2]}
  fileName <- paste( output_name[1], wd,".dat", sep = "")
  utils::capture.output(data.frame(S1), file = fileName)


  # Organize the output
  dist3 <-  dplyr::select(dist2,id,x)
  dist3 <-  dplyr::mutate(dist3,type="population")
  S2   <-   dplyr::mutate(S1,type="sample")
  result <-  rbind(dist3,S2)

  # to capture the graph
  graph <-      lattice::histogram(~x|type,data=result,xlab="Score",
                  nint = n_break,
                 scales = list(x = list(tick.number = 5,relation = "free")))

  lattice::trellis.device(device="png",
           filename=paste( output_name[1], wd,".png", sep = ""))
  print(graph)
  grDevices::dev.off()

  desc <-  rbind(psych::describe(x),
                  psych::describe(reference_v4),
                  psych::describe(S1$x))[,c(2:4,8:9,11:12)]
  rownames(desc) <- c("population","reference","sample")
  # output with three components
  output <- list(desc =desc ,
                  sample = tibble::as_tibble(data.frame(S1)),
                  graph = lattice::trellis.last.object()
  )

  return(output)
}
```

With the *while* loop in the *for* loop that repeats as many as the number of categories, the number of individuals required to be included in each category of the reference vector divided into categories continues until the required number of people is reached by selecting one by one from the relevant category of the data set with the *sample_n()* function. The code in the *while* loop works in two different ways depending on the value of the *"*replacement" *argument*. If the value of the argument is FALSE, the person selected for sampling once is not included in the resampling and is removed from the data (line 96 to 98); if the value of the argument is TRUE, this part of the code will not work. Each line of the "new_sample" and "ID_list" objects formed within the *while* loop contains the "NA" value (max (x_counts) - x_counts [i] values), except for the amount of data that should be in each category. In the *for* loop, the missing data of the objects in the matrices formed in the *while* loop are removed and assigned to the empty list objects formed in 86-87 lines. After the loops were completed, lists containing information about each individual were selected for the sampling, and that person was assigned to the vector objects (Lines 107-108). After that, these vectors are combined in the S1 data set to form the desired sample.

Between lines 141 and 144, the output of the function is formed. The output, which is a three-component list, consists of descriptive statistics of the data and sample, the sample formed and the histogram graphs of the data, and the distribution of the sample. Descriptive statistics, which are the first component of the list, were formed between lines 136-139 by using the *describe()* function in the *psych* package (Revelle, 2018), the *graph*, which is the third component, was formed by using the *histogram()* function in the "lattice" package (Sarkar, 2008) between 120-134 lines. The *desc* component consisting of descriptive statistics information is a matrix. This matrix includes the mean, standard deviation, skewness, and kurtosis of the population, sample and the reference distribution. The second component is called *sample*, and it is from the *tibble* package (Wickham, Francois, and Müller; 2016). It is situated between 112-117 lines required to extract this data. It includes ids and x scores which are sampled. The third component is called "graph," and it includes two histogram graphs one is for "population" (imported data), and one is for the "sample" (extracted data). The third component of the output is also extracted.

EXAMPLES WITH REAL DATA

In the examples, related functions and outputs are presented based on the "Science and Technology" and "Social Studies" subtests data of the 6th Grade Public Boarding and Scholarship Examinations (PBSE) in 2013. At the secondary school level, the PBSE test consists of 100 multiple-choice test items, which include 25 items in each subtest (Turkish, Mathematics, Science and Technology, and Social Studies). It was administered in two booklet types, A and B (MoNE, 2013).

In 2013, 242,598 students participated in PBSE at the 6th-grade level, and 121,523 (50.09%) received booklet A. Of the students, 133,866 (55.18%) were female and 108,732 (44.82%) were male students. Within the scope of the study, randomly selected 5,000 students taking booklet A were considered as the "population." Of this group, 2,745 (54.90%) are female students. The data were obtained by the Directorate General for Measurement, Assessment and Examination Services of the Ministry of National Education in accordance with written permission. The total score distributions for each test were examined. Then, two datasets were used for the demonstration. The Science and Technology subtest was chosen as an example of left-skewed distribution. The Social Studies subtest was used as an example of platykurtic distribution. In each example, a sample of 500 students was drawn from the population for the related subtests. That the samples have the desired properties in terms of distribution type was taken into consideration. The functions and outputs for this process were given in Tables 8-14 and Figures 1-3.

In the first two examples, particular importance has been given to draw samples with a normal distribution and both negatively skewed and leptokurtic distribution from the data of the science and technology subtest, respectively. The command for the first example is shown in Table 8.

Table 8. Draw the Sample with a Normal Distribution of 500 Students from the Total Score Distribution of Science and Technology

```
install.packages("drawsample")
library(drawsample)
data(example_data)
Score1_normal <- drawsample::draw_sample(dist= example_data [,c(1,2)],
 n=500, skew = 0, kurts = 0,output_name = c("sample","1"))
```

First, the package *drawsample* is installed and then loaded. After then the object "example_data" which is automatically provided by the package is loaded. It has three columns including the total scores of the PBSE subtests of 5,000 students and IDs. The first column contains IDs (1: 5000), the second column contains the total scores of "Score_1 (Science and Tecnology subtest) ", and the third column contains "Score_2 (Social Studies subtest)" respectively.

In the function *draw_sample()* given in Table 8, the value of the "dist" argument contains the first and the second columns of "example_data [,c(1,2)]". In the output produced by the function in this model, the IDs and total scores of 500 students are recorded in the working directory of "Sample1.dat" extracted by "Sample1.png", which includes histogram graphs of the "population" of 5,000 students and the of 500 students with a distribution close to normal distribution. Table 9 shows the descriptive statistics of the distribution of 500 students drawn from the total score distribution of Science and Technology and the output of some of the students in the sample extracted.

Table 9. Outputs of the Distribution of 500 Students Drawn from the Total Score Distribution of Science and Technology

```
> Score1_normal
$desc
              n  mean   sd min max  skew kurtosis
population 5000 14.61 4.90   0  25 -0.40    -0.35
reference   500 13.21 4.58   0  25  0.04    -0.03
sample      500 13.73 4.59   0  25  0.01    -0.13

$sample
# A tibble: 500 x 2
      id     x
   <dbl> <dbl>
 1   416     2
 2   456     2
 3  3169     0
 4  4918     2
 5  4411     3
 6  4847     4
 7  4752     3
 8  1159     3
 9  4018     4
10  2963     4
# ... with 490 more rows

$graph
```

When the descriptive statistics in Table 9 were examined, the distribution of the total score of the Science and Technology subtest of 5,000 students was left-skewed, and the drawn sample was very close to the normal distribution. Figure 1 shows the "Sample1.png" which includes histogram graphs.

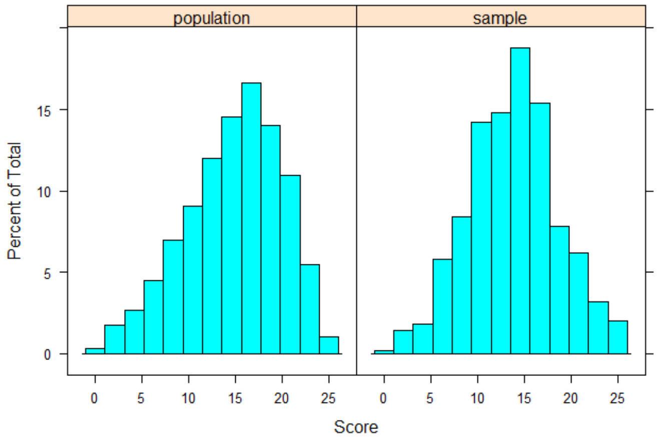


Figure 1. Histograms for the “population” and the “sample” Desired to Have Normal Distribution (Sample1.png)

As seen in Figure 1, the drawn sample distribution given according to the total scores of Science and Technology subtest was very close to the normal distribution. The command for this second example is shown in Table 10. In the second example, different from example 1, the value of skew and kurts are changed to -1 and 2, respectively.

Table 10. A sampling of 500 Students with Left Skewed and Leptokurtic Distribution from the Total Score Distribution of Science and Technology

```
Score1_nskew_lepto<- draw_sample(dist= example_data [,c(1,2)], n=500,skew = -1, kurts =
5,
 output_name = c("sample","2"))
```

Table 11 shows the descriptive statistics of the distribution of 500 students from the total score distribution of Science and Technology and the output of some of the students in the sample drawn.

Table 11. Outputs of the Distribution of 500 Students Drawn From the Total Score Distribution of Science and Technology.

```
Score1_nskew_lepto
## $desc
##                n  mean   sd min max  skew kurtosis
## population 5000 14.61 4.90   0  25 -0.40    -0.35
## reference   500 15.71 2.40   0  25 -0.95     5.03
## sample      500 16.24 2.43   2  25 -0.75     3.48
##
## $sample
## # A tibble: 500 x 2
##       id     x
##    <dbl> <dbl>
##  1  1124     2
##  2   768     8
##  3    82     8
##  4  2748     7
##  5  3196    10
##  6  3049     9
##  7  3456    10
##  8  3319    10
##  9  3942     9
## 10  2558    10
## # ... with 490 more rows
##
## $graph
```

When Table 11 is examined, although the skewness of “sample 2” is almost the same as the skewness of the “reference”, the “sample 2” is flattened than the “reference”. Figure 2 shows “Sample2.png” which includes histogram graphs for this model.

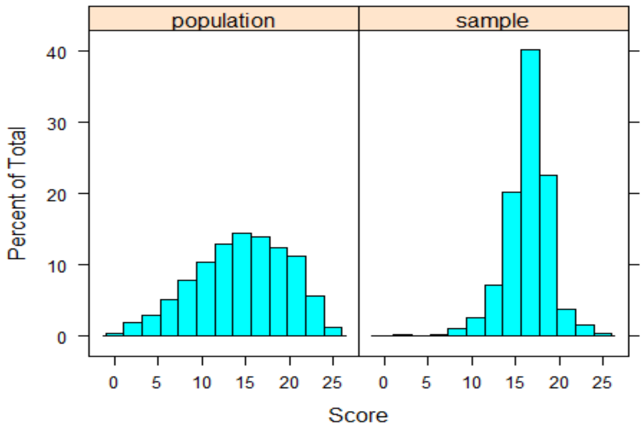


Figure 2. Histograms for the “Population” and the “Sample” Desired with Left Skewed and Leptokurtic Distribution (Sample2.Png)

The next two examples for real data were to draw sample with right-skewed and leptokurtic distribution (skewness value is =1.5 and kurtosis value=3) drawn from the distribution given according to the total scores of Social Sciences subtest. The command required for this situation is presented in Table 12.

Table 12. A sampling of 500 Students with Right Skewed and Leptokurtic Distribution Drawn from the Sum Score Distribution of Social Sciences

```
Score2_pskew_lepto<- draw_sample(dist= example_data [,c(1,3)], n=500,skew = 1.5, kurts =
3, output_name = c("sample","3"))
```

When the code in Table 12 is set to work, since the function cannot draw the data with the desired properties from the provided data without resampling, it gives an error and suggests allowing resampling. The argument "replacement," which is FALSE by default, has been replaced to meet the distribution conditions set out in Table 13.

Table 13. Sampling with Replacement of 500 Students with Right Skewed and Leptokurtic Distribution Drawn from the Sum Score Distribution of Social Sciences

```
Score2_pskew_lepto<- drawsample::draw_sample(dist= example_data [,c(1,3)], n=500,skew =
1.5, kurts = 3,replacement = TRUE,
 output_name = c("sample","3"))
```

Resampling is allowed when "TRUE" is entered in the "replacement" argument. In other words, an individual selected from "population" to "sample" is allowed to be repeatedly selected to provide the desired distribution.

Table 14 shows the descriptive statistics of the distribution of 500 students drawn from the total score distribution of Social Sciences and the output of some of the students in the sample extracted. In this case, the dist data frame contains the columns "ID" and "Score_2", which are used for defining the student identity and total score of the Social Studies subtest.

Table 14. Outputs of the Distribution of 500 Students Drawn from the Total Score Distribution of Social Sciences

```
> Score2_pskew_lepto
$desc
              n  mean   sd min max  skew kurtosis
population 5000 12.78 5.22   0  25 -0.17    -0.88
reference   500  6.79 3.72   0  25  1.49     2.96
sample      500  7.32 3.78   0  25  1.39     2.75

$sample
# A tibble: 500 x 2
      id     x
   <dbl> <dbl>
 1  3756     1
 2   390     1
 3  1483     1
 4  2855     2
 5  4792     1
 6  3660     2
 7  3937     1
 8  4280     2
 9   930     0
10  4324     2
# ... with 490 more rows

$graph
```

When the descriptive statistics in Table 14 were examined, the total score distribution of the Social Sciences subtest of 5,000 students (population) is slightly left-skewed; the desired sample is right-skewed. On the other hand, the "population" s distribution shape is flatty, but the "sample" distribution shape is leptokurtic. Figure 3 shows the "Sample3.png" which includes histogram graphs.

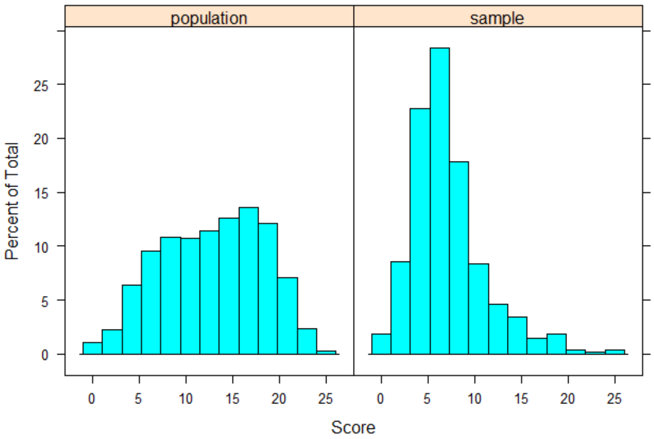


Figure 3. Histograms for the "population" and the "sample" Desired to have Skewed Distribution with Replacement (Sample3.png)

Evaluating the Function's Stability

Measures of kurtosis and skewness are used to determine if indicators met normality assumptions (Kline, 2005). The extent to which a frequency distribution diverges from symmetry is described as skewness. Kurtosis is a measure of how flat the top of a symmetric distribution is when compared to a normal distribution of the same variance. A perfect symmetrical distribution will have a skewness of 0 and a kurtosis of -3 ('excess' kurtosis of 0). The original kurtosis value is sometimes called kurtosis (proper), and West, Finch, & Curran (1995) proposed a reference of substantial departure from normality as an absolute kurtosis (proper) value > 7. Most statistical packages such as SPSS provide 'excess' kurtosis obtained by subtracting 3 from the kurtosis (proper). In this study, 'excess' kurtosis is used for practical reasons. Distributions that are more flat-topped than normal distributions are called platykurtic, and their kurtosis values are less than 3. Distributions that are less flat-topped than normal distributions are called leptokurtic, and their kurtosis values are more than 3 (Flott, 1995; Wuensch, 2005).

There is no consensus about the skewness and kurtosis values which indicate normality in the literature. It is widely accepted that absolute skew and kurtosis values up to one provide normality. (Büyüköztürk, Çokluk, & Köklü, 2014; Huck, 2012; Ramos et al., 2018). Furthermore, there are some suggestions that much larger values of skewness and kurtosis indicate normality (Brown, 2006; Kim, 2013; West et al., 1995). Furthermore, kurtosis is generally interesting only when dealing with approximately symmetrical distributions. Skewed distributions are always leptokurtic. Besides, kurtosis can be thought of as a measurement which adjusts to remove the effect of skewness (Blest, 2003). Moreover, social science researchers are concerned with the deviation of the distribution from symmetry rather than its flatness. In addition, high kurtosis should be considered for the researcher to look for outliers in one or both tails of the distribution (Wuensch, 2005). For this reason, although the possible skewness and kurtosis values can be selected in the *draw_sample()* function, the data provided by the function provides very close results in the skewness values, but not in the kurtosis values. We recommend that users should choose kurtosis values closest to 0 for normal distributions and higher than 3 for leptokurtic distributions, and lower than 3 for platykurtic distributions. If the aim of the researcher is to obtain data with outliers, the value of kurtosis can be increased up to 20 according to the number of outliers.

In order to determine how close the drawn sample to the reference distribution, a function called *draw_sampleRMSE()* is written. This function can take samples from the data with different set.seed values as much as the specified number of replications. The functions' output is the skewness and kurtossis values for each replication of *draw_sample() function*. R commands for *draw_sampleRMSE()* function are given in Table 15.

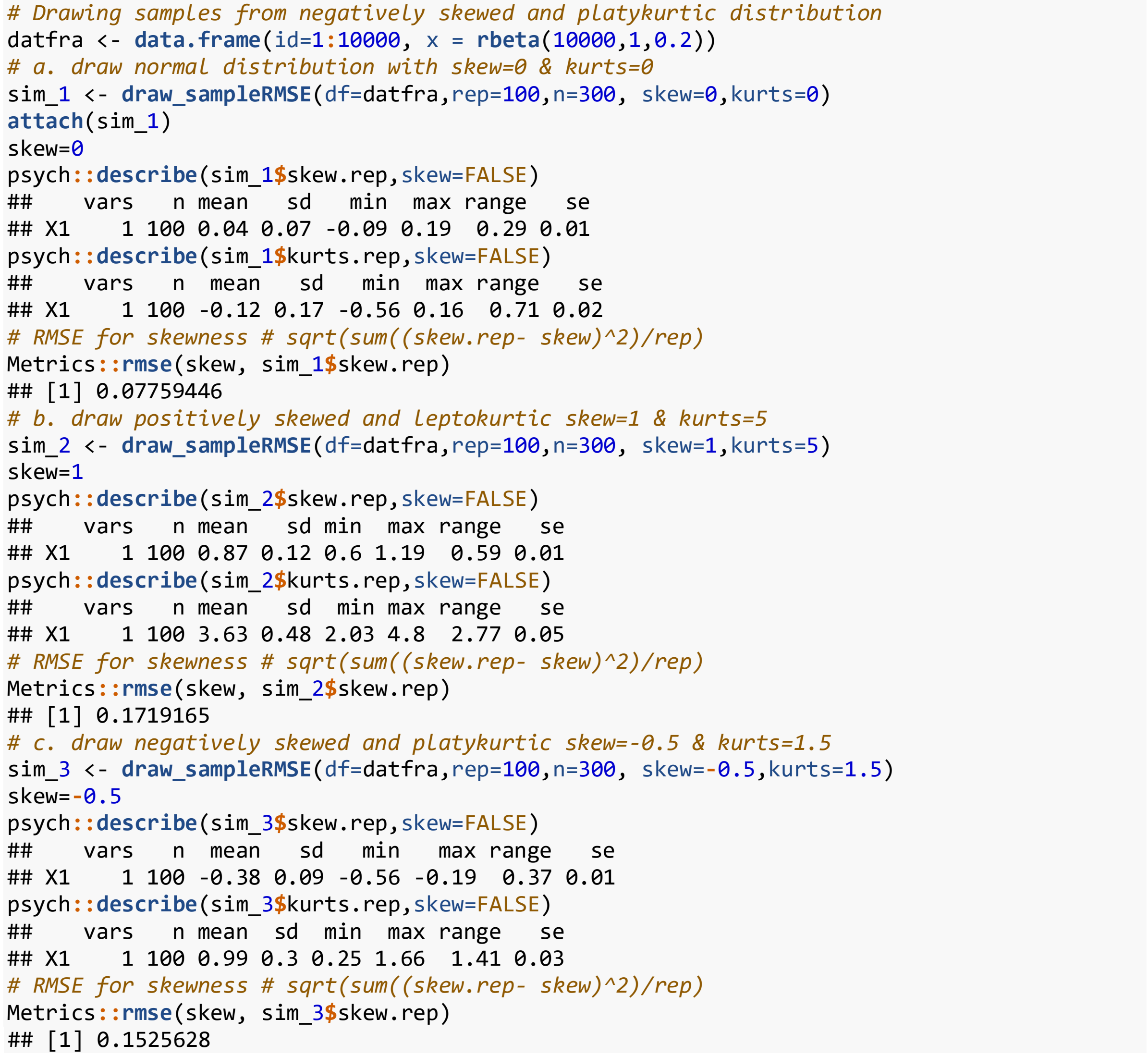
Table 15. R Commands for *draw_sampleRMSE()* Function

```
## A function for running the draw_sample() function
draw_sampleRMSE <- function(df,rep=rep,n=n, skew=skew,kurts=kurts){
  skew.rep <- c();kurts.rep <-c()
  i=1
  while(i < (rep+1)) {
    skip_to_next <- FALSE
    tryCatch({
      set.seed(sample(1:10000,1))
      result <- drawsample::draw_sample(dist =df, n = n,
                                        skew = skew,kurts = kurts, output_name = c("samp
le",paste(i)))$desc },
      error = function(e) { skip_to_next <<- TRUE})
    if(skip_to_next) { next }
      if(is.na(result[3,6])==FALSE){
      skew.rep[i] <- result[3,6]
      kurts.rep[i] <- result[3,7]
      i= i +1
    }else{i=i}
  }
  return(data.frame(skew.rep,kurts.rep))
}
```

To illustrate the stability of *draw_samples()*, two simulated datasets are used. First, negatively skewed and platykurtic data was generated with a sample size of 10000 by using *rbeta()* function, called "datfra".

Then, 100 different samples were drawn from "datfra" with a different set.seed values with *draw_sampleRMSE()* function. After calculating the skewness and kurtosis values for each sample, the RMSE values and descriptive statistics were presented in Table 16 for skewness values, and only descriptive statistics were presented for kurtosis

Table 16. Drawing Samples from Negatively Skewed and Platykurtic Distribution



```
# Drawing samples from negatively skewed and platykurtic distribution
datfra <- data.frame(id=1:10000, x = rbeta(10000,1,0.2))
# a. draw normal distribution with skew=0 & kurts=0
sim_1 <- draw_sampleRMSE(df=datfra,rep=100,n=300, skew=0,kurts=0)
attach(sim_1)
skew=0
psych::describe(sim_1$skew.rep,skew=FALSE)
##    vars   n mean   sd   min  max range   se
## X1    1 100 0.04 0.07 -0.09 0.19  0.29 0.01
psych::describe(sim_1$kurts.rep,skew=FALSE)
##    vars   n  mean   sd   min  max range   se
## X1    1 100 -0.12 0.17 -0.56 0.16  0.71 0.02
# RMSE for skewness # sqrt(sum((skew.rep- skew)^2)/rep)
Metrics::rmse(skew, sim_1$skew.rep)
## [1] 0.07759446
# b. draw positively skewed and leptokurtic skew=1 & kurts=5
sim_2 <- draw_sampleRMSE(df=datfra,rep=100,n=300, skew=1,kurts=5)
skew=1
psych::describe(sim_2$skew.rep,skew=FALSE)
##    vars   n mean   sd min  max range   se
## X1    1 100 0.87 0.12 0.6 1.19  0.59 0.01
psych::describe(sim_2$kurts.rep,skew=FALSE)
##    vars   n mean   sd  min max range   se
## X1    1 100 3.63 0.48 2.03 4.8  2.77 0.05
# RMSE for skewness # sqrt(sum((skew.rep- skew)^2)/rep)
Metrics::rmse(skew, sim_2$skew.rep)
## [1] 0.1719165
# c. draw negatively skewed and platykurtic skew=-0.5 & kurts=1.5
sim_3 <- draw_sampleRMSE(df=datfra,rep=100,n=300, skew=-0.5,kurts=1.5)
skew=-0.5
psych::describe(sim_3$skew.rep,skew=FALSE)
##    vars   n  mean   sd   min   max range   se
## X1    1 100 -0.38 0.09 -0.56 -0.19  0.37 0.01
psych::describe(sim_3$kurts.rep,skew=FALSE)
##    vars   n mean  sd  min  max range   se
## X1    1 100 0.99 0.3 0.25 1.66  1.41 0.03
# RMSE for skewness # sqrt(sum((skew.rep- skew)^2)/rep)
Metrics::rmse(skew, sim_3$skew.rep)
## [1] 0.1525628
```

In the first example in Table 16, normal distributions are drawn from the negatively skewed and leptokurtic distribution. It is seen that the mean of skewness and kurtosis values of the distributions produced in this example are quite close to the determined value, 0. The skewness values vary between -0.09 and 0.19, and kurtosis varies between -0.56 and 0.16. RMSE calculated for the skewness value was determined as 0.078.

In the second example in Table 16, positively skewed and leptokurtic distributions are drawn from the negatively skewed and leptokurtic distribution. It is seen that the mean skewness value of the distributions produced in this example is quite close to the determined value, 1. However, the mean kurtosis value of the distributions produced in this example is larger than 3, as expected for leptokurtic distributions. The skewness values vary between 0.6 and 1.19, and RMSE calculated for the skewness value was determined as 0.172.

In the third example in Table 16, negatively skewed and platykurtic distributions are drawn from the negatively skewed and leptokurtic distribution. It is seen that the mean skewness value of the distributions produced in this example is quite close to the determined value, -0.5. However, the mean kurtosis value of the distributions produced in this example is smaller than 3, as expected for platykurtic distributions. The skewness values vary between -0.56 and -0.19, and RMSE calculated for the skewness value was determined as 0.153.

Second, positively skewed and platykurtic data was generated with a sample size of 10000 by using *rbeta()* function, called "datfra2".Then, 100 different samples were drawn from the "datfra2" with a different set.seed values with *draw_sampleRMSE()* function. After calculating the skewness and kurtosis values for each sample, the RMSE values and descriptive statistics were presented in Table 17 for skewness values, and only descriptive statistics were presented for kurtosis.

Table 17. Drawing Samples from Negatively Skewed and Platykurtic Distribution

```
# Drawing samples from positively skewed and platykurtic distribution
datfra2 <- data.frame(id=1:10000, x = rbeta(10000,1,6))
# d. draw positively skewed and leptokurtic skew=2 & kurts=5
sim_4 <- draw_sampleRMSE(df=datfra2,rep=100,n=300, skew=2,kurts=5)
skew=2
psych::describe(sim_4$skew.rep,skew=FALSE)
##    vars   n mean   sd  min  max range   se
## X1    1 100    2 0.17 1.51 2.27  0.76 0.02
psych::describe(sim_4$kurts.rep,skew=FALSE)
##    vars   n mean   sd  min  max range   se
## X1    1 100 5.21 0.77 3.01 6.66  3.66 0.08
# RMSE for skewness # sqrt(sum((skew.rep- skew)^2)/rep)
Metrics::rmse(skew, sim_4$skew.rep)
 ## [1] 0.1740071
```

In Table 17, positively skewed and leptokurtic distributions are drawn from the positively skewed and leptokurtic distribution. It is seen that the mean of skewness values of the distributions produced in this example are quite close to the determined value, 2. The skewness values vary between 1.51 and 2.27, and kurtosis values are higher than 3. RMSE calculated for the skewness value was determined as 0.174. As a result, it was found that the function gives more consistent results at more common skewness values (between -1 + 1).

INSTALLING THE drawsample PACKAGE

The R package *drawsample* can be installed from CRAN with *install.packages("drawsample")* command. The package *drawsample* automatically provides the example data set "example_data". Additionally, package's files are available from the GitHub repository https://github.com/atalay-k/drawsample.

FINAL REMARKS

In this study, an R package *drawsample* has been developed to draw samples with desired properties from a given distribution. Contrary to simulation studies, the importance given to studies with real data has increased in recent years. It is thought that using the drawn samples obtained from the real data with *drawsample* package will provide an alternative to simulation studies as well as a complement for these studies. In addition, since the real data is used instead of the simulation studies, the descriptive characteristics of the study groups can be examined. Thus, it may be possible to examine the demographic characteristics of the individuals making up the sample.

In this study, four examples with real data are presented. It can be inferred from the examples in the study; the sample drawn from the real data is very close to the desired properties. However, it should be noted that it is not so easy to draw samples that perfectly match the desired properties in real data sets to draw sample from simulation data sets. Apart from the examples discussed in the study, two simulation data were genareted to evaluate the stability of the of *draw_sample()*. Then samples were drawn from these data sets under four cases. For each case in the the *draw_sample()*, 100 replications were performed and RMSE values are reported. As a limitation, *draw_sample()* yields more inconsistent results at less common skew values. In addition, this inconsistency is directly related to the characteristics of the data from which the sample is taken. For example, the size of the population and its similarity (distribution shape) with the desired data are directly related to the amount of inconsistency. Due to the nature of random assignment, the function will get different samples in each time, even for the same values. Users are advised to run the function several times if they cannot obtain samples with the desired properties the first time.

Researchers can access the web-wide data sets provided by the "https://toolbox.google.com/datasetsearch" search engine, as well as they can access large public data such as TIMSS (Trends in International Mathematics and Science Study), PIRLS (The Progress in International Reading Literacy Study), and PISA (The Program for International Student Assessment). Various studies can be done by drawing samples using the data sets mentioned above based on distribution properties. In situations like this, a good approach would be to draw a sample of the population. As authors, we are open to all kinds of suggestions in the development of the *drawsample* package.

REFERENCES

Abdel-fattah, A.-F. A. (1994, April). *Comparing BILOG and LOGIST estimates for normal, truncated normal and beta ability distributions.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

American Educational Research Association. (2020). *Journal of Educational and Behavioral Statistics*. Retrieved from https://journals.sagepub.com/description/jeb

Bahry L. M. (2012). *Polytomous item response theory parameter recovery: An investigation of non-normal distributions and small sample size* (Unpublished Master's disertation). University of Alberta, Edmonton, Canada.

Bıkmaz Bilgen, Ö., & Doğan, N. (2017). Çok kategorili parametrik ve parametrik olmayan madde tepki kuramı modellerinin karşılaştırılması [Comparison of Polytomous Parametric and Nonparametric Item Response Theory Models]. *Journal of Measurement and Evaluation in Education and Psychology, 8*(4), 354-372. DOI: 10.21031/epod.346650

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology,9*(2), 78–84. DOI: 10.1027/1614-2241/a000057

Blanca, M., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option?. *Psicothema*, *29*(4), 552-557. DOI: 10.7334/psicothema2016.383

Blest, D. C. (2003). A new measure of kurtosis adjusted for skewness. *Australian & New Zealand Journal of Statistics*, *45*(2), 175-179.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* New York, NY: Guilford Press.

Büyüköztürk, Ş., Çokluk, Ö., & Köklü, N. (2014). *Sosyal Bilimler için istatistik.* Ankara: Pegem Akademi.

Çelikten, S., & Çakan, M. (2019). Bayesian ve nonbayesian kestirim yöntemlerine dayali olarak siniflama indekslerinin TIMSS 2015 matematik testi üzerinde incelenmesi. [Investigation of classification indices on TIMSS 2015 mathematic-subtest through Bayesian and nonbayesian estimation methods]. *Necatibey Eğitim Fakültesi Elektronik Fen ve Matematik Eğitimi Dergisi*, *13*(1), 105-124.

Courville, T. G. (2004). *An empirical comparison of item response theory and classical test theory item/person statistics* (Doctoral dissertation, Texas A&M University).

Custer, M., Omar, M. H., & Pomplun, M. (2006). Vertical scaling with the Rasch model utilizing default and tight convergence settings with WINSTEPS and BILOG-MG. *Applied Measurement in Education, 19*(2), 133-149. DOI: 10.1207/s15324818ame1902_4

D'agostino, R. B., Belanger, A., & D'Agostino Jr, R. B. (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician*, *44*(4), 316-321. DOI: 10.2307/2684359

Doğan, N. & Tezbaşaran, A. A. (2003). Klasik test kuramı ve örtük özellikler kuramının örneklemler bağlamında karşılaştırılması. [Comparison of classical test theory and latent traits theory by samples]. *Hacettepe University Journal of Education, 25,* 58–67. DOI: 10.17860/efd.86348

Doğan, N., & Kılıç, A. F. (2018). The Effects of Sample Size, Correlation Technique, and Factor Extraction Method on Reliability Coefficients. *Kastamonu Eğitim Dergisi*, *26*(3), 697-706. DOI: 10.24106/kefdergi.413303

Dolma, S. (2009). *Çok ihtimalli rasch modeli ile derecelendirilmiş yanıt modelinin örtük özellikleri tahminleme performansı açısından simülasyon yöntemiyle karşılaştırılması* [*A simulation study for the comparison of* *the polytomous Rasch model and graded response model according to their performance on recovering the latent traits*]. (Unpublished Doctoral dissertation). İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul, Turkey.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist, 63*(7), 591–601. doi: 10.1037/0003-066X.63.7.591

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58,* 357-381. doi: 10.1177/0013164498058003001

Fialkowski, A. C. (2018). SimMultiCorrData: Simulation of Correlated Data with Multiple. Retrieved from: https://cran.r-project.org/web/packages/SimMultiCorrData/index.html

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In Hancock, G.R. & Mueller R. O. (Eds.), *Structural equation modeling: A second course*, (pp. 269-314). Information Age Publishing, U.S.A.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*, 521-532. doi: 10.1007/BF02293811

Flott, L. W. (1995). Quality control: Measurement error. *Metal Finishing*, *93*(9), 72-75.

Geary, R. C. (1947). Testing for normality. *Biometrika, 34*(3/4), 209-242. DOI: 10.1093/biomet/34.3-4.209

Gelbal, S. (1994). *P madde güçlük indeksi ile Rasch modelinin b parametresi ve bunlara dayalı yetenek ölçüleri üzerine bir karşılaştırma* [*A comparison of item difficulty index P and Rasch model b parameters and their ability measures based on them*].Doctoral disertation, Hacettepe University, Ankara. Retrieved from

Gotzmann, A. J. (2011). *Comparison of vertical scaling methods in the context of NCLB.* (Doctoral dissertation, University of Alberta, Alberta). Retrieved from https://era.library.ualberta.ca/items/04a8d59c-791d-435b-bde6-7a6de3012169

Hallgren, K. A. (2013). Conducting simulation studies in the R programming environment. *Tutorials in Quantitative Methods for Psychology*, *9*(2), 43–60. DOI:10.20982/tqmp.09.2.p043

Han, K. T. (2007). WinGen: Windows software thatgenerates IRT parameters and item responses. *Applied Psychological Measurement, 31*(5), 457-459. doi: 10.1177/0146621607299271

Han, K. T., & Hambleton, R. K. (2007). *User's Manual: WinGen (Center for Educational Assessment Report No. 642*). Amherst, MA: University of Massachusetts, School of Education.

Headrick, T. C. 2002. Fast fifth-order polynomial transforms for generating univariate and multivariate non-normal distributions. *Computational Statistics & Data Analysis. 40*(1),685–711. doi: 10.1016/S0167-9473(02)00072-5

Huck, S. W. (2012). *Reading statistics and research* (6th ed). Boston: Pearson.

John Wiley & Sons, Inc.-a. (2019). *Educational Measurement: Issues and Practice.* Retrieved from https://onlinelibrary.wiley.com/page/journal/17453992/homepage/productinformation.html

Kaya, Y., Leite, W. L., & Miller, M. D. (2015). A comparison of logistic regression models for DIF detection in polytomous items: The effect of small sample sizes and non-normality of ability distributions. *International Journal of Assessment Tools in Education*, *2*(1), 22-39. doi: 10.21449/ijate.239563

Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, *36*(5), 399-419. DOI: 10.1177/0146621612446170

Kim, H. Y. (2013). Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restorative dentistry & endodontics*, *38*(1), 52-54.

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

Kline, R. B. (2005). *Principles and practice of structural equations modeling.* New York: Guilford.

Kogar, H . (2018). Effects of Various Simulation Conditions on Latent-Trait Estimates: A Simulation Study. International Journal of Assessment Tools in Education , 5 (2) , 263-273. DOI: 10.21449/ijate.377138

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

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156–166. DOI: 10.1037/0033-2909.105.1.156

Ministry of National Education. (2013). *Parasız Yatılılık Ve Bursluluk Sınavı (PYBS) Sınav Kılavuzu* [*Guide of Public Boarding And Scholarship Examination (PBSE)*]. Retrieved from http://www.meb.gov.tr/sinavlar/dokumanlar/2013/kilavuz/2013_PYBS_2.pdf

Ministry of National Education. (2020). *Ortaöğretim Kurumlarına İlişkin Merkezi Sınav Kılavuzu [Guide for Central Examination Secondary Education Institution]*. Retrieved from: http://www.meb.gov.tr/meb_iys_dosyalar/2020_07/17104126_2020_Ortaogretim_Kurumlarina_Iliskin_Merkezi_Sinav.pdf

Nartgün, Z. (2002). *Aynı tutumu ölçmeye yönelik likert tipi ölçek ile metrik ölçeğin madde ve ölçek özelliklerinin klasik test kuramı ve örtük özellikler kuramına göre incelenmesi [Examination of item and scale properties of likert type scale and metric scale to measure the same attitude according to classical test theory and item response theory]*. (Unpublished doctoral dissertation). Hacettepe University Social Sciences Institute, Ankara.

Olivier, J., & Norberg, M. M. (2010). Positively skewed data: revisiting the box-cox power transformation. *International Journal of Psychological Research*, *3*(1), 68-95. DOI: 10.21500/20112084.846

Pearson, E.S. (1932). The analysis of variance in cases of non-normal variation. *Biometrika. 23*, 114-133.

Pomplun, M., Omar, M. H., & Custer, M. (2004). A comparison of WINSTEPS and BILOG-MG for vertical scaling with the Rasch model. *Educational and Psychological Measurement*, *64*(4), 600-616. doi: 10.1177/0013164403261761

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved September 10, 2019, from http://www.R-project.org/

Ramos, C., Costa, P. A., Rudnicki, T., Marôco, A. L., Leal, I., Guimarães, R., ... & Tedeschi, R. G. (2018). The effectiveness of a group intervention to facilitate posttraumatic growth among women with breast cancer. *Psycho*‐*oncology*, *27*(1), 258-264. DOI:10.1002/pon.4501

Revelle, W. (2018) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.8.12.

Reyhanlıoğlu Keceoğlu, Ç. (2018). *Parametrik ve Parametrik Olmayan Madde Tepki Kuramında Farklı Örneklem Büyüklüklerine ve Boyutluluklarına Göre Parametre Değişmezliğinin İncelenmesi.* Unpublished doctoral dissertation). Hacettepe University Social Sciences Institute, Ankara.

Şahin, M. G., & Yıldırım, Y. (2018). The examination of item difficulty distribution, test length and sample size in different ability distribution. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *9(3),* 277-294. DOI: 10.21031/epod.385000

Sarkar, D. (2008). *lattice: Multivariate Data Visualization with R.* Springer-Verlag, New York.

SAS Institute Inc. (2009). *SAS/Stat User's Guide, version 9.2,* (Version 9.2). Cary, NC.

Sen, S., Cohen, A. S., & Kim, S. H. (2014, November). Robustness of mixture IRT models to violations of latent normality. In *Quantitative Psychology Research: The 78th Annual Meeting of the Psychometric Society* (Vol. 89, p. 27). Springer.

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

Sireci, S. G. (1991). *"Sample-Independent" Item Parameters? An Investigation of the Stability of IRT Item Parameters Estimated from Small Data Sets.* Paper presented at the annual Conference of Northeastern Educational Research Association, New York, NY.

SSCP (2019). *2019 YKS Değerlendirme Raporu [2019 Examinations of the Council of Higher Education Assessment Report]*. Retrieved from https://dokuman.osym.gov.tr/pdfdokuman/2019/GENEL/yksDegRaporweb03092019.pdf

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

Urry, V. W. (1974). Approximations to item parameters of mental test models and their uses. Educational and Psychological Measurement, *34*, 253-269. doi: 10.1177/001316447403400206

Uysal, İ. (2014). *Comparison of irt test equating methods for mixed format tests.* [*Madde tepki kuramına dayalı test eşitleme yöntemlerinin karma modeller üzerinde karşılaştırılması*]. (Master disertation, Bolu Abant Izzet Baysal University, Bolu). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/
Variable Types. R package version 0.2.2.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with non-normal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (p. 56–75). Newbery Park, CA: Sage

Wickham, H., François, R., Henry, L., & Müller, K. (2016) *tibble: Simple data frames.* Retrieved from https://CRAN.Rproject.org/package=tibble. R package version 3.0.3

Wickham, H., François, R., Henry, L., & Müller, K. (2019). dplyr: A Grammar of data manipulation. R package version 0.8.0.1. https://CRAN.R-project.org/package=dplyr

Wicklin, R. (2013). *Simulating data with SAS*. SAS Institute.

Wuensch, K. L. (2005). *Kurtosis. Encyclopedia of Statistics in Behavioral Science*. doi:10.1002/0470013192.bsa334

Yıldırım, H., Uysal-Saraç, M., & Büyüköztürk, Ş. (2018). Farklı örneklem büyüklüğü ve dağılımı Koşullarında WLS ve Robust WLS yöntemlerinin karşılaştırılması. *Ilkogretim Online, 17*(1), 431-439. doi: 10.17051/ilkonline.2018.413794

Yıldırım, Y. (2015). *Derecelendirilmiş tepki modeli temelli parametre kestiriminde normallik sayıltısı ihlalinin ölçme kesinliğine etkisi* [*The effect of normality violation in the process of parameter estimation based upon Graded Response Model on measurement precision*]. (Master disertation, Gazi University, Ankara). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/

Yoes, M. E. (1993*). A comparison of the effectiveness of item parameter estimation techniques used with the three-parameter logistic item response theory model. (Volumes I and II).* Unpublished Ph.D., University of Minnesota, Minneapolis/St. Paul, MN.

R'da "*drawsample"* Paketi ile Evrenden İstenilen Özelliklere Sahip Örneklem Çekme

Giriş

Eğitimde ve psikolojide ölçme ve değerlendirme alanında, puanların dağılımı grupların betimlenmesinde önemli bir role sahiptir. Grupların betimlenmesine ek olarak, normallik varsayımına dayanan birçok anlam çıkarıcı istatistiksel teknikleri kullanmak için normallik varsayımını test etmek çok önemlidir. Ancak Erceg-Hurn ve Mirosevich'in (2008) belirttiği gibi, gerçek veriler analiz edilirken normallik varsayımı nadiren karşılanmaktadır. Yapılan birçok araştırmada belirtildiği gibi, normal olmayan dağılımlar normal dağılıma göre daha yaygındır (Blanca, Arnau, López-Montiel, Bono ve Bendayan, 2013; Geary, 1947; Micceri, 1989; Olivier ve Norberg, 2010; Pearson, 1932).

Normallik varsayımının karşılanamaması, normalliğin ihlali ve dağılım türleri; test eşitleme, bilgisayarda bireyselleştirilmiş test uygulamaları, değişen madde fonksiyonu, sınıflandırma ve gizil puan kestirimleri gibi önemli konularda birçok araştırmacının odak noktası olmuştur (Custer, Omar, &Pomplun, 2006; Finney & DiStefano, 2006; Gotzmann, 2011; Kieftenbeld & Natesan, 2012; Kirisci, Hsu, & Yu, 2001; Kolen, 1985; Kogar, 2018; Seong, 1990; Uysal, 2014; Yıldırım, 2015). Normal dağılım ile normal olmayan dağılım türlerini elde etmede simülasyon ile veri üretmeye çok sık başvurulmaktadır (Abdel-fattah, 1994; Bıkmaz-Bilgen & Doğan, 2017; Dolma, 2009; Kaya, Leite, & Miller, 2015; Urry, 1974; Yıldırım, Uysal-Saraç, & Büyüköztürk, 2018; Yoes, 1993). Araştırmacılar farklı dağılım türlerinde veri üretilmesinde ise çeşitli yazılımlardan yararlanmıştır. Bahry (2012), WinGen 3.1 (Han, 2007) ile beta dağılımı ile çarpıklık katsayısı 0; 0,5 ve 1,00 olan örneklem büyüklüğü 100 ile 3000 arasında olan örneklemler üretmiştir. WinGen'e benzer şekilde SAS yazılımı (SAS Enstitüsü, 2009) ve R'da (R Core Team, 2014) farklı dağılım türleri elde etmede kullanılabilir. Örneğin Gotzmann (2011), normal dağılım gösteren dağılımı üretmede SAS'daki "Normal Distribution Function" dan ve çarpık dağılım gösteren dağılımı üretmede "RAND Beta Distribution Function" dan yararlanarak iki durum için 2000000 birey parametresi üretmiş ve bu verilerden yetenek parametresi ortalamaları araştırmanın amacına uygun olacak şekilde belirlenmiş ve farklı örneklem büyüklüklerinde (1500, 3000) rastgele veri setleri seçilmiştir. Veri üretmede, beta dağılımlarının kullanımı, çarpık puan dağılımlarını üretmeyi kolaylaştırmaktadır (Han ve Hambleton; 2007). Beta dağılımının bileşenleri α ve β parametreleridir. Bazı araştırmacılar ise simülasyon verisinden istendik özelliklere sahip örneklem çekmektedirler. Bu amaç doğrultusunda, orijinal veri setinden çarpık dağılıma sahip örneklem çekmede, Fleishman'ın (1978) güç yöntemi uygundur (Blanca, Alarcón, Arnau, Bono ve Bendayan, 2017; Stone, 1992; Kieftenbeld ve Natesan, 2012; Sen, Cohen ve Kim; 2014).

Simülasyon yöntemleri esnektir ve elde edilmesi mümkün olmayan problemlere nicel yanıtların sağlanması için uygulanabilir (Hallgren, 2013). Simülasyon güçlü bir teknik olmasına rağmen, sonuçları genelleme, düzenleme ve gerçek verilere uygulama güçlüğü gibi bazı sınırlıkları vardır (Wicklin, 2013). Simülasyon verileri, gerçek verilerde ulaşılamayan mükemmel bir uyum sağlar. Hallgren'in (2013) belirttiği gibi, gerçek dünya veri setleri, simülasyon çalışmalarında oluşturulan ve genellikle mükemmel uyum olarak adlandırılabilecek idealist koşullar altında üretilen "temiz" veri setlerinden daha "kirli" olacaktır. Sireci (1991) gerçek test verileri kullanılmadığında, üretilen veriler pratikte karşılaşılan ilgili durumun özelliklerini doğru olarak yansıtıp yansıtmadığı bilinemeyeceğini ve geçerliği test edilemeyeceğini ifade etmiştir. Bu yüzden çalışmalarda gerçek veri kullanımı çalışmaların önemini artırmaktır. Ayrıca, Educational Measurement: Questions and Practice (EM: IP) ve Journal of Educational and Behavioral Statistics (JEBS) gibi bazı prestijli dergiler, uzun bir süre zarfında simülasyon çalışmalarını kabul etseler de, günümüzde simülasyon temelli çalışmaların "uygun olmayan makale konu örneklerinden" veya "düşük önceliğe sahip" olarak kabul edildiğini belirtmiştir (John Wiley & Sons Inc., 2019; American Educational Research Association, 2020).

Gerçek uygulamalarda veri toplama süreci zorluklarla doludur. Elde edilen örneklemler evren dağılımını temsil etmiyor, normal dağılmıyor veya istenen dağılıma uygun olmayan bir halde olabilir. Araştırma problemlerine dayalı olarak normalliği sağlamada, normal dağılımdan uzaklaştıran verilerin çıkarıldığı ya da ayrıştırıldığı bazı araştırmalara rastlanılmıştır. Gelbal (1994) yaptığı araştırmada, araştırmasının amacına uygun olarak, 2072 beşinci sınıf öğrencisine uygulanan Türkçe testi ve 2077 beşinci sınıf öğrencisine uygulanan Matematik testine ait verileri incelemiş test puanları normal dağılım göstermediği için Türkçe testinden 506 öğrenciyi, Matematik testinden 521 öğrenciyi çıkarmış ve normal dağılım gösteren iki yeni veri seti elde ederek test puanlarının hem normal dağıldığı hem de normal dağılmadığı durumlar üzerinde çalışmıştır. Doğan ve Tezbaşaran (2003) yaptıkları çalışmada amaçları doğrultusunda istenen dağılımı sağlayan verilerin örnekleme alınmasını sağlamışlardır. Örneklemlerin seçiminde random ve kasıtlı örnekleme tekniklerinin kullanıldığını ifade etmişlerdir. Araştırmacılar; amaçları doğrultusunda Orta Öğretim Kurumları Öğrenci Seçme ve Yerleştirme Sınavı (ÖSYS) 2001 giren bireylerin oluşturduğu evrenden örneklem büyüklükleri 2353 ile 29244 arasında değişen random, sağa çarpık, sola çarpık, basık ve normal dağılım gösteren beş örneklem seçmiştir. Çarpık örneklemlerde, yapılacak karşılaştırmaların isabetliliğini artırmak için çarpıklık mutlak değerleri (1,00) ve basıklık değerleri (1,37) eşit tutulmuş ve basık dağılımda çarpıklık katsayısının 0 olması sağlanmıştır. Doğan ve Tezbaşaran'ın (2003) çalışmasına benzer şekilde, Şahin ve Yıldırım (2018), başlangıçta sağa çarpık bir veri seti olan Seviye Belirleme Sınavı (SBS) verilerinden (çarpıklık katsayısı = 1,05) hem sağa çarpık hem de sola çarpık yetenek dağılımları elde etmiştir. Sola çarpık veri setleri için, amaçlanan örneklemeyle istenen örneklem dağılımı sağlanmış ve araştırma kapsamında ele alınan tüm örneklem büyüklüklerinde örneklemler için çarpıklık katsayısı ≈-1,00 olan gruplar seçilmiştir. Yukarıdakilere ek olarak, literatürde birçok araştırmacı, yaptıkları çalışmaların amacına uygun olacak şekilde gerçek veri setinden (evren) örneklem almayı seçmiştir (Courville, 2004; Doğan ve Kılıç, 2018; Fan, 1998; Nartgün, 2002; Reyhanlıoğlu Keçeoğlu, 2018). Evrenden örnekleme sürecinde, gelecekteki çalışmalar için örneklem seçimini kolaylaştıran ve istenen özelliklere yaklaştıran bir araca sahip olmanın önemli olacağı düşünülmektedir. Öyle ki normal dağılımdan farklı yetenek dağılımlarına sahip örneklemler üzerinde çalışılması literatürde bazı araştırmaların sonucunda önerilmiştir (Çelikten ve Çakan, 2019). Araştırmalar incelendiğinde, araştırmacıların geniş bir veri setinden istenen özelliklere sahip örneklem seçimini sağlayacak bir araca ihtiyaç olduğu sonucuna varılmıştır.

Bu çalışmanın amacı gerçek bir veri setinden istendik özelliklere sahip örneklemlerin seçilmesinde yararlanılacak *drawsample* adlı bir R paketinin geliştirilmesidir. Bu amaç doğrultusunda seçilecek örneklemin sahip olması beklenen özelliklerden normallikten sapma ölçüleri (çarpıklık ve basıklık) ve örneklem büyüklüğü gibi koşulların bir veya birden fazlasının belirlenmesi ile kasıtlı örnekleme yapılabilmektedir. Çalışmanın amacı için Fleishman'ın (1978) güç yönteminden yararlanılmıştır.

Geliştirilen *drawsample* paketinde yer alan *draw_sample()* fonksiyonu ile evrenden ya da büyük bir örneklemden, istendik özelliklere sahip örneklem çekilmesinin olabildiğince kullanılışlı olması beklenmekte ve gerçek veriler ile yapılan araştırmaların önem kazandığı bu dönemde simülasyon çalışmalarına alternatif oluşturarak, dağılıma ilişkin gerçek verilere dayalı farklı konularda yapılacak olan çalışmalara katkı sağlayacağı düşünülmektedir.

R'da draw_sample() Fonksiyonu

İstenilen özelliklerde örneklem seçmek amacıyla 6 argümanlı bir *draw_sample()* fonksiyonu yazılmıştır. Fonksiyonun argümanlarına Tablo 1'de yer verilmiştir.

Tablo 1: *draw_sample()* Fonksiyonun Argümanları

Argümanlar	
dist	Very seti: ID ve puanları içeren iki sütunlu bir veri seti
n	Sayısal: Örneklem büyüklüğü
skew	Sayısal: Çarpıklık değeri
kurt	Sayıısal: Basıklık değeri
replacement	Mantıksal: Yeniden örnekleme yapılsın mı? (varsayılan FALSE'dur)
output_name	Karakter: İki bileşenli çıktı dosyasının adı

Bu fonksiyonda, "dist" argümanı örneklemlerin çekileceği veri setidir. Veri seti ilk sütununu öğrenci kimlik numaraları (ID) ve ikinci sütununu öğrencinin toplam test puanını veya yetenek puanı (theta) içerecek şekilde iki sütundan oluşmalıdır. Fonksiyonun argümanlarından istenen örneklem büyüklüğü olan "n" örneklemlerin çekileceği veri setinin uzunluğundan büyükse şu hatayı verir: "Cannot take a sample larger than the length of the data". Örneğin, içe aktarılan verilerin örneklem büyüklüğü 1000 olmasına rağmen ve kullanıcılar örneklem büyüklüğü 2000 olan bir örneklem almak isterlerse fonksiyon hata verir ve çalışmayı durdurur.

Tablo 1'deki argümanlardan "skew " ve "kurts" belirlenirken, Fleishman Güç Yöntemi Ağırlıkları tablosuna başvurulmalıdır. Örneğin çarpıklık değeri 1 ve basıklık değeri 0 gibi bazı kombinasyonlara karşılık gelen Fleishman katsayıları yoktur. Eğer kişinin verdiği çarpıklık değeri tabloda yoksa veya çarpıklık değeri olsa bile o değere karşılık gelen basıklık değeri tabloda bulunmuyorsa fonksiyon "No valid power method constants could be found for the specified values. Change the values" hatası vererek çalışmayı durdurur.

Fonksiyon ile kullanıcının girdiği çarpıklık ve basıklık değerlerine sahip olan bir referans dağılım oluşturulmakta ve daha sonra evrenden referans dağılım baz alınarak bir örnekleme yapılmaktadır. Bu kısımda kullanıcının örneklemi çekmek istediği veri setinde yer alan dağılımın minumum ve maksimum değerlerine göre yeniden ölçeklenmektedir. Veri setinde oluşturulan her bir kategoride bulunan birey sayısı, oluşturulan referans dağılımının her bir kategorisinde bulunundan daha fazla olduğunda fonksiyon, *replacement* argümanının varsayılan değeri olarak yeniden örnekleme yapılmadan yürütülebilir. Ancak *draw_sample()* veri setindeki en az bir kategoride bulunan birey sayısı, referans dağılımının ilgili kategorisinde bulunan birey sayısından az ise "Cannot take a sample form that data without replacement. Please change replacement=TRUE" hatasını verir. Bu durumda, fonksiyon *replacement* argümanının değeri değiştirilerek (FALSE) kullanılabilir.

draw_sample() fonksiyonu"output" olarak üç farklı liste içermektedir. Bunlardan biri "desc" olarak adlandırılan bir çıktıdır. Bu çıktıda "population (tüm veri, içe aktarılan veriler)", "reference (referans alınan dağılım)" ve "sample (çekilen örneklem)" dağılımlarının ortalamasını, standart sapmasını, çarpıklığını ve basıklığını içerir. Diğeri "graph" olarak adlandırılırmaktadır ve sol tarafta "popülasyon (içe aktarılan veriler)" ve sağ tarafta "sample" (çıkarılan veriler) olmak üzere iki histogram grafiği içermektedir. Bir diğeri ise "sample" olarak adlandırılan ve ilk sütununu seçilen örneklemdeki bireylerin kimlik numaraları (ID) ve diğer sütununu bireylerin puanını içerecen iki sütunlu bir veri setidir.

Bu çalışmada geliştirilen fonksiyonun kullanımını göstermek amacıyla 2013 yılı 6. Sınıf Parasız Yatılılık ve Bursluluk Sınavı'na (PYBS) ait Fen ve Teknoloji ile Sosyal Bilimler alt testleri verilerine dayalı olarak örnek dört uygulama sunulmuştur. Yapılan bu örnek uygulamalarda 5000 kişilik veri setlerinden 500 kişilik örneklemler çekilmiştir. İlk iki örnek uygulamada normal dağılımdan daha sola çarpık bir dağılım şekli gösteren Fen ve Teknoloji alt testi (Score_1) verilerinden sırasıyla normal dağılıma sahip ve hem sola çarpık hem de sivri bir dağılıma sahip örneklemler çekilmiştir. Üçüncü örnek uygulamada ise normal dağılımdan daha basık bir dağılım şekli gösteren Sosyal Bilimler alt testi (Score_2) verisinden sağa çarpık ve sivri dağılıma sahip bir örneklem çekilmesi için komutlar verilmiştir. Ancak fonksiyon burada çalışmamış ve yeniden örneklemeye olanak sağlacak şekilde terar komut verilerek istenen özellikte bir örneklem elde edilmiştir. Elde edilen örneklemlerin dağılım türü açısından istenilen özelliklere sahip olduğu görülmüşür. Fonksiyonun tutarlılığının incelenmesi amacıyla *rbeta()* fonksiyonu ile üretilen iki veri setinden çekilen örneklemlere ilişkin yapılan replikasyonlar sonucu çarpıklık değerlerine ilişkin ortalama RMSE değerleri raporlanmıştır. Fonksiyonun daha sık rastlanan çarpıklık değerlerinde (-1 ve +1 arasında) daha tutarlı sonuçlar verdiği bulgusuna ulaşılmıştır.

Sonuç ve Tartışma

Bu çalışmada, belirli bir dağılımdan istenen özelliklere sahip örneklemler elde etmek için R'da *draw_sample()* adlı bir fonksiyon ve örnek veri seti sunan *drawsample* paketi geliştirilmiştir. Simülasyon çalışmalarının aksine gerçek verilerle yapılan çalışmalara verilen önem son yıllarda artmıştır. *draw_sample()* fonksiyonu ile gerçek verilerden çekilen örneklerin kullanılmasının simülasyon çalışmalarına alternatif oluşturacağı gibi bu çalışmaları tamamlayacağı düşünülmektedir. Ayrıca örneklemler gerçek kişilerden oluşacağı için çalışma gruplarının betimleyici özellikleri incelenebilir. Böylece örneklemi oluşturan bireylerin demografik özelliklerini incelemek mümkün olabilir.

Bu çalışmada hem gerçek veriden, hem de simülasyon verisine dayalı olarak örnek uygulamalar sunulmuştur. Çalışmadaki örneklerden anlaşılacağı üzere simülasyon verilerinden alınan örneklem istenilen özelliklere çok yakındır. Bununla birlikte, gerçek veri setinden istenen özelliklere mükemmel bir şekilde uyan örneklemler seçmenin, özellikle yeniden örnekleme yapılmadığında, kolay olmadığı unutulmamalıdır. Sunulan bu örnek uygulamaların dışında fonksiyonun tutarlılığının değerlendirilmesi amacıyla üretilen simülasyon verisinden tekrarlı çekilen örneklemlere dayalı olarak *draw_sample()* fonksiyonunun, daha az rastlanan büyük çarpıklık değerlerinde daha tutarsız sonuçlar verdiği görülmüştür. Ayrıca bu tutarsızlık, örneklemin çekildiği evrenin özellikleriyle de doğrudan ilişkilidir. Örneğin evren ile çekilecek örneklemin büyüklükleri ve istenen özelliklere sahip örneklemin evren ile dağılım türlerinin benzerliği tutarsızlık miktarı ile doğrudan ilişkilidir. Rastgele atamanın doğası gereği, fonksiyondaki argümanların aynı değerleri için bile her seferinde farklı örneklem çekebilir. Kullanıcılar istenilen özelliklere sahip örneklemi ilk seferde çekemezlerse, kullanıcılara fonksiyonu birkaç kez çalıştırmaları önerilir.

Araştırmacılar, "https://toolbox.google.com/datasetsearch" arama motoru tarafından sağlanan web genelindeki veri kümeleri ile TIMSS (Trends in International Mathematics and Science Study), PIRLS (The Progress in International Reading Literacy Study) ve PISA (The Program for International Student Assessment) gibi açık büyük verilere erişebilirler. Dağılım özelliklerine göre yukarıda belirtilen veri setlerinden örneklemler alınarak çeşitli çalışmalar yapılabilir. Bu gibi durumlarda, *drawsample* paketi ile evrenden örneklem çekmek iyi bir yaklaşım olacaktır. Yazarlar olarak *drawsample* paketinin geliştirilmesinde her türlü öneriye açık olduğumuzu bildiririz.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(4); 430-450

Silent Predictors of Test Disengagement in PIAAC 2012

Münevver İLGÜN DİBEK *

Abstract

Although the effects of test disengagement on the validity of the scores obtained from the data set have been examined in many studies, the predictors of the disengaged behaviors received relatively limited scholarly attention in low-stakes assessment, in particular, in international comparison studies. As such, the present study with a twofold purpose sets out to determine the best fitted explanatory item response theory model and examine the predictors of test disengagement. The data were collected by using items measuring literacy and numeracy skills of adults from different countries such as Norway, Austria, Ireland, France, Denmark, Germany, and Finland participated in PIAAC 2012. The results of the model with item and person characteristics demonstrated that adults tended to be disengaged on very difficult items. Similarly, age has a negative effect on test-taking engagement for adults in several countries such as France and Ireland, while several predictors such as educational attainment, readiness to learn, and the use of ICT skills at home and work had positive effects on test engagement. In addition, females exhibit a higher level of engagement in Norway. Overall, the findings suggested that the effect of the predictors on disengagement depended on the domain and country. So, this study brings further attention that the role of test disengagement should be a prerequisite practice before reaching a conclusion from international large-stake assessments.

Key Words: Explanatory item response theory model, low-stakes assessment, PIAAC, test disengagement.

INTRODUCTION

Examinees are not always motivated to put their full effort into responding to test items, especially in low-stakes settings, such as the Programme for the International Assessment of Adult Competencies (PIAAC). (e.g., Finn, 2015; Wise & DeMars, 2010). The reason why low test motivation is often seen in low-stakes assessments can be revealed by expectancy-value models (e.g., Eccles & Wigfield, 2002). More specifically, as indicated by these models, achievement motivation is closely affected by factors, such as expectancy and value. The former factor is defined as the individual’s expectation of achievement in responding to the test items and will be low if the item is too difficult relative to the ability of the individual. In the most general sense, the latter factor is related to the perceived importance and usefulness of the test. However, there is not a straightforward explanation since there are different aspects of value components, such as attainment value, intrinsic value, utility value, and perceived costs (Eccles & Wigfield, 2002). Both the combination of them and each aspect separately is considered to be low in low-stakes assessments. This is because, although there is a need to make a sufficient effort to respond to the test items correctly, the intrinsic motivation of some of the respondents is low, and the results obtained from the test are not vital for the respondents. Therefore, this results in a contradiction. There will be serious problems when the lower levels of motivation of individuals give rise to a low test effort (Wise & DeMars, 2010). These invalid responses cause construct-irrelevant variance and distortion of psychometric features (e.g., Rios, Guo, Mao, & Liu, 2017), leading to the misinterpretation of the results obtained from the data set (Nagy, Nagengast, Becker, Rose, & Frey, 2018). To put it in different words, the true scores of the individuals are contaminated by a systematic source of error due to their level of engagement in the test (Braun, Kirsch, Yamamoto, Park, & Eagan, 2011). In addition, disengagement gives rise to (a) inflated item difficulties, as well as deflated item discriminations (e.g., van Barnevald, 2007), (b) biased item and test information estimates (e.g., van

* Ass. Prof. Dr., TED University, Faculty of Education, Ankara-Turkey, munevver.ilgun@tedu.edu.tr, ORCID ID:0000-0002-7098-0118

To cite this article:
İlgün-Dibek, M. (2020). Silent predictors of test disengagement in PIAAC 2012. *Journal of Measurement and Evaluation in Education and Psychology, 11*(4), 430-450. doi: 10.21031/epod.796626.

Received: 17.09.2020
Accepted: 13.12.2020

Barnevald, 2007), (c) inflation of reliability estimates based on classical test theory (CTT) (e.g., Wise & DeMars, 2009), (d) erroneous flagging of differential item functioning (e.g., Wise & DeMars, 2010), and (e) decreased correlations with external variables (e.g., Wise, 2009).

Although test disengagement has been characterized in different ways in the literature, rapid guessing is the most widely used and validated one (Wise, 2015). According to Schnipke and Scrams (2002), rapid guessing behavior is the fast response of the test takers to the test items in a way that does not allow them to understand the content of the item. To determine whether this method is being implemented, Schnipke and Scrams (2002) proposed that respondents are divided into two groups according to their solution behavior or rapid guessing behavior. In this approach, the focus is on time elapsed between presenting the item to the respondents and the respondent's response to the item. If the test-taker responds in a period below a certain response time threshold, it means that s/he is displaying rapid guessing behavior. The main challenge in this situation is determining which responses to items are rapid guessing and which responses are solution behaviors.

Alternative Methods for Measuring Test Disengagement

Whatever the reason for the occurrence of disengaged behavior, measuring this behavior accurately and efficiently is crucial given the sizable validity problems that occur due to test disengagement. Test disengagement is determined by computing item response time thresholds that differentiate engaged and disengaged responses. To determine test disengagement, constant threshold and item-specific thresholds are proposed in the literature. For example, as a constant threshold, the frequently used method is the three-second rule (Kong, Wise & Bhola, 2007; Lee & Jia, 2014). The amount of time required to answer the item may vary from item to item (Lee & Jia, 2014). As an example, while respondents can answer an easy item that measures numerical skills faster, they can answer an item that includes long texts with a high reading load and measures verbal skills in a longer time. Thus, researchers tend to use item-specific thresholds, with one of the earliest and most basic approach being the visual inspection method (DeMars, 2007; Wise & Kong, 2005). For each item, the notion is to define the threshold as the judged endpoint of the short time spike in a bimodal response time distribution. In this process, the distributions of the response time of test-takers responding rapidly and those responding more slowly are presented. Although the visual inspection method has various advantages, such as easy interpretation and being evidence-based, there are disadvantages; e.g., being subjective, time-consuming, and not applicable in cases where there is no bi-model distribution (Lee & Jia, 2014; Rios et al., 2017).

Another method for determining item-specific thresholds was the one used by Lee and Jia (2014) on items in multiple-choice format. For each item, the proportion correct conditional on the response time is determined. The response time threshold is defined as proportion correct greater than the chance level for obtaining a correct answer. Since the items included in the PIAAC assessment vary in difficulty and complexity, the amount of time required to give the correct answer will differ. Therefore, considering the advantages of item-specific response time thresholds shown in previous research (e.g., Wise, 2006), the current study adopted this approach.

While much is known about the impact of disengagement on observed test scores, little is known about the impact of an item and personal characteristics on the disengagement of individuals. Some individuals consistently exhibit more disengaged behaviors than others. Determining the person and item as a source of variation could be used for examining individual differences.

Relationship Between Test Disengagement and Person- and Item-Level Variables

Considering the effects of test disengagement on the observed scores of individuals, the reasons for individuals' disengagement have become the focus of attention. Differences between individuals in terms of test disengagement show that it is crucial to take the person as a source of variation in disengagement (Wise, 2009). Therefore, examining the role of person-level variables on test disengagement is beneficial in terms of explaining these differences. To evaluate this situation in terms of large-scale applications, the results of these applications are not of vital importance for individuals

(Asseburg & Frey, 2013; Sundre & Kitsantas 2004; Wise, 2009). Therefore, according to the expectancy-value theory, individuals will attribute the same value to the areas measured in these practices. Consequently, there will be no individual differences in terms of test engagement. However, individuals' perceived expectations about their ability to answer items correctly change from one person to another, depending on the several characteristics that they have. In this regard, gender can affect their perception of the capability, and thus their engagement. Several studies in the literature indicate that males exhibit disengaged behaviors more frequently than females (e.g., DeMars, Bashkov & Socha 2013). Females tend to spend more time answering the items (Setzer, Wise, van den Heuvel, & Ling, 2013).

Although the education level and age of individuals may have a significant effect on the time they spend responding to an item in the test, it has been observed that the literature does not focus on this issue sufficiently. The investigation of this effect would help shed light on solving some unanswered questions in education. For example, highly educated individuals are committed to achieving several tasks and thus have sufficient competency (Organisation for Economic Co-operation and Development [OECD], 2016b); therefore, they may spend more time responding to an item. In addition, older adults may have the necessary knowledge and skills and tend to respond faster to items due to biological factors, such as fatigue and boredom so that they can complete the assessment as soon as possible (Xie, 2003).

Individuals' readiness to learn has an effect on their disengagement levels. It is closely related to whether adults have sufficient motivation, cognitive skills, and learning strategies to learn a task, feel curious about it, are interested in learning, look for associations among ideas, and believe that they can cope with a problem that they face (Smith, Rose, Smith & Ross-Gordon, 2015). Although the extent to which individuals have the characteristic to be measured by that test plays an important role in responding to a test item, in some cases, various factors also have a critical effect on responding behaviors. When these factors are not taken into account, invalid interpretations can be obtained by only looking at test scores (Nagy et al., 2018), At this point, considering that the test items in the PIAAC are given in a computer environment regardless of which domain measurement, the familiarity of the individuals with various technological elements such as computers and the internet will also have an effect on the individuals' behavior of responding to the test items as if they were insidious, silent factors. In other words, as a source of variation in the engagement levels of respondents, familiarity with information and communications technology (ICT) can also affect respondents' engagement. The frequent use of the ICT skills of individuals makes them familiar with computers, which increases the motivation, concentration and achievement of individuals in computer-based assessments (Mastuti & Handoyo, 2017). In addition, the extent to which the individuals use various skills at home and work can have an effect on how much effort they applied when responding to tests.

In the literature, it has been stated that several item-level variables have an impact on individuals' disengagement levels. According to the expectancy-value theory, if the individuals perceive an item as difficult by taking into consideration their competence, their engagement in the testing situation will be negatively affected. Some studies revealed that individuals put more effort into items which had moderate difficulty relative to their ability (Asseburg & Frey, 2013).

In conclusion, the importance of addressing these variables can be explained by analogy with the area above and below an iceberg. While there is only a small part of the total mass above the iceberg, there is a large part of it below, and this controls all the movements of the iceberg. At this point, the same logic can be used to explain the disengagement behaviors of individuals. In other words, in this study, these variables that make up the area under the disengagement as an iceberg will play an important role in explaining the disengagement behavior of individuals. To narrow the focus even more, when the effect of these person and item-level variables on disengagement is ignored, the difference in test scores due to disengagement could not be determined correctly (Braun et al., 2011). Thus, investigation of to what extent these variables explain the disengagement behavior is crucial.

It seems, however, that there has been extensive research on the topic of test-taking effort. Many of these endeavors possess several limitations: focusing on relatively homogenous populations based in a single country (Goldhammer, Martens & Lüdtke, 2017). To date, there have been very few studies that have examined potential differences in test-taking effort between countries in international assessments

(Rios & Guo, 2020), although their personal characteristics largely differ by culture/country (Brown & Harris, 2016). Also, regardless of the number of response categories, studies using traditional IRT models provide information on various individual or item-related characteristics such as respondents' abilities, cognitive levels, achievements, or difficulty and discrimination. Still, they are insufficient to identify systematic effects resulting from the design of the measurement process. In other words, they do not reveal common variability across items or individuals depending on the design of the measurement process or measurement tool. However, this information is very important in determining construct-irrelevant variance originating from various reasons such as cognitive, cultural, and biological factors (AERA, APA, NCME, 2014). Since data were collected in a nested design in the PIAAC study, analyses were done using explanatory item response theory models (EIRT), which allow to include several item and person characteristics as first-level and second-level units, respectively. Thus, this study begins to close this gap in the literature taking a closer look at the predictors of the test disengagement of adults from different countries. Examination of predictors provides the opportunity to obtain more detailed and appropriate results about the factors behind the disengagement of examines.

Purpose of the Study

The aim of this study was to examine the role of several item- and person-level variables on engaged responses in the domains of literacy and numeracy assessed in PIAAC 2012. Investigation of examines' responses on these domains is crucial since, in the most basic sense, the skills regarding numeracy and literacy contribute to the development of various high-level thinking skills, such as analytical thinking, understanding the information in a particular field. In particular, numeracy means more in everyday life than the mathematics we learn at school. In addition, the skills in these areas are used in many areas, from real life to education, business life, and communication with authorized persons (OECD, 2013c). Thus, in order to investigate examines' responses in terms of their engagement in tests requiring numeracy and literacy skills, the answers to the two related research questions were sought:

1. Which of the explanatory item response theory (EIRT) models (baseline model, a model with person characteristics, a model with the item characteristic, and a model with all person and item characteristics and the interaction between them) is best fitted to the PIAAC 2012 subdata?

2. To what extent does the engagement of adults in responding to items included in PIAAC 2012 be explained by person and item characteristics?

METHOD

Sample and Population

The target population of this study included all non-institutionalized adults between age 16 and 65 residing in the country at the time of data collection and participated in Round 1 of PIAAC 2012. In this study, the reason for the selection of countries participating in Round 1 is the high number of countries participated in this round and to increase the representation and generalizability of the results. Another reason for choosing Round 1 is that the t-disengagement rates of the countries participating in only this round are clearly examined in relation to each other in the official report (OECD, 2019), which ensures that the selection of data sets is based on evidence.

In PIAAC, probability sampling was used (OECD, 2013b). In the present study, countries were selected according to their rates of t-disengagement, which represents situations where a respondent spends less time than specified as an item-specific threshold (OECD, 2019). Therefore, in the term "t-disengagement", "t" stands for threshold. More precisely, the percentage of individuals showing t-disengagement in countries participating PIAAC 2012 varies between 8.4% and 33.4%. In the grouping of countries, the percentage of individuals showing t-disengagement in a country is compared to the average percentage of individuals showing t-disengagement in all countries participating in PIAAC 2012. For example, if the percentage of the individuals showing t-disengagement in a country is above the average percentage of t-disengagement, which is 15.70%, this country is classified as the country with a high percentage of t-disengagement. Accordingly, in addition to two countries such as France (21.50%) and Ireland (20.40%) with the highest percentage of individuals with t-disengagement, two countries such as Denmark (14.50%) and Germany (12.30%) where the percentage of individuals with t-disengagement is close to the average were selected. Also, three countries with the least percentage of individuals showing t-disengagement were selected (OECD, 2017) to represent better the pattern observed in the countries that participated in PIAAC 2012. From these examines, the ones who took the computer-based assessment of PIAAC 2012 were included as participants of this study. As a result, the sample of the current study includes 29959 adults from seven countries in total. Specifically, the frequency of these participants by the variables of the interest and countries were presented in Table 1.

Table 1. Frequency of Adults by Variable and Country

	Austria	*Denmark*	*Germany*	*France*	*Ireland*	*Norway*	*Finland*
Variables	*n(3830)*	*n (6048)*	*n(4510)*	*n(2758)*	*n (4058)*	*n(4292)*	*n(4463)*
Gender							
Male	1932	2942	2271	1372	1831	2942	2230
Female	1898	3106	2239	1386	2227	3106	2233
Highest level of schooling							
Less than high school	-	888	695	193	534	888	632
High school	-	2407	1899	1063	847	2407	1659
Above high school	-	2668	1876	1484	2656	2668	2156
Not definable	-	85	40	18	21	85	16
CBA Core score for stage 2							
3	51	104	75	64	75	104	65
4	249	378	296	216	283	378	246
5	927	1422	1120	817	1042	1422	1040
6	2603	4144	3019	1661	2658	4144	3112
Age in 10-year bands							
24 or less	825	965	1023	258	657	965	849
25-34	822	851	921	690	1113	851	995
35-44	899	1170	943	784	1217	1170	874
45-54	832	1182	1026	696	639	1182	908
55 plus	452	1880	597	330	432	1880	837
Index of readiness to learn							
All zero response	1	5	1		1	5	3
Lowest to 20%	463	374	563	135	428	374	167
More than 20% to 40%	840	1052	1232	437	790	1052	545
More than 40% to 60%	841	1348	1107	726	857	1348	967
More than 60% to 80%	829	1542	869	790	954	1542	1381
More than 80%	856	1727	738	670	1028	1727	1400
Index of use of ICT skills at work							
All zero response	153	188	200	162	113	188	188
Lowest to 20%	450	668	487	460	362	668	668
More than 20% to 40%	515	893	571	535	443	893	893
More than 40% to 60%	581	923	690	638	437	923	923
More than 60% to 80%	555	796	639	586	477	796	796
More than 80%	404	912	366	377	582	912	912
Valid skip	1172	1668	1557		1644	1668	1668
Index of use of ICT skills at home							
All zero response	19	10	19	3	10	4	117
Lowest to 20%	581	479	566	257	579	337	579
More than 20% to 40%	708	879	762	689	820	711	830
More than 40% to 60%	826	1290	1002	708	788	1045	708
More than 60% to 80%	848	1529	1101	630	766	1186	592
More than 80%	710	1742	893	471	763	938	428
Valid skip	138	119	167	-	332	71	1209

Data Collection Instruments

In PIAAC 2012, whether the surveys to be used as data collection tools will be applied in the computer environment or in the form of paper and pencil is determined according to the success of the respondents in two tests that measure their ICT skills. If the respondents fail to reach a certain level in the first stage, they will be redirected to the paper-based core section. Furthermore, if the respondents who were successful in the first task fail the subsequent short test, they only participate in the paper-based assessment. To participate in the computer-based assessments, the respondents must pass both tests.

The data collection instrument of the present study contained the literacy and numeracy surveys administered in the computer-based assessment of PIAAC 2012 (Round 1). Fifty-eight items were included in the literacy survey assessing adults' ability to read digital texts, as well as traditional print-based texts. Additionally, 56 items were included in the numeracy survey assessing the adults' ability to use, apply, interpret, and communicate mathematical information. For each domain, the distribution of items by context was presented in Table 2 (OECD, 2016a).

Table 2. Distribution of Items by Context

Survey	Context	Number	%
Literacy	Work	10	17
	Personal	29	50
	Community	13	23
	Education	6	10
	Total	58	100
Numeracy	Everyday life	25	45
	Work-related	13	23
	Society and community	14	25
	Further learning	4	7
	Total	56	100

In order to get evidence for the reliability of the test scores, how much variance is explained by the model for each cognitive domain was computed. Accordingly, reliability coefficients of the results obtained from literacy and numeracy domains range from .86 to .90 (OECD, 2013b). These values are found to be acceptable because they are more than .60, which is the minimum cut-off criteria in social sciences (Zikmund, Babin, Carr, & Griffin, 2010).

Explanatory item-level and individual-level variables

Studies (Bridgeman & Cline, 2000; Masters, Schnipke, & Connor, 2005; and Yang, O'Neill, & Kramer, 2002) examining the factors that have an influence on the time individuals spend on responding to a test item have considered item difficulty, item type, content area, degree of abstraction, etc. as an item level variable. However, in this study, since not all items and thus their characteristics are released by the OECD, only the item difficulty variable (OECD, 2013b) is considered the item-level variable as taken by the similar study of Goldhammer et al. (2017).

The cognitive pre-test is a kind of short test given to examinees to determine whether they are directed to full computer-based assessment of PIAAC. It includes three literacy and three numeracy items of low difficulty. If the examines failed from this test, they will be given the reading components of the assessment. On the other hand, if they achieve this test, they will take the full assessment (OECD, 2013b).

In PIAAC, there are several demographic variables regarding examinees. One of them is gender. More precisely, in this assessment, examinees are required to provide information about their gender. Also, there is an item which assesses examinees' age in 10-year bands such as 24 or less, 24-34, 35-44, 45-54, and over 55.   Another demographic variable assessed in PIAAC is educational attainment, which refers to the highest level of schooling. This categorical variable includes categories such as less than high school, high school, and above high school.

In PIAAC, examinees' readiness to learn is also measured. Specifically, there are six items focusing on the extent to which the examinees deal with problems and tasks they encounter. With these questions, they are asked how often they relate a new idea to the real-life situation and what they learned before, they are willing to learn something new, try to learn hard things in all details, and search for additional information to make it understandable when something they don't understand (Perry, Helmschrott, Konradt, & Maehler, 2017).

One of the variables measured in PIAAC is the use of ICT at work. There are a set of questions about the frequency of the use of computers or the internet as part of their job. More precisely, these questions focus on the use of e-mail, the internet for understanding job-related issues, conducting transactions on the internet, participating in real-time discussions on the internet, and the use of spreadsheets and word processing and the use of a programming language to program or write computer code. For measuring the use of ICT at home, the same questions were exposed to the examinees. However, this time these questions focus on the frequency of doing these activities in everyday life. All in all, examinees are divided into subcategories according to their frequency of using ICT at work, from those who use it least to those who use it most (OECD, 2015).

Data Analysis

The following procedure was followed to identify disengaged behaviors. If the time taken to respond to an item is below the threshold, it is considered that insufficient effort has been made for that item. To compute item-specific thresholds, the proportion correct greater than zero (P+>0%) method was used. Before seeking answers for the research questions, the time spent on the item was converted to a dichotomous engagement indicator (0 = disengaged, 1 = engaged) as an item response variable depending on whether the response time was below or above the response time thresholds. The variables cognitive pre-test score and item difficulty were centered and scaled to make a more meaningful interpretation of interaction effects.

Validity checks

In the present study, two validity checks were used to ensure that the threshold procedure employed accurately identified disengaged responses. In the first validity check, the engaged and disengaged response behaviors were compared in terms of their proportion correct (e.g., Wise & Kong, 2005; Wise & Ma, 2012). In order for the threshold determination process to be valid, the proportion correct for engaged behavior should be higher than the chance level, and the proportion correct for disengaged behavior should be at the level of chance. Considering that the items measuring verbal and numerical skills of adults in the PIAAC application have many response options, the probability of finding the correct answer by chance is very close to zero or zero. In the present study, the distributions of the observed proportion correct for responses classified as engaged or disengaged using the proportion correct conditional method (P+>0%) were examined for each domain and country. Accordingly, it was proven that the proportion correct for disengaged response behavior was found to be close to zero or zero, whereas the proportion correct for engaged response behavior was much higher. As an example, the distribution of the proportion correct scores of the engaged and disengaged individuals in Norway for each domain is presented in Figure 1.

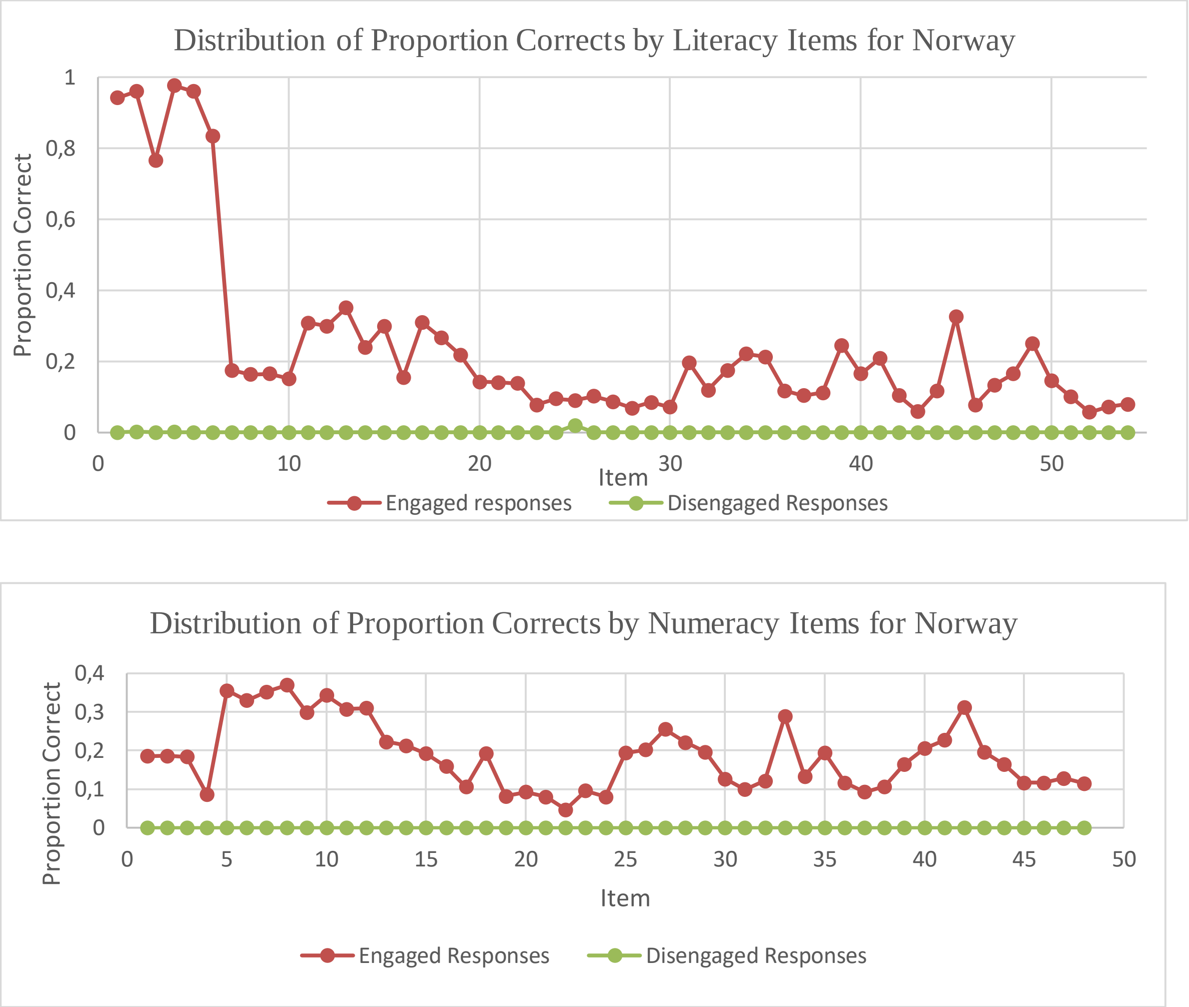


Figure 1. Distributions of the Proportion Correct Scores of Engaged and Disengaged Responses

In the upper part of Figure 1, the red line shows the proportion correct for engaged response behavior while the lower green line represents the corresponding proportion correct for disengaged response behavior. Figure 1 clearly shows that the proportion correct scores of the engaged individuals were higher than those of the disengaged individuals in Norway. A similar pattern was also observed in the other selected countries.

Another validity check for each item and domain was the examination of the association between the proficiency scores of individuals and the proportion correct of engaged and disengaged behaviors (e.g., Lee & Jia, 2014). According to the proficiency scores, individuals are divided into different groups referred to as score groups. In order for the threshold determination process to be valid, it is expected that there must be a positive relationship between the proportion correct and proficiency scores of the engaged responses for each item. No such relationship is expected for disengaged behaviors.

In the current study, the participants were divided into six score groups ranging from low competency to high competency as defined by PIAAC competency levels (OECD, 2013a) for both domains. Regardless of which plausible value is taken for examinees, individuals are at the same competency level defined by PIAAC. Furthermore, the plausible values were not used in the main analysis, but only as a proof of validity check. Therefore, in order to provide ease in calculations and interpretations, in assigning people to score groups, the mean of the adults' 10 plausible values regarding both domains was used. For each item, the relationship between the proficiency scores of the participants (i.e., an average of plausible values) and the proportion correct scores of engaged and disengaged response behaviors were investigated. Figure 2 shows the related findings for the selected literacy and numeracy items.

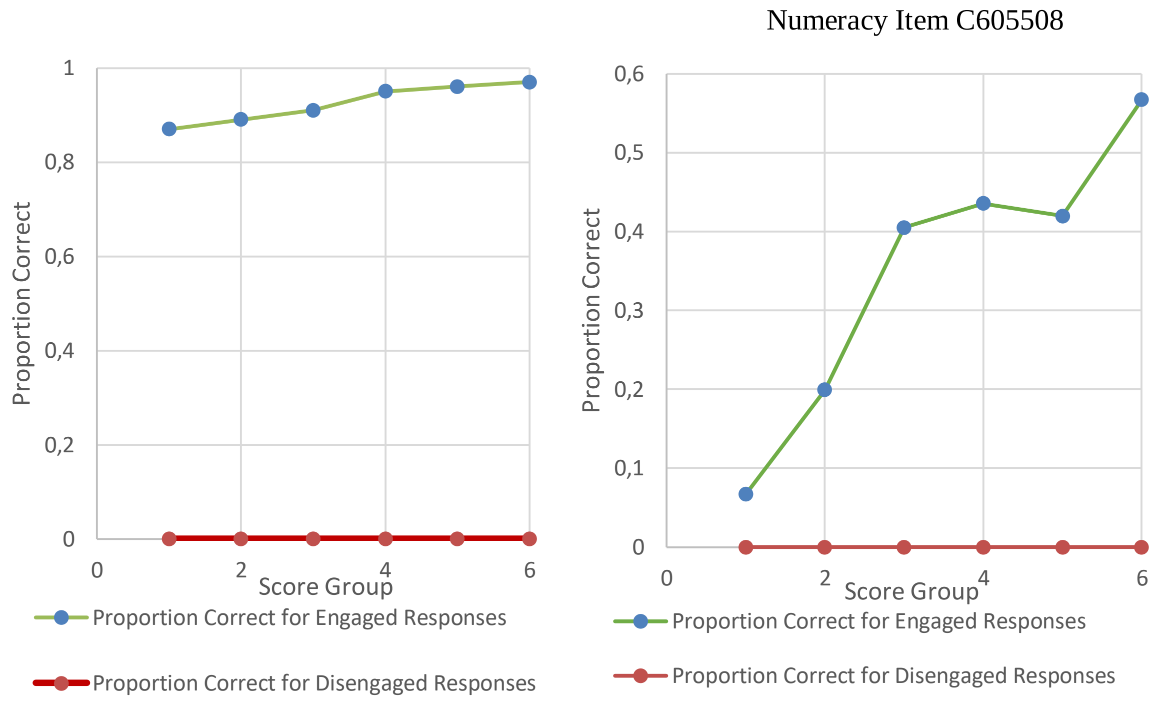


Figure 2. Association between the Score Groups and Proportion Correct Scores in Selected Literacy (C301C05) and Numeracy Items (C605508)

In both figures, the upper and lower lines show the association between the score groups and proportion correct scores for engaged and disengaged response behaviors, respectively. As expected, the association between the score group (plausible values) and the proportion correct for engaged response behavior was positive for all items in both domains.

Once the validity of the procedure for determining a threshold was proven, a 1-parameter logistic (1PL) item response model for each domain with dichotomous engagement indicators (0 = disengaged, 1 = engaged) was tested as an item response variable. 1PL models assume uni-dimensionality and equal discriminations across items. To determine the item fit, information-weighted (Infit) and unweighted (Outfit) mean-squared residual-based item fit statistics were inspected. If the infit and outfit values are between .5 and 1.5, it shows that the item fits the data (de Ayala, 2009). Thus, for each country and domain, very few items that did not fit the data were removed from the data set, which will not distort the representativeness of items. Specifically, for the countries Norway, Austria, Denmark, Germany, and Ireland, nine items were removed from the literacy survey, while seven items were removed from the numeracy survey. Furthermore, for Finland, three items were not included in the analysis of the responses to the literacy survey, while seven of the items were removed from the numeracy survey. Lastly, for France, the numbers of the items excluded from the data sets regarding the domain of literacy and numeracy were six and four, respectively.

Different EIRT models were constructed due to their flexibility to include the effect of the item and person-level variables simultaneously (Briggs, 2008). These models can be used for measurement and explanation purposes. The EIRM approach defines individuals as clusters, items as the repeated observations, and item responses as the dependent variable within a multilevel structure. In other words, the EIRT is of the multilevel models in which individuals' item responses are considered as the first-level factors, individuals are considered as second-level factors, and the individuals’ and/or items’ characteristics are included as predictors (De Boeck & Wilson, 2004).

Accordingly, after testing the baseline model, Model 0 and Model 1 with personal characteristics, such as educational attainment, gender, age group, cognitive skill, readiness to learn, and use of ICT skills at home and work were tested. Model 2 included only item difficulty since an item characteristic was being tested. Finally, the full Model 3 was tested with item- and person-level variables and the interaction of item difficulty with cognitive skill. After running all models, likelihood-based fit statistics, such as the likelihood-ratio (LL) statistics, Akaike’s information criterion (AIC), and the Bayesian information criterion (BIC), were determined. All models were estimated in the R environment (R Core Team, 2016). The TAM package (Kiefer, from the “lme4” package (Bates, Maechler, Bolker & Walker, 2015) was used to test explanatory item response models. The intra-class correlation (ICC) for each domain and country was computed to determine the proportion of variance in the dependent variable and the test-taking engagement that is attributed to personal differences. ICC is calculated by dividing the random effect variance by the total variance (Hox, 2002).

RESULTS

Model-Fit

For both literacy and numeracy domains, four explanatory IRT models were tested, and the LL, BIC, and AIC values were examined to determine the most appropriate IRT model for PIAAC 2012. There is no general rule about which model (the most complex or simpler) will fit the data. Therefore, in this study, although it was not predicted that Model 3 would definitely fit better before, it is predicted that item and individual-level variables may be effective on individuals' engagement levels. When the results were examined, it was found that Model 3 fitted the PIAAC 2012 data best because of the lower values of these indices. Therefore, the results of Model 3 were taken into consideration in this study. The model-fit results were presented in Table 3.

Table 3. Model-fit Results for Literacy and Numeracy Domains

		Literacy			*Numeracy*		
Country	*Model*	*AIC*	*BIC*	*LL*	*AIC*	*BIC*	*LL*
Austria	Model 0	172115.3	172145.7	-86054.6	172105.0	172135.5	-86049.5
	Model 1	172120.5	172394.4	-86033.3	172120.4	172394.2	-86033.2
	Model 2	170062.5	170123.3	-85025.2	169870.9	169931.8	-84929.4
	Model 3	169761	170065.3	-84850.5	169543.1	169847.4	-84741.6
Denmark	Model 0	265707.4	265739.2	-132851	264451.8	264483.6	-132222.9
	Model 1	265736.3	266054.2	-132838	264481.8	264799.8	-132210.9
	Model 2	262570.4	262634	-131279	261320.8	261384.4	-130654.4
	Model 3	261618.6	261968.3	-130776	260473.2	260823.0	-130203.6
Germany	Model 0	203498.3	203529.2	-101746	201059.9	201090.8	-100527.0
	Model 1	203523.5	203832.7	-101732	201082.9	201392.1	-100511.4
	Model 2	201159.3	201221.1	-100574	198410.1	198472.0	-99199.1
	Model 3	200581.1	200921.1	-100258	197927.6	198267.7	-98930.8
France	Model 0	166548.8	166578.4	-83271.4	166045.9	166075.7	-83020.0
	Model 1	166367.6	166634.2	-83156.8	165999.2	166267.3	-82972.6
	Model 2	166427.4	166476.8	-83208.7	157061.9	157111.3	-78526.0
	Model 3	165945.2	166231.5	-82943.6	157029.8	157316.1	-78485.9
Ireland	Model 0	177264.2	177294.8	-88629.1	181819.0	181849.6	-90906.5
	Model 1	177291.4	177587.2	-88616.7	181843.9	182170.3	-90889.9
	Model 2	174959.1	175020.3	-87473.6	179729.1	179790.3	-89858.6
	Model 3	174546.4	174883.0	-87240.2	179283.9	179620.5	-89608.9
Finland	Model 0	182698.3	182729.4	-91346.2	181819.0	181849.6	-90906.5
	Model 1	180477.6	180788.8	-90208.8	181843.8	182170.3	-90889.9
	Model 2	182685.4	182747.6	-91336.7	179729.1	179790.3	-89858.6
	Model 3	180465.1	180807.4	-90199.5	179283.9	179620.5	-89608.9
Norway	Model 0	185127.2	185158.0	-92560.6	189895.0	189925.8	-94944.5
	Model 1	185146.7	185454.4	-92543.4	189924.3	190232.0	-94932.2
	Model 2	182599.0	182660.5	-91293.5	187528.2	187589.7	-93758.1
	Model 3	182067.7	182406.2	-91000.9	186871.6	187210.1	-93402.8

Differences in Test Engagement

For each country, the results regarding the effects of the item- and person-level factors on test-taking engagement are presented in Tables 4 and 5.

Table 4. Results regarding Items Assessing Literacy

Variables	Subcategory	Austria	Finland	Norway	Denmark	Germany	France	Ireland
Intercept		-13.9**	-20.86**	-1.08**	-1.13*	-12.98**	-.09**	-10.73**
Difficulty		-	-2.68**	-			-.83**	-
Cognitive pre-test x difficulty		.11**		.15**	.16**	.14**	.16**	.19**
Age	35-44	-	-	-	-		-.07**	-
	45-54	-	-	-	-		-.05**	-
	Over 55	-	-	-.23'	-		-.10**	-
Educational attainment	Above high school	-	.35'	-	-	.22'	-	-
Readiness to learn	Lowest to 20%	12.43**	-3.34**	-	-	12.25**	-	9.45**
	More than 20% to 40%	12.41**	-3.14**	-	-	12.36**	.05'	9.41**
	More than 40% to 60%	12.36**	-2.60**		-	12.2**	.05'	9.54**
	More than 60% to 80%	12.21**	-2.48**	-	-	12.22**	-	9.59**
	More than 80%	12.44**	-2.62**	-	-	12.32**	-	9.59**
Use of ICT at home	lowest to 20%	1.12**	-	-	-	-	-	
	More than 20% to 40%	1.1**	-	-	-	-	-	.78'
	More than 40% to 60%	1.09**	-	-	-	-	-	-
	More than 60% to 80%	1.27**	-	-	-	-	-	-
	More than 80%	1.22**	-.81'	-	-		-	-
Use of ICT at work	lowest to 20%	-	18.16**	-	-	-		-
	More than 20% to 40%	-	23.42**	-	-	-	-	-
	More than 40% to 60%	-	23.42**	-	-	-	-	-
	More than 60% to 80%	-	23.24**	-	-		-	-
	More than 80%		23.06**	-			-	-
ICC		.49	.48	.50	.49	.48	.50	.50

** $p < .001$, *$p < .01$, ' $p < .05$

Table 5. Results regarding Items Assessing Numeracy

Variables	Subcategory	Austria	Finland	Norway	Denmark	Germany	France	Ireland
Intercept		-	-1.84**	-3.06**	-1.72**	-1.46*	-	-1.87**
Difficulty	-	-	-	-	-	-	-	-
Cognitive pretest	-	.09'	.07*	-	.08'	-	.03**	.07*
Cognitive pre-test x difficulty		.39**	.13**	-	.51**	.43**	-	.13**
Age	25-34	-	-	.21'	-	-		-
	35-44	-	-	-	-	-	-.06*	-
	45-54	-	-.25*	-	-	-	-.04'	-.24*
	Over 55	-		-			-.09**	
Gender	Female		-	.15*	-	-	-	-
Readiness to learn	lowest to 20%	-	9.92**	-	-	-	-	9.95**
	More than 20% to 40%	-	9.82**	-	-	-	.05'	9.84**
	More than 40% to 60%	-	9.80**	-	-	-	.05'	9.83**
	More than 60% to 80%	-	9.80**	-	-	-	.05'	9.83**
	More than 80%	-	9.81**	-	-	-	-	9.84**
Use of ICT at home	lowest to 20%	-	-	1.28*	-	-	-	-
	More than 20% to 40%	-	-.79'	1.42**	-	-	-	-.79'
	More than 40% to 60%	-	-	1.43**	-	-	-	-
	More than 60% to 80%	-	-	1.47**	-	-	-	-
	More than 80%	-	-.86'	1.4*	-	-	-	-.86'
Use of ICT at work	lowest to 20%	-.35*	-	-	-	-	-	-
ICC		.49	.50	.50	.51	.48	.50	.49

** $p < .001$, *$p < .01$, ' $p < .05$

As shown in Table 4, the difficulty of items measuring literacy had a negative effect on the engagement of participants in France (-.93) and Finland (-2.68), showing that when the item difficulty increased, adults tended not to give sufficient time to the items. On the other hand, the difficulty of items measuring numeracy was found to have no significant effect on the engagement of the adults. In addition to the main effect of item difficulty on engagement, the interaction between item difficulty and cognitive skill was also significant. Specifically, the effect of item difficulty on engagement was higher among strong test-takers who put more effort into solving items than poor test-takers who did not put sufficient effort into items.

Age had a statistically significant on the engagement of participants in literacy items in France and Norway. Specifically, as the age of the French participants increased, they tended to be disengaged. Additionally, there was a particularly strong decrease in the engagement rate of the oldest group, participants aged 55 or above in Norway (-.23). A similar pattern was also found for the domain of numeracy. Moreover, the significant negative effect of age on the engagement of the adults taking the numeracy items was observed in the countries of Ireland and Finland.

The highest level of educational attainment was associated with higher engagement in Germany (.22) and Finland (.35). In other words, individuals with a high level of education in Germany spent more time answering the questions. When the results were examined in terms of the numeracy domain, as shown in Table 5, it was found that educational attainment had no significant effect on engagement.

As it was clearly seen in Table 4, for Austria, Germany, France, and Ireland, the adults' readiness to engage in learning activities had a positive effect on their engagement on items addressing literacy skills. However, this was not the case for the participants from Finland. The adults' readiness to engage in learning activities which require the use of literacy skills had a negative effect on their engagement. The finding was that the adults who were highly ready to learn put insufficient effort into answering the items. For the domain of numeracy, as s presented in Table 5, a similar pattern observed for literacy domain was also found in France and Ireland in terms of the effect of adults' readiness to learn on their engagement levels. That is, as the level of readiness to learn of the adults increased, their test-engagement levels also increased when responding to the items assessing numeracy items.

For the literacy domain, Table 5 shows that the effect of the use of ICT skills at home of individuals from each category in Austria on their engagement levels was positive and significant, suggesting that the test-takers who more frequently used ICT skills at home exhibited a higher level of engagement. In contrast, the use of ICT skills at home was negatively associated with the adults' engagement in numeracy in Ireland (-.79) and Finland (-.79), but the use of ICT skills at home for each category of the individuals in Norway was positively related to the students' engagement in numeracy.

When the effect of the use of the ICT skills of individuals at work was examined across all countries, according to Table 4, it was found that in Finland, those who more frequently used ICT skills at work tended to be more engaged while responding to the items measuring literacy. On the other hand, this was not the case for the field of numeracy. A negative and significant effect (-.35) of the use of ICT skills at work on the engagement of individuals in Austria was found, suggesting that the adults who used ICT skills frequently at work tended to be disengaged when answering the items in the test. When the findings regarding gender were considered, it was determined that for only the field of numeracy, in Norway, being female (.15) was found to be positively related to test-taking engagement.

For each country and domain, as presented in Tables 4 and 5, the ICC values taking into account the adults' test-taking engagement differences at the person level were found to be similar to each other. Specifically, approximately 50% of the variation in engagement levels of individuals was attributable to differences between subjects.

DISCUSSION and CONCLUSION

This study aimed to determine which of the explanatory IRT model was the best fit for the analysis of the PIAAC sub-data. In addition, the present study aimed to investigate the effect of person- and item-level factors depending on the analysis of the model that best fitted the data. To achieve these aims, predictions were created utilizing different models for the domains of literacy and numeracy.

The conclusion of this study is that there is increasing disengagement in more difficult items measuring literacy skills, thus indicating that individuals spend little time on very difficult items (OECD, 2013a). When individuals perceive an item to be very difficult, they may tend to stop trying to understand and respond to the item very quickly. Considering that the data in this study belonged to the low stake assessment, the low motivation of the participants may have played a role in this outcome. Furthermore, whether a particular item is perceived as 'too difficult' depends on the cognitive level of the adult. The reason behind this finding is that there is a significant and positive effect of the interaction between cognitive pre-test and item difficulty on test engagement (Wise & Kingsbury, 2015). In other words, the significant effect of the interaction between item difficulty and cognitive pre-test shows that individuals tend to engage in relation to their cognitive skills.

Older adults tend to exhibit a higher propensity to disengage in both fields. Increasing disengagement by older test-takers in items in technology-rich environments may be related to their lower levels of ICT experience and skills (OECD, 2013a). They have more difficulty than their younger counterparts in using computers due to age-associated changes in visual, perceptual, psychomotor, and cognitive abilities. Older people with insufficient experience with computers may also have a negative attitude toward computer usage (Xie, 2003), which may cause disengaged behaviors in testing.

Additionally, the present study revealed that more educated individuals were more engaged in the items assessing literacy. This finding is supported by the study of Goldhammer, Martens, Christoph, and Lüdtke (2016), in which the effect of educational attainment on the individual' disengagement was investigated. There may be several reasons for this result. Firstly, compared to individuals who are less educated, highly educated individuals are relatively more proficient and more likely to respond to more difficult items. Secondly, since those with higher education are more accustomed to testing and assessment environments; thus, they may get less tired than test takers with lower education levels. As a result, the former do not stop trying to give an answer to an item. Lastly, people with a high level of education may have a stronger sense of commitment to completing the assessment, which makes them put more effort into solving the items. Those people with a low level of education may have difficulty in understanding the items. They may not have sufficient literacy and numeracy skills (OECD, 2019), which can result in a tendency to respond to items quickly.

Individuals who are more ready to learn tend to exhibit more engagement in the items. The reason behind these results might be related to the composite feature of the readiness to learn, which consists of attitudinal or emotional, cognitive, behavioral, and, to a lesser extent, personality or dispositional components (Smith, Rose, Ross-Gordon & Smith, 2015). Therefore, individuals who are more ready to learn are more attentive, willing, and motivated to learn. Thus, they can easily concentrate on the items and complete them without getting bored (Eccles & Wigfield, 2002).

The current study concluded that adults who frequently used ICT skills at home and work engaged more than the adults that rarely used ICT skills. This finding is in line with the literature that suggests individuals with strong ICT skills engage more in a technology-enriched environment (Bergdahl, Nouri & Fors, 2019). This can be explained by familiarity with ICT which has an effect on the motivation and engagement of individuals (OECD, 2019).

It is concluded that gender has a significant effect on adults' engagement in items assessing numeracy skills, suggesting that engagement can be seen as a domain-specific construct (Goldhammer et al., 2016); for example, in Norway, females exhibit a higher level of engagement. This finding is also supported by the study of Marrs and Sigler (2012). They found that females tended to engage in the material at a deeper level, whereas males tended to display minimal effort.

Interpreting the results regarding literacy obtained from this study in terms of country groups according to t-disengagement percentages shows that the use of ICT skill had no effect, except for the test-taking engagements of countries with a low t-disengagement percentage. On the other hand, for the numeracy domain, there were several similarities in the effect of person-level factors on the same country groups. For example, the effect of age and readiness to learn on countries with a high t-disengagement percentage was similar. For the numeracy domain, age had a negative effect on test-taking engagement for adults in both France and Ireland, whereas readiness to learn had a positive effect. Additionally, it was concluded that some personal-level variables (age, gender, readiness to learn, and use of ICT skills at home and work) did not have an effect on the test-taking engagement of countries with a relatively moderate t-disengagement percentage.

To make more accurate evaluations, it is suggested that assessment practitioners should manage disengagement by identifying disengaged responses when obtaining test scores and filtering such responses in the data. Additionally, adults can be provided with valuable feedback regarding their performance (DeMars et al., 2013). One or more of these methods can be used for the validity of the results obtained from low-stake assessments. Underestimating disengaged responses may have significant negative consequences due to the potential high-stakes nature of international assessments for educational stakeholders and policymakers. By demonstrating the differential predictors of disengaged responses by country, this study revealed the potential for educational stakeholders to make inaccurate inferences when comparing subgroup performance across countries. For example, when comparing performance by gender, it is possible that score differences observed between males and females across countries may be confused with test-taking effort as opposed to true differences. Since such effects may be investigated as a basis for constructing national education policy reform, it is crucial that disengaged responses are identified and filtered before performing operational analyses (e.g., item analyses) and research analyses (Rios & Guo, 2020). These recommendations are some examples of how the results of this new study can be used and how they can benefit practitioners. However, in any case, the most important message that can be derived from this study is that the source of the differences in the scores of individuals in low-stake assessments may be their disengagement levels. Future research can be conducted to explore the extent to which these factors developed in recent years are effective in disengagement under low-stakes conditions.

The findings from this study offer practical uses; however, they are limited in a number of ways. Firstly, in this study, a selection was made from countries with different levels of disengagement, but not all countries participating in PIAAC 2012 were included. The findings of the present study cannot be generalized to adults; thus, further similar research is required. Secondly, this study used only one method to determine response time thresholds. Since there are many other methods to detect disengaged behaviors, future research can be conducted to compare the effectiveness of these methods. Despite the limitations of this study, it is considered that it draws further attention to the role of test-taking effort in international assessments and contributes to the discussion of investigating test-takers' effort as part of standard operational practices.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, & Joint Committee on Standards for Educational and

Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling, 55*(1), 92–104.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48

Bergdahl, N., Nouri, J., & Fors, U. (2019). Disengagement, engagement and digital skills in technology-enhanced learning. *Education and Information Technologies*, 149. doi:10.1007/s10639-019-09998-w.

Braun, H., Kirsch, I., Yamamoto, K., Park, J., & Eagan, M. K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP reading assessment. *Teachers College Record, 113*(11), 2309–2344.

Briggs, D. C. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education, 21*(2), 89-118. doi: 10.1080/08957340801926086.

Bridgeman, B., & Cline, F. (2000). *Variations in mean response time for questions on the computer-adaptive GRE General Test: Implications for fair assessment.* GRE Board Professional Report No. 96-20P. Princeton, NJ: Educational Testing Service.

Brown, G. T. L., & Harris, L. R. (2016). *Handbook of human and social conditions in assessment*. New York, NY: Routledge.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press

Demars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment, 12*(1), 23–45. doi:10.1080/10627190709336946

DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research and Practices in Assessment, 8*, 69-82.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*(1), 109–132. doi:10.1146/annurev.psych.53.100901.135153.

Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report Series, 2015(2),*1–17. doi: 10.1002/ets2.12067

Goldhammer, F., Martens, T., & Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC: an exploratory IRT modelling approach considering person and item characteristics. *Large-scale Assessments in Education, 5*(18), 1-25. doi: 10.1186/s40536-017-0051-9.

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC.* Vol. 133. In: OECD Education Working Papers. Paris: OECD Publishing.

Hox J. 2002. *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.

Kiefer, T., Robitzsch, A., & Wu, M. (2016). *TAM: Test analysis modules*. R package version 1.99–6. Retrieved from http:// CRAN.R-project.org/package=TAM

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement, 67*(4), 606–619.

Lee, Y.-H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-scale Assessments in Education, 2*(1), 1–24. doi: 10.1186/s40536-014-0008-1.

Marrs, H., & Sigler, E. A. (2012). Male academic performance in college: The possible role of study strategies. *Psychology of Men & Masculinity, 13*(2), 227-241.

Masters, J., Schnipke, D. L., & Connor, C. (2005, April). *Comparing item response times and difficulty for calculation items*. Paper presented at the annual meeting of the American Educational Research Association, Montréal, Canada.

Mastuti, E., & Handoyo, S. (2017, October). *Effects of individual differences on the performance in computer-based test (CBT).* Paper presented at the 3rd ASEAN Conference on Psychology, Counselling, and Humanities (ACPCH). Malang, Indonesia.

Nagy, G., Nagengast, B., Becker, M., Rose, N., & Frey, A. (2018). Item position effects in a reading comprehension test: an IRT study of individual differences and individual correlates. *Psychological Test and Assessment Modeling, 60*(2), 165–187.

Organisation for Economic Co-operation and Development. (2013a). *OECD skills outlook 2013: First results from the survey of adult skills.* Paris: OECD Publishing.

Organisation for Economic Co-operation and Development. (2013b), "The methodology of the Survey of Adult Skills (PIAAC) and the quality of data", in *The*
Survey of Adult Skills: Reader's Companion, OECD Publishing, Paris.

Organisation for Economic Co-operation and Development. (2013c). *What the Survey of Adult Skills (PIAAC) measures in The Survey of Adult Skills: Reader's Companion*, OECD Publishing, Paris. doi: https://doi.org/10.1787/9789264204027-4-en

Organisation for Economic Co-operation and Development. (2015). *Adults, Computers and Problem Solving: What's the Problem?, OECD Skills Studies*, OECD Publishing, Paris, https://doi.org/10.1787/9789264236844-en.

Organisation for Economic Co-operation and Development. (2016a).Technical report of the Survey of Adult Skills (PIAAC) (2nd edition). OECD, Paris,

Organisation for Economic Co-operation and Development. (2016b). *The Survey of Adult Skills: Reader's Companion*, Second Edition, OECD Skills Studies, OECD Publishing, Paris. doi: 10.1787/9789264258075-en.

Organisation for Economic Co-operation and Development. (2017). *Programme for the International Assessment of Adult Competencies (PIAAC), Log Files*, GESIS Data Archive, Cologne, doi:10.4232/1.12955.

Organisation for Economic Co-operation and Development. (2019), *Beyond Proficiency: Using Log Files to Understand Respondent Behaviour in the Survey of Adult Skills*, OECD Skills Studies, OECD Publishing, Paris. doi: 10.1787/0b1414ed-en.

Perry, A., Helmschrott, S., Konradt, I., & Maehler, D. B. (2017). *User Guide for the German PIAAC Scientific Use File*: Version II. (GESIS Papers, 2017/23). Köln: GESIS - Leibniz-Institut für Sozialwissenschaften. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-54438-v2-7

Rios, J. A., & Guo, H. (2020). Can culture be a salient predictor of test-taking engagement? An analysis of differential noneffortful responding on an international college-level assessment of critical thinking. *Applied Measurement in Education*. Advance online publication. doi: 10.1080/08957347.2020.1789141

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated scores: To filter unmotivated examinees or not? *International Journal of Testing, 17*(1),74-104. doi: 10.1080/15305058.2016.1231193

Smith, M C., Rose, A.D., Smith, T. J.& Ross-Gordon, J. M. (2015, May). *Adults' readiness to learn and skill acquisition and use: An analysis of PIAAC*. Paper presented at the 56th Annual Adult Education Research Conference. Manhattan, KS.

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237-266). Mahwah, NJ: Lawrence Erlbaum Associates.

Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a largescale assessment. *Applied Measurement in Education, 26*(1), 34–49. doi: 10.1080/08957347.2013.739453

Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology, 29*(1), 6–26. doi:10.1016/S0361-476X(02)00063-2.

Team, R. C. (2016). *R: A language and environment for statistical computing* (Version 3.1.3). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/.

Xie, B. (2003). Older adults, computers, and the Internet: Future directions. *Gerontechnology,2*(4), 289-305.

van Barnevald, C. (2007). The effect of examinee motivation on test construction within an IRT framework. *Applied Psychological Measurement, 31*(1), 31–46. doi:10.1177/0146621606286206

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes, computer-based test. *Applied Measurement in Education, 19*(2), 25-114.

Wise, S. L. (2009). Strategies for managing the problem of unmotivated examinees in low-stakes testing programs. *The Journal of General Education, 58*(3), 152–166. doi:10.1353/jge.0.0042

Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education, 28*(3), 237–252. doi:10.1080/08957347.2015.1042155

Wise, S. L. , & DeMars, C. E. (2009). A Clarification of the effects of rapid guessing on Coefficient α: A note on Attali's "reliability of speeded number-right multiple-choice tests". *Applied Psychological Measurement , 33*(6), 488–490. doi:10.1177/0146621607304655

Wise S. L. & DeMars C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment, 15*(1), 27-41.

Wise, S. L., & Kingsbury, G. G. (2015). *Modeling student test-taking motivation in the context of an adaptive achievement test.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183. doi:10.1207/s15324818ame1802_2

Yang, C. L., O'Neill, T. R., & Kramer, G. A. (2002). Examining item difficulty and response time on perceptual ability test items. *Journal of Applied Measurement, 3*(3), 282-299.

Zikmund, W.G., Babin, B.J., Carr, J.C. and Griffin, M. (2010). *Business research methods,* Canada:South-Western Cengage Learning.

PIAAC 2012'de Test Katılımının Sessiz Yordayıcıları

Giriş

Bireylerin düşük riskli uluslararası değerlendirmelerde güdülerinin düşük olması testteki maddeleri cevaplamaya yeterince zaman ayırmamalarına neden olmaktadır (Wise ve DeMars, 2010). Bu durum testin psikometrik özelliklerin bozulmasına (Rios, Guo, Mao, & Liu, 2017)v e veri setinden elde edilen sonuçların yanlış yorumlanmasına yol açmaktadır (Nagy, Nagengast, Becker, Rose ve Frey, 2018). Daha doğrusu, bireylerin gerçek puanlarına, teste katılım seviyelerine bağlı olarak sistematik bir hata karışmaktadır (Braun, Kirsch, Yamamoto, Park ve Eagan, 2011). Bunun yanı sıra, bireylerin teste yeterince zaman ayırmamaları (a) madde güçlük ve ayırıcılık parametrelerinin olduğundan daha yüksek (van Barnevald, 2007) (b) madde ve test bilgi fonksiyonlarının yanlı olarak (van Barnevald, 2007), (c) klasik test teorisine dayalı güvenilirlik tahminlerinin olduğundan yüksek (Wise & DeMars, 2009), (d) değişen madde fonksiyonun yanlış (Wise & DeMars, 2010) ve (e) değişkenler arası korelasyonların daha düşük (Wise, 2009) kestirilmesine neden olmaktadır.

Bireylerin testteki maddelere yeterince zaman ayırmamasının nedeni, bu davranışın doğru ve verimli bir şekilde ölçülmesi, testteki maddelere yeterince zaman ayırmamadan kaynaklanan büyük geçerlilik sorunları göz önüne alındığında çok önemlidir. Geniş ölçekli uygulamalardan biri olan PIAAC değerlendirmesine dâhil edilen maddeler zorluk ve karmaşıklık açısından farklılık gösterdiğinden, doğru cevabı vermek için gereken süre birbirinden farklı olacaktır. Bu nedenle, avantajları göz önünde bulundurularak, bu çalışmada testteki maddelere katılım gösteren ve göstermeyen davranışları belirlemede maddeye özgü tepki süresi eşikleri kullanılmıştır (Wise, 2006).

Bireylerin testteki madde üzerinde harcadıkları zaman konusunda kapsamlı araştırmalar yapılmış olsa da, bu çabaların çoğu tek bir ülkede bulunan nispeten homojen popülasyonlara odaklanmıştır (Goldhammer, Martens & Lüdtke, 2017). Kişisel özellikler kültüre veya ülkeye göre büyük ölçüde farklılık gösterse de (Brown ve Harris, 2016) uluslararası değerlendirmelerde ülkeler arasında ülkelerin teste harcadıkları zaman açısından potansiyel farklılıkları inceleyen çok az çalışma yapılmıştır (Rios ve Guo, 2020). Genel olarak, bu çalışma, farklı ülkelerden yetişkinlerin katılımını etkileyen faktörleri daha yakından inceleyerek alan yazındaki bu boşluğu kapatmaya katkıda bulunmaktadır. Bu bağlamda, bu çalışma, PIAAC uygulamasında ele alınan sözel ve sayısal becerilerle ilgili alanlara ilişkin maddelere harcanan zaman üzerindeki çeşitli madde ve birey düzeyindeki değişkenlerinin rolünü incelemeyi amaçlamaktadır. Bu doğrultuda, bu çalışmada cevap aranan araştırma soruları şu şekildedir:

1) Açımlayıcı madde tepki modellerinden hangisi (temel model, birey düzeyindeki değişkenlerinin dâhil edildiği model, madde düzeyindeki değişkenin dâhil edildiği model ve bütün madde ve birey düzeyindeki değişkenlerin ve bunlar arasındaki etkileşimin dâhil edildiği model) PIAAC alt verilerine en iyi uyumu sağlamaktadır?

2) Maddelere katılım gösteren yanıtlar birey ve madde düzeyindeki değişkenlerle açıklanabilir mi?

Yöntem

Çalışmanın hedef evreni veri toplama sırasında ülkede ikamet eden ve PIAAC 2012'ye katılan 16 ila 65 yaşları arasındaki yetişkinler içermektedir. Olasılıklı örnekleme yöntemi kullanılmıştır. Çalışmanın örneklemini teste katılmama düzeylerine seçilen ülkeler oluşturmaktadır. Buna göre, katılmama düzeyi yüksek olan ülke grubundan iki ülke (Fransa ve İrlanda), orta olan iki ülke (Danimarka ve Almanya) ve düşük olan üç ülke (Avusturya, Finlandiya ve Norveç) çalışmaya dâhil edilmiştir.

Çalışmada veri toplama aracı olarak sözel ve sayısal becerileri ölçen anketler kullanılmıştır. Bilgisayar tabanlı değerlendirmeye katılan yetişkinlerin dijital metinleri okuma becerilerinin yanı sıra geleneksel basılı metinleri de değerlendiren sözel becerileri ölçen ankette 58 madde dâhil edilmiştir. Ek olarak, yetişkinlerin matematiksel bilgileri kullanma, uygulama, yorumlama ve iletme yeteneklerini değerlendiren sayısal becerileri ölçen ankette 56 madde dâhil edilmiştir (OECD, 2016).

Maddelere özgü eşik parametrelerini belirlemek için sıfırdan büyük doğru cevaplama oranı (P +>% 0) yöntemi kullanılmıştır. Araştırma sorularına cevap aramadan önce ikili puanlanan yeni bir değişken tanımlanmıştır. Buna göre, maddeye harcanan zaman, madde eşik parametresinin altında veya üstünde bir değer almasına göre yeniden kodlanmıştır (0= katılım göstermemiş, 1= katılım göstermiş). Madde güçlük parametreleri kestirimlerin kolaylaşması açısından 100'e bölünerek yeniden ölçeklendirilmiştir. Etkileşim etkisini belirlemek için açıklayıcı madde tepki modellerinin analizleri sırasında bilişsel ön test puanları ve madde güçlük parametreleri ölçeklendirilmiştir.

Madde eşik parametrelerinin belirlenmesi sürecinin testteki maddelere yeteri kadar katılım göstermeyen ve gösteren bireyleri doğru bir şekilde ayırıp ayırmadığını belirlemek için iki tane geçerlik kontrolü yapılmıştır. Birinci geçerlik kontrolünde, doğru cevaplama oranları katılım göstermiş ve göstermemiş bireyler açısından karşılaştırılmıştır. Geçerli bir belirleme sürecinde, PIAAC uygulamasındaki maddelerin çok sayıda tepki seçeneklerinin olduğu düşünüldüğümde katılım gösteren bireylerin doğru cevaplama oranlarının dağılımı sıfırdan büyük iken katılım göstermeyen bireylerin doğru cevaplama oranlarının dağılımının sıfır veya sıfıra çok yakın olması beklenir. Bu çalışmada da bu durum doğrulanmıştır. Bir diğer geçerlik kanıtı olarak ise farklı yeterlik gruplarında katılım göstermemiş ve göstermiş bireylerin doğru cevaplama oranları karşılaştırılmıştır. Maddelere katılım gösteren bireyler için yeterlik puanları ile doğru cevaplama oranları arasında pozitif yönde ilişki çıkması beklenirken katılım göstermeyen bireyler için manidar bir ilişkinin çıkması beklenmez. Bu çalışma da bu durum doğrulanmıştır.

Her bir alan için 1-parametreli lojistik modeller bireylerin katılım düzeylerini gösteren ve yeniden oluşturulan ikili puanlanan değişkenin varlığında test edilmiştir. Modele uyum sağlamayan maddeler veri setinden çıkarılmıştır. Dört farklı açıklayıcı madde tepki kuramı modeli madde ve birey düzeyindeki değişkenlerin etkilerini aynı anda incelenmesini sağlaması nedeniyle test edilmiştir. Veriye uyum sağlayan modelin belirlenmesinde çeşitli uyum iyiliği indekslerinden yararlanılmıştır. Verilerin analizinde tek boyutluluğu belirlemede R yazılımında "TAM" paketi (Kiefer et al., 2016) ve açıklayıcı madde tepki kuramı modellerinin analizinde ise “lme4” paketi (Bates et al., 2015) kullanılmıştır. Yetişkinlerin maddelere katılım düzeylerindeki varyansı açıklamada bireyler arası farklılıkların etkisini belirlemek için sınıf içi korelasyon katsayıları her bir ülke ve her bir alan için hesaplanmıştır.

Sonuç ve Tartışma

Veriye en iyi uyum sağlayan modelin hem madde düzeyinde hem de birey düzeyindeki değişkenlerin dâhil edildiği model olduğu bulunmuştur. Bu çalışmada, bireylerin çok zor maddelere çok az zaman ayırdıkları sonucuna ulaşılmıştır (OECD, 2013). Kişiler maddenin çok zor olduğunu algıladıklarında, denemeyi bırakıp maddeye çok çabuk cevap verme eğiliminde olabilirler. Bu çalışmadaki verilerin düşük riskli bir değerlendirmeye ait olduğu düşünüldüğünde, bu durumda katılımcıların düşük motivasyonu rol oynamış olabilir. Ayrıca belirli bir maddenin “çok zor” olarak algılanıp algılanmaması yetişkinlerin bilişsel düzeyine bağlıdır. Bu durum, bu çalışmadan elde edilen sonuçlardan biri olan bilişsel ön test ile madde zorluğu arasındaki etkileşimin test katılımı üzerinde anlamlı ve olumlu bir etkisinin olmasıyla da desteklenmektedir (Wise ve Kingsbury, 2015).

Daha yaşlı yetişkinlerin her iki alanda da daha yüksek düzeyde katılmama eğilimi gösterdikleri sonucuna varılmıştır. Teknoloji açısından zengin ortamlardaki değerlendirmelerde nispeten yaşı büyük olan katılımcılarının artan ilgisizliği, teknolojiyle ilgili deneyim ve becerilerinin daha düşük olmasıyla açıklanabilir (OECD, 2013). Bu yüzden özellikle bilgisayar kullanıma yönelik olumsuz tutuma sahip olabilir (Xie, 2003). Bu durum ise onların testteki maddelere yeteri düzeyde katılmamalarına neden olabilir.

Ayrıca bu çalışma, daha eğitimli bireylerin sözel becerileri değerlendiren maddelere daha fazla zaman harcadıklarını ortaya çıkarmıştır. Bu bulgu, Goldhammer, Martens, Christoph ve Lüdtke'nin (2016) eğitim düzeyinin bireyin testteki maddelere katılmamaları üzerindeki etkisinin araştırıldığı çalışmasıyla desteklenmektedir. Bu sonucun birkaç nedeni olabilir. İlk olarak, eğitim düzeyi yüksek olan bireyler, eğitim düzeyi düşük olan bireylere göre görece daha yetkin olduklarından, onlardan daha zor maddelere cevap vermeleri istenebilir. İkinci olarak, test ve değerlendirme ortamlarına daha alışkın oldukları için diğer katılımcılara göre daha az yorulabilirler. Sonuç olarak, maddeye cevap vermeye çalışmaktan vazgeçmeme eğilimi gösterebilir.

Öğrenmeye daha hazır olan bireyler, maddeleri cevaplamada yeterince zaman harcamaktadırlar. Bu durum, öğrenmeye daha hazır olan bireylerin daha dikkatli, daha istekli ve öğrenmeye güdülü olmasıyla açıklanabilir. Böylece maddeler üzerinde kolayca odaklanabilir ve sıkılmadan tamamlayabilirler (Eccles ve Wigfield, 2002).Mevcut çalışmada, BİT becerilerini evde ve işte sıklıkla kullanan yetişkinlerin, BİT becerilerini nadiren kullanan yetişkinlere kıyasla testte yer alan maddeleri cevaplamada yeterince zaman harcadıkları sonucuna varılmıştır. Bu bulgu, yüksek düzeyde BİT becerilerine sahip bireylerin teknolojiyle zenginleştirilmiş ortamlarda daha fazla katıldıklarını belirten alan yazınla paralellik göstermektedir. (Bergdahl, Nouri & Fors, 2019). Bu, bireylerin güdüsü ve katılımı üzerinde etkisi olan BİT'e olan aşinalık ile açıklanabilir (OECD, 2019).

Bu çalışmada cinsiyetin, yetişkinlerin sayısal becerilerini değerlendiren maddelere katılımı üzerinde önemli bir etkisinin olduğu sonucuna varılmıştır ve bu, maddelere katılımın alana özgü bir yapı olduğunu göstermektedir (Goldhammer, Martens & Lüdtke, 2016). Daha açık olarak belirtmek gerekirse, Norveç'teki kadınlar maddelere cevap vermede daha yüksek düzeyde katılım sergilemektedir. Bu bulgu, Marrs ve Sigler'in (2012), kadınların kendilerine verilen göreve daha yüksek düzeyde katılma, erkeklerin ise minimum çaba gösterme eğiliminde olduğunu belirten çalışmasıyla uyumludur.

Daha doğru değerlendirmeler yapmak için, uygulayıcılar test puanlarını hesaplarken ve verilerdeki bu tür yanıtları belirleyerek filtreleyebilir. Ayrıca yetişkinlere performanslarıyla ilgili değerli geri bildirimler de sunulabilir (DeMars, Bashkov & Socha, 2013). Düşük riskli değerlendirmelerden elde edilen sonuçların geçerliliği için bu yöntemlerden bir veya daha fazlası kullanılabilir. Bununla birlikte, her durumda, bu çalışmadan çıkarılabilecek en önemli mesaj, bireylerin sonucuna dayalı olarak önemli kararların alınmadığı (geçti-kaldı, veya seviye atlama gibi) değerlendirmelerdeki puanlarındaki farklılıkların kaynağının, bireylerin maddelere yeterince zaman ayırmama davranışı olabileceğidir.

Bu çalışmadan elde edilen bulgular, pratik açıdan faydalı olmasına rağmen birkaç yönden sınırlıdır. İlk olarak, bu çalışmada, farklı seviyelerde katılmama düzeyindeki ülkelerden bir seçim yapılmasına rağmen, PIAAC 2012'ye katılan tüm ülkeler bu çalışmaya dahil edilmemiştir. Bu çalışmanın bulguları bütün yetişkinlere genellenemeyebilir. Bu nedenle, bulgular gelecekteki araştırmalarda tekrarlanmalıdır. İkinci olarak, bu çalışmada tepki süresi eşiklerini belirlemek ve böylece katılmama ve katılma davranışlarını sergileyen bireyleri ayırt etmek için yalnızca bir yöntem kullanılmıştır. Katılmama davranışı sergileyen bireyleri tespit etmek için başka birçok yöntem vardır. Dolayısıyla, bu yöntemlerin etkinliğini karşılaştırmak için araştırmalar yapılabilir.