Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN:1309-6575 *İlkbahar* 2019 *Cilt: 10- Sayı: 1*
Spring 2019 *Volume: 10- Issue: 1*

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.

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.

İletişim
e-posta: epod@epod-online.org

Contact
e-mail: epod@epod-online.org
Web:http://epod-online.org

Dizinleme / Abstracting & Indexing
Emerging Sources Citation Index (ESCI), DOAJ (Directory of Open Access Journals), TÜBİTAK TR DIZIN

Hakem Kurulu / Referee Board

Ahmet Salih ŞİMŞEK (Cumhuriyet Üni.)
Ahmet TURHAN (American Institute Research)
Akif AVCU (Marmara Üni.)
Asiye Şengül Avşar (Recep Tayyip Erdoğan Üni.)
Ayfer SAYIN (Gazi Üni.)
Ayşegül ALTUN (Ondokuz Mayıs Üni.)
Arif ÖZER (Hacettepe Üni.)
Aylin ALBAYRAK SARI (Hacettepe Üni.)
Bahar Şahin Sarkın (İstanbul Okan Üni.)
Belgin DEMİRUS (MEB)
Bengu BORKAN (Boğaziçi Üni.)
Betül ALATLI (Gaziosmanpaşa Üni.)
Beyza AKSU DÜNYA (Bartın Üni.)
Bilge GÖK (Hacettepe Üni.)
Bilge BAŞUSTA UZUN (Mersin Üni.)
Burak AYDIN (Recep Tayyip Erdoğan Üni.)
Burcu ATAR (Hacettepe Üni.)
Burhanettin ÖZDEMİR (Siirt Üni.)
Cem Oktay GÜZELLER (Akdeniz Üni.)
Cenk AKAY (Mersin Üni.)
Ceylan GÜNDEĞER (Hacettepe Üni.)
Çiğdem Reyhanlioğlu Keçeoğlu
Cindy M. WALKER (Duquesne University)
Çiğdem AKIN ARIKAN (Hacettepe Üni.)
David KAPLAN (University of Wisconsin)
Deniz GÜLLEROĞLU (Ankara Üni.)
Derya ÇAKICI ESER (Kırıkkale Üni)
Derya ÇOBANOĞLU AKTAN (Hacettepe Üni.)
Didem KEPIR SAVOLY
Didem ÖZDOĞAN (İstanbul Kültür Üni.)
Dilara BAKAN KALAYCIOĞLU (ÖSYM)
Dilek GENÇTANRIM (Kırşehir Ahi Evran Üni.)
Durmuş ÖZBAŞI (Çanakkele Onsekiz Mart Üni)
Duygu Gizem ERTOPRAK (Amasya Üni)
Duygu KOÇAK (Alanya Alaaddin Keykubat Üni.)
Ebru DOĞRUÖZ (Çankırı Karatekin Üni.)
Elif Bengi ÜNSAL ÖZBERK (Trakya Üni.)
Emine ÖNEN (Gazi Üni.)
Emrah GÜL (Hakkari Üni.)
Emre ÇETİN (Doğu Akdeniz Üni.)
Emre TOPRAK (Erciyes Üni.)
Eren Halil Özberk (Trakya Üni.)
Ergül DEMİR (Ankara Üni.)
Erkan ATALMIS (Kahramanmaras Sutcu Imam Uni.)
Esin TEZBAŞARAN (İstanbul Üni)
Esin YILMAZ KOĞAR (Niğde Ömer Halisdemir Üni.)
Esra Eminoğlu ÖZMERCAN (MEB)
Fatih KEZER (Kocaeli Üni.)
Fatih ORCAN (Karadeniz Teknik Üni)
Fatma BAYRAK (Hacettepe Üni.)
Fazilet TAŞDEMİR (Recep Tayyip Erdoğan Üni.)
Funda NALBANTOĞLU YILMAZ (Nevşehir Üni.)
Gizem UYUMAZ (Giresun Üni.)
Gonca Usta (Cumhuriyet Üni.)
Gökhan AKSU (Adnan Menderes Üni.)
Gül GÜLER (İstanbul Aydın Üni.)
Gülden KAYA UYANIK (Sakarya Üni.)
Gülşen TAŞDELEN TEKER (Sakarya Üni.)
Hakan KOĞAR (Akdeniz Üni.)
Hakan Sarıçam (Dumlupınar Üni.)
Hakan Yavuz ATAR (Gazi Üni.)
Halil YURDUGÜL (Hacettepe Üni.)
Hatice KUMANDAŞ (Artvin Çoruh Üni.)
Hülya KELECİOĞLU (Hacettepe Üni.)
Hülya YÜREKLI (Yıldız Teknik Üni.)
İbrahim Alper KÖSE (Abant İzzet Baysal Üni.)
İlhan KOYUNCU (Adıyaman Üni.)
İlkay AŞKIN TEKKOL (Kastamonu Üni.)
İlker KALENDER (Bilkent Üni.)
Kübra ATALAY KABASAKAL (Hacettepe Üni.)
Levent YAKAR (Hacettepe. Üni.)
Mehmet KAPLAN (MEB)
Melek Gülşah ŞAHİN (Gazi Üni.)
Meltem ACAR GÜVENDİR (Trakya Üni.)
Meltem YURTÇU (Hcettepe Üni.)
Metin BULUŞ (Adıyaman Üni)
Murat Doğan ŞAHİN (Anadolu Üni.)
Mustafa ASİL (University of Otago)
Mustafa İLHAN (Dicle Üni.)
Nagihan BOZTUNÇ ÖZTÜRK (Hacettepe Üni.)
Neşe GÜLER (İzmir Demokrasi Üni.)
Neşe ÖZTÜRK GÜBEŞ (Mehmet Akif Ersoy Üni.)
Nuri DOĞAN (Hacettepe Üni.)
Nükhet DEMİRTAŞLI (Emekli Öğretim Üyesi)
Okan BULUT (University of Alberta)
Onur ÖZMEN (TED Üniversitesi)
Ömer KUTLU (Ankara Üni.)
Ömür Kaya KALKAN (Pamukkale Üni.)
Önder SÜNBÜL (Mersin Üni.)
Özge ALTINTAS (Ankara Üni.)
Özge BIKMAZ BİLGEN (Adnan Menderes Üni.)
Özlem ULAŞ (Giresun Üni.)
Recep GÜR (Erzincan Üni.)
Ragıp Terzi (Harran Üni.)
Recep Serkan ARIK (Dumlupınar Üni.)
Sakine GÖÇER ŞAHİN (University of Wisconsin Madison)

Hakem Kurulu / Referee Board

Seçil ÖMÜR SÜNBÜL (Mersin Üni.)
Sedat ŞEN (Harran Üni.)
Seher YALÇIN (Ankara Üni.)
Selahattin GELBAL (Hacettepe Üni.)
Selen Demirtaş ZORBAZ (Ordu Üni)
Selma Şenel(Balıkesir Üni.)
Sema SULAK (Bartın Üni.)
Semirhan GÖKÇE (Niğde Ömer Halisdemir Üni.)
Serkan ARIKAN (Muğla Sıtkı Koçman Üni.)
Seval KIZILDAĞ (Adıyaman Üni.)
Sevda ÇETİN (Hacettepe Üni.)
Sevilay KİLMEN (Abant İzzet Baysal Üni.)
Sinem Evin AKBAY (Mersin Üni.)
Sümeyra SOYSAL
Şeref TAN (Gazi Üni.)
Şeyma UYAR (Mehmet Akif Ersoy Üni.)
Tahsin Oğuz BAŞOKÇU (Ege Üni.)
Terry A. ACKERMAN (University of Iowa)
Tuğba KARADAVUT AVCI (Kilis 7 Aralık Üni.)
Tuncay ÖĞRETMEN (Ege Üni.)
Tülin ACAR (Parantez Eğitim)
Türkan DOĞAN (Hacettepe Üni.)
Yavuz AKPINAR (Boğaziçi Üni.)
Yeşim ÖZER ÖZKAN (Gaziantep Üni.)
Zekeriya NARTGÜN (Abant İzzet Baysal Üni.)
Zeynep ŞEN AKÇAY (Hacettepe Üni)

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

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (Mart 2019, Sayı: 10-1)

Journal of Measurement and Evaluation in Education and Psychology (March 2019, Issue: 10-1)

İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(1);1-11

A Preliminary Study to Evaluate the Reproducibility of Factor Analysis Results: The Case of Educational Research Journals in Turkey*

Burak AYDIN ** Mehmet KAPLAN *** Hakan ATILGAN ****
Sungur GÜREL *****

Abstract
In quantitative research, an attempt to reproduce previously reported results requires at least a transparent definition of the population, sampling method, and the analyses procedures used in the prior studies. Focusing on the articles published between 2010 and 2017 by the four prestigious educational research journals in Turkey, this study aimed to investigate the reproducibility of the factor analysis results from a theoretical perspective. A total of 275 articles were subject to descriptive content analysis. Results showed that 77.8% of the studies did not include an explicit definition of the population under interest, and in 50.9% of the studies, the sampling method was either not clear or reported to be convenience sampling. Moreover, information about the missing data or a missing data dealing technique was absent in the 76% of the articles. Approximately, half of the studies were found to have inadequate model fit. Furthermore, in almost all studies, it could not be determined whether the item types (i.e., levels of measurement scales) were taken into consideration during the analyses. In conclusion, the majority of the investigated factor analysis results were evaluated to be non-reproducible in practice.

Key Words: Reproducibility, factor analysis, descriptive content analysis

INTRODUCTION

The Open Science Collaboration (OSC) team reviewed several academic articles published in three respected psychology journals, investigated the reproducibility of the reported results in a total of 100 experimental or correlational studies, and stated that most of the results in those articles could not be obtained again (Open Science Collaboration, 2015). This reproducibility crisis was subject to both negative criticisms (e.g., Gilbert, King, Pettigrew, & Wilson, 2016) and supportive reports (e.g., Anderson et al., 2016). The negative criticism by Gilbert et al. (2016) stated that the reproducibility study by the OSC team had three issues that are sampling error, low statistical power, and bias. Hence the authors concluded that the OSC team seriously underestimated the reproducibility. This conclusion however criticized by Anderson et al. (2016) stating that Gilbert et al. (2016)'s study was very optimistic and based on statistical misconceptions and selective interpretations. Following the crisis, several steps such as journal policies that encourage to share data sets and the software scripts, and academic collaborations that promote open science (e.g., Moshontz et al., 2018) have been taken into consideration to overcome reproducibility issues in scientific research. Especially in social science, negligence in appropriate use of sample selection procedures and data analysis are the two main sources of error that may reduce the reproducibility rates of the results.

* Preliminary results of this work were presented at the 5th Congress on Measurement and Evaluation in Education and Psychology, Antalya 2016.

** Adjunct Professor, Recep Tayyip Erdoğan University, Department of Education, Rize-Turkey, burak.r.aydin@gmail.com, ORCID ID: 0000-0003-4462-1784
***Assistant Professor, Artvin Çoruh University, Department of Education, Artvin-Turkey, mehmet.kaplan2@gmail.com, ORCID ID: 0000-0002-4175-3899
**** Associate Professor, Ege University, Department of Education, Izmir-Turkey, hakan.atilgan@ege.edu.tr , ORCID ID: 0000-0002-5562-3446
*****Assistant Professor, Siirt Üniversitesi, Department of Education, Siirt-Turkey, s.gurel@siirt.edu.tr, ORCID ID: 0000-0003-3425-858X

To cite this article:
Aydın, B., Kaplan, M., Atılgan, H., & Gürel, S. (2019). A preliminary study to evaluate the reproducibility of factor analysis results: The case of educational research journals in Turkey. *Journal of Measurement and Evaluation in Education and Psychology, 10*(1), 1-11. DOI: 0.21031/epod.482393

Received: 13.11.2018
Accepted: 13.02.2019

The sampling method is an important part of quantitative research because inaccurate representation of the population can threaten the external validity of the study. Sampling methods can be classified in various ways (e.g., Balcı, 2000, Lavrakas, 2008; Kish, 1965; Levy & Lemeshow, 2013; Neuman, 2013); however, a most common categorization is known as probability sampling or non-probability sampling methods. Regardless of the sampling method, the use of inadequately small sample size and the existence of non-response or response bias (Lewis-Beck, Bryman & Liao, 2004) can result in non-reproducible findings in quantitative research. In addition, selection bias resulting from the non-probability-based methods is also another source of non-reproducible results. It is also important to take the sampling method into consideration when analyzing the data. Sterba (2009) discussed Neyman's and Fisher's frameworks to address sampling techniques when making statistical inferences. Fisher's framework requires three prerequisites with non-probability-based sampling methods, a correct statistical model, a valid distributional assumption, and conditionality. The conditionality assumption is not satisfied if the sampling technique is not taken into consideration (i.e., clustered or stratified sampling) and if the non-random sample fails to mimic a random sample due to disproportionally selected cases. On the other hand, Neyman's framework was created exclusively for random sampling methods (Sterba, 2009). Thus, the appropriate selection of sampling method and adequate data analysis play a vital role to increase reproducibility of research findings.

Inspired by the OSC's work (Open Science Collaboration, 2015) and their definition of the direct replication as an attempt to recreate the conditions to obtain previous findings, this study aims to show whether a team of researchers will have difficulties if they attempt to recreate the conditions and reproduce the results in the educational research articles published by the journals headquartered in Turkey. Hence, a preliminary study was designed to conduct a descriptive content analysis to investigate the sampling methods and data analysis procedures in these journals. To create a manageable study, the content was narrowed to factor analyses.

Factor Analysis in Educational Research

Educational researchers might reach conclusions using scores derived from a measurement tool, and in such cases, the validity of the conclusions is not independent of the validity of the scores. Scale validity is a unitary concept; however, evidence to support validity can be sought through several dimensions. One of these dimensions is known as construct validity (Atılgan, Kan, & Aydın, 2017; Nunnally & Bernstein, 1994). A psychological construct cannot be defined unless it is measurable (Crocker & Algina, 1986; Lord & Novick, 1968) and one of the procedures to provide evidence for the construct validity is the factor analysis. The use of factor analysis in educational research has been popular when developing a new scale or adapting a scale for cross-validation using confirmatory factor analysis (CFA) or explanatory factor analysis (EFA). CFA is also common when using a developed scale in quantitative research. For example, Göktaş et al. (2012), focusing on the studies conducted in Turkey, investigated 2111 articles published in 19 journals between 2005 and 2009 and identified a measurement tool in 1794 studies. A similar finding was reported by Karadağ (2011) who examined 211 doctoral dissertations completed between 2003 and 2007. Yılmaz and Altınkurt (2012), Sözbilir, Güler, and Çiltaş (2012), Selçuk, Palancı, Kandemir, and Dündar (2014), Kozikoğlu and Senemoğlu (2016), Yalçın, Yavuz, and Dibek (2016), and Gökmen et al. (2017) also noticed the common use of measurement tools both in national and international journals. Scale development and adaptation studies are also common in national journals. For example, Öztürk, Eroğlu, and Kelecioğlu (2015) identified 108 adaptation studies published in 10 journals between 2005 and 2014. The common use of scale development and adaptation was also noticed by Gül and Sözbilir (2015). Readers interested in further details about factor analysis and their role in educational research are referred to Acar (2014), Büyüköztürk (2002), Çüm and Koç (2013), Erkuş (2016), Güvendir and Özkan (2015), Kline (2015), Öztürk, Eroğlu, and Kelecioğlu (2015), Prudon (2015), Yurdugül and Bayrak (2012), Worthington and Whittaker (2006), and Wright (2017).

Results obtained with factor analysis are not independent from the sample. For example, Simon (1979) completed one of the studies that revealed the importance of sample selection in factor analysis. The

author wanted to draw attention that an attitude scale validated with a sample of university students could work differently for non-university students. His first sample consisted of 188 students from a single university, while the second sample consisted of 188 different individuals with the help of a foundation operating on a national basis. The author used the same factor analysis techniques on two different samples and reached different factor structures. At this point, it should be noted that, in the factor analysis, the sample should not represent a country, a territory, or a society, but it needs to represent the behaviors to be measured. Another study, which put forth the importance of sample selection in EFA, was completed by Gaskin, Orellana, Bowe, and Lambert (2017). The authors studied the construct validity of a scale used by the World Health Organization to determine whether individuals were generally healthy. In a study, in which 31251 individuals over the age of 50 from six different countries were considered as the population, the authors tested two different sampling methods. In the first approach, 1000 different samples were selected using simple random sampling to reflect the skewed distribution of the 31251 individuals' total health scores. In the second approach, 1000 different samples were selected with stratified random sampling to reach normally distributed scores. Exploratory factor analyses were performed on selected samples. With random sampling, generally a single factor solution was reached, whereas with the stratified sample a two-factor structure was reached. The authors found the structure obtained by stratified random sampling to be more defensible. These results showed that the sample can support different factor structures even when using probability-based methods. In addition, these results emphasize the importance of using prior knowledge about the population in sampling (Smith, 1983). From the sample perspective, one of the factors that make reproducibility difficult is using convenience sampling. The convenience sampling method can compromise the accuracy of the results in exchange for saving time and money (Balcı, 2015). The probability that a sample reached by the convenience sampling method is representative of any population greater than itself is usually very low. The validity of the results obtained by convenience sampling method has a high degree of concern, and this has been the subject of several academic studies (Bornstein, Jager, & Putnick, 2013; Delice, 2010; Landers & Behrend, 2015; Peterson & Merunka, 2014; Tyrer & Heyman, 2016).

After determining a sampling method that can represent the population, another important issue for reproducibility is the sample size. The sample needs to be sufficiently large to achieve unbiased estimates in factor analysis. Using an appropriate sample size may vary depending on the complexity of the factor structure, the magnitude of the factor loadings and the missing data. To determine the appropriate sample size in their studies, researchers can use the Monte Carlo simulation studies (Wolf, Harrington, Clark, & Miller, 2013). In other words, the definition of the population, choice of the sampling method and the sample size play an important role in factor analysis, and they affect the accuracy of the psychometric properties of the measurement tool. The factors obtained by factor analysis are affected by the sample (Kline, 2015; Thompson, 2004).

From a technical point of view, factor analysis is a dimension reduction process. The responses to n different questions in a scale form an n × n covariance matrix, and the factor analysis searches for a solution to produce this matrix using a smaller number of variables (Crocker & Algina, 1986). In other words, the variance with the n different variables is tried to be represented by a smaller number of variables, i.e., factors. This dimension reduction process can be quite complex depending on, for example, the number of questions, the relationship between items, how the missing data is handled, and the characteristics of the estimation method. Several sources address all the technical parts of factor analysis (e.g., Büyüköztürk, 2002; Crocker & Algina, 1986; Kline, 2015; Prudon, 2015; Thompson, 2004). A structure revealed by an EFA or CFA may not be reproduced with a similar sample if the missing data technique is not known (Akbaş & Tavşancıl, 2015; Çüm & Gelbal, 2015; Kürşad & Nartgün, 2015) and if the estimation method is not clearly defined (Beauducel & Herzberg, 2006; Hox, 1995). In addition, it should be clear whether the items were treated as categorical or continuous variables (Rhemtulla, Brosseau-Liard, & Savalei, 2012, Yang-Wallentin, Jöreskog, & Luo, 2010). Model-data fit information can also provide clues for reproducible findings (Prudon, 2015).

Overall, any attempt to reproduce results of a factor analysis requires detailed information about the sampling method and the analysis procedure. As stated earlier, the purpose of this study is to show

whether a team of researchers will have difficulties if they attempt to recreate the conditions and reproduce the factor analysis results reported in the educational research articles published by the journals headquartered in Turkey. The research questions are set to be:

1. Is the definition of the population explicit?
2. Which sampling methods are used?
3. What are the sample sizes, number of items and factors?
4. How is the missing data handled?
5. Which software is used?
6. Are the levels of measurement scales (categorical or continuous) taken into consideration and which estimators were used?
7. What is the reported data-model fit information?

METHOD

The scope of the study was limited to four internationally indexed educational research journals headquartered in Turkey, namely, Eurasian Journal of Educational Research (EJER), Educational Sciences Theory and Practice (ESTP), Hacettepe University Journal of Education (HUJE), and Education and Science (ES). Because it was not feasible to examine all the studies published in these journals with a small research team, the boundaries of this study were limited by the publication date and research topic. Specifically, the articles published between January 2010 and December 2017 including the keywords related to the factor analysis, which is one of the most commonly used data analysis method in educational research, were selected to be reviewed in this study. More specifically, to identify articles that reported factor analysis in the specified date range, keywords of *development, adaptation, factor analysis, structural equation modeling, validity, reliability, confirmatory, exploratory, CFA, EFA, Cronbach* or their Turkish translations were searched and a total of 341 academic articles were downloaded to be reviwed for the purpose of this study. Articles in each journal were examined by one of the four authors in our research team, and it was narrowed down to 275 out of 341 articles where CFA, EFA, or Principal Component Analysis (PCA) were used for the data analysis. These 275 articles were then investigated in a descriptive content analysis framework. The descriptive content analysis is one of the quantitative data analysis methods and usually includes reporting of basic statistics such as frequency, average, median, and variance (Gall, Gall, & Borg, 1996; Stapleton & Leite, 2005).

Data Collection

Title, publication year, publishing journal, and general purpose in 275 articles selected for this study were recorded. Specifically, the general purpose of the study was coded as scale development, scale adaptation, or other. The sampling characteristics, sampling method and the clear definition of the population were considered as the first dimension of reproducibility. The content of the sample used in those studies was coded as *students, teachers or prospective teachers, academicians, administrators*, or *other*. The data analysis procedures, which were considered as the second dimension of the reproducibility, were also examined in this study. Specifically, the following criteria were recorded: (i) whether the missing data was reported, (ii) whether the missing data was handled using an appropriate technique, (iii) whether an EFA and CFA were performed using the same sample, (iv) sample size, (v) number of items in scales, (vi) number of factors found, (vii) items types (e.g., Likert or yes/no), (viii) software, and (ix) model-data fit information. The data analysis techniques were coded as explanatory or confirmatory. It is worth to note that PCA was considered as an exploratory technique (Bryant & Yarnold, 1995). For the model fit information, the ratio of the chi-square to the degrees of freedom, the root of the square error of approximation (RMSEA), standardized root mean square residual (SRMR), comparative fit index (CFI), Tucker-Lewis index (TLI or NNFI), normative fit index (NFI), goodness of fit index (GFI), adjusted GFI (AGFI), incremental fit index (IFI), and relative fit index (RFI) were recorded. In addition, if more than one scale was used in an article, number of items, number of factors, type of the items, and fit information were recorded on a different row for the same article. Also, if more

than one model was tested for the same scale, only the information of the final model was recorded. As a result, the final data set consisted of 448 rows in total.

FINDINGS

The number of published articles selected for this study was 35 (12.7%) in 2010, 32 (11.6%) in 2011, 35 (12.7%) in 2012, 46 (16.7%) in 2013, 53 (19.3%) in 2014, 28 (10.2%) in 2015, 18 (6.5%) in 2016, and 28 (10.2%) in 2017. In addition, the frequency of the articles by the journals was 94 (34.2%), 56 (20.4%), 40 (14.5%), and 85 (30.9%) for the ES, EJER, HUJE, and ESTP, respectively. The frequency of studies in scale development was 108 (39.3%), in scale adaptation was 99 (36.0%), and in other topics was 68 (24.7%). Table 1 shows the frequencies of the 275 articles by year, journal, and research purpose.

Table 1. The Frequencies of the 275 Articles by Year, Journal, and Study Purpose.

Year	ES			EJER			HUJE			ESTP			Total
	SD	SA	O	SD	SA	O	SD	SA	O	SD	SA	O	
2010	3	5	5	1	3	1	4	4	0	6	3	0	35
2011	3	8	3	1	0	1	3	1	0	6	5	1	32
2012	4	3	2	2	3	0	3	6	0	9	1	2	35
2013	5	11	2	2	3	1	3	5	0	6	6	2	46
2014	12	12	4	5	0	5	3	1	0	5	3	3	53
2015	1	0	2	1	2	7	4	1	2	2	0	6	28
2016	0	2	2	2	2	4	0	0	0	2	4	0	18
2017	1	2	2	5	2	3	0	0	0	4	1	8	28
Total	29	43	22	19	15	22	20	18	2	40	23	22	275

Note: SD = Scale development, SA = Scale adaptation, O = Other.

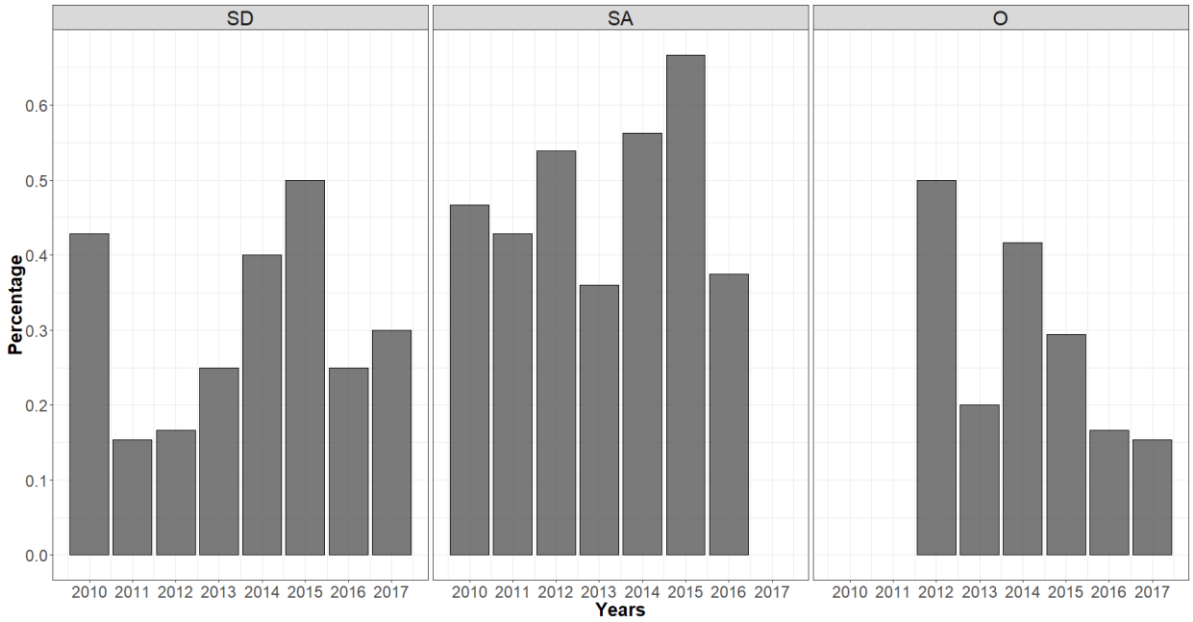
Definition of Population and Sampling Method

A clear definition of the population and the appropriate selection of the sampling method in quantitative research are important for ensuring the validity of the results. Based on the results, only 61 (22.2%) of the 275 articles reviewed in this study provided a clear definition of the population in their research. Table 2 shows the percentage of the studies that explicitly reported the population definition by year and study purpose. Scale development and adaptation studies included a clear definition approximately 1 in every 5 studies, whereas other studies had a rate of 1 in every 3. On the contrary to unclear definition of the population, the sampling method, whether probability-based or non-probability-based, was determined in 227 (82.6%) of the articles. More specifically, 169 of those 227 studies used a non-probability-based sampling, and 58 used a probability-based sampling. Of the 169 studies, the sampling technique was clearly stated in 112 articles where 92 of them were convenience, 11 of them were purposeful, 5 of them were stratified, 2 of them were maximum variation, 1 of them was snowball, and 1 of them was typical case sampling. In general, 48 (17.4%) of the 275 studies did not have a clear definition of the sampling method and 92 (33.5%) of the 275 stated that convenience sampling was used. Figure 1 shows the percentage of convenience sampling across years and study purpose. Overall 31%, 49% and 29% of the studies reported the use of convenience sampling for scale development, scale adaptation and other purposes respectively. In addition, the content of the sample was clearly defined in all articles. Specifically, 205 (74.6%) of the studies included only students, 40 (14.6%) of them included teachers or prospective teachers, 8 (2.9%) of them included only academicians, 4 (1.4%) of them included only administrators, and the remaining 18 (6.4%) of them included at least two of these groups or other individuals (e.g., parents and adults).

Table 2. Population and Missing Data Information of the 275 Articles by Year and Study Purpose

Year	Population information percentage			Missing data information percentage		
	SD	SA	O	SD	SA	O
2010	21	33	17	29	13	33
2011	15	21	40	31	21	40
2012	28	8	50	22	23	25
2013	25	12	40	13	20	0
2014	12	6	25	20	13	25
2015	0	67	41	25	67	29
2016	0	38	50	75	25	17
2017	40	20	8	10	40	46
All years	19	19	31	23	21	29
Overall	22			24		

Note: SD = Scale development, SA = Scale adaptation, O = Other.



Note: SD = Scale development, SA = Scale adaptation, O = Other.

Figure 1. Percentage of Convenience Sampling Across Years and Study Purpose

Sample Size, Number of Items, Number of Factors, and Item Types

The sample size, number of items, number of factors, and item types were recorded separately for 448 analyses in 275 articles. The median values of the observed sample sizes, the number of items used in the scale, and the number of obtained factors were 398, 25, and 3, respectively. In addition, the median value of the sample size per item was 14.8, and the number of items per factor was distributed with a median of 7. Table 3 shows the median values for sample size, item per factor and sample size per item by year, and study purpose. Item per factor median values were similar across years and purpose. However, sample size median values across all years were slightly lower for the scale development and adaptation, 381 and 400 respectively, compared to the median value for the other purposes which was 459. The sample size per item median values across all years were similar for the scale adaptation and other studies, 16.6 and 17.7, respectively, slightly larger compared to scale development values which was 12.2. Items with more than two categories (e.g., Likert) were employed in 228 (82.9%) of the 275 articles, whereas 7 (2.5%) studies used binary, 4 (1.5%) studies used continuously scaled items, and the item type could not be determined for the remaining 36 (13.1%) studies. Furthermore, a total of 318

individual analyses out of 448 reported the item type, and out of these 318, 304 used items with more than two categories. The most preferred items (i.e., in 209 analyses) were the ones with five categories. Items with three, four, six, seven, nine, and ten categories were also used in 11, 36, 16, 24, 5, and 3 analyses, respectively.

Missing Data and Analysis Procedure

Of the 275 articles reviewed, only 66 (24%) reported how the missing data were handled. Of these 66 studies, 62 utilized listwise deletion and 3 utilized an imputation method. In addition, it was reported that there was no missing data at all in 1 study. Table 2 shows the percentage of the studies that included missing data information by year and study purpose. Similar to population definition rates, scale development and adaptation studies had a lower rate, 23% and 21% respectively, compared to other studies, 29%.

For the data analysis method, it was determined that 84 (30.6%) of the studies employed only CFA, 57 (20.7%) employed only EFA, and 134 (48.7%) employed both EFA and CFA. 90 of 134 articles that employed both EFA and CFA conducted analyses using the same sample, or they divided the study sample into halves. The software information could be identified in 183 of the 275 articles. Specifically, SPSS and Lisrel together, Lisrel, SPSS, AMOS, SPSS and AMOS together, M*plus*, and EQS were used in 71, 49, 29, 19, 12, 2, and 1 studies, respectively.

Table 3. Median Values of Sample Size, Item per Factor and Sample Size per Factor of the 448 Analyses by Year and General Purpose

Year	Sample size median			Item per factor median			Sample size per item median		
	SD	SA	O	SD	SA	O	SD	SA	O
2010	464	358	367	8.9	8.3	8.5	12.2	21.3	10.0
2011	461	341	214	8.5	7.6	4.0	12.4	12.8	16.6
2012	336	529	258	6.1	7.0	12.5	10.8	13.6	6.0
2013	388	407	605	6.5	6.0	4.0	10.7	21.9	97.9
2014	317	436	256	6.0	5.0	7.0	12.9	25.6	14.9
2015	384	357	657	10.7	6.3	6.1	13.0	9.4	49.2
2016	330	462	556	4.3	5.3	7.0	12.3	15.5	20.4
2017	303	270	719	7.3	5.7	6.6	11.0	16.4	27.8
All years	381	400	459	7.3	6.5	7.0	12.2	16.6	17.7
Overall	398			7				14.8	

Note: SD = Scale development, SA = Scale adaptation, O = Other.

Data-model Fit Information

The ratio of chi-square by the degrees of freedom was reported in 183 studies, and it ranged between 1.01 and 9.45 with a median value of 2.66; RMSEA was reported in 245 studies ranged between 0 and 0.44 with a median of 0.06; SRMR was reported in 131 studies ranged between 0.004 and 0.11 with a median of 0.05; CFI was reported in 233 studies ranged between 0.70 and 1 with a median of 0.96; TLI was reported in 143 studies between 0.69 and 1 with a median of 0.96; NFI was reported in 146 studies ranged between 0.64 and 1 with a median of 0.95; GFI was reported in 197 studies ranged between 0.47 and 1 with a median of 0.92; AGFI was reported in 157 studies ranged between 0.07 and 1 with a median of 0.90; IFI was reported in 54 studies ranged between 0.81 and 1 with a median of 0.95; and finally RFI was reported in 41 studies ranged between 0.62 and 1 with a median of 0.96. The estimator was determined only in 39 analyses, and 30 of them utilized maximum likelihood, 7 used robust maximum likelihood, and 2 used least squares methods.

DISCUSSION and CONCLUSION

Inspired by the reproducibility crisis in psychology research (Open Science Collaboration, 2015), this preliminary study aimed to evaluate the reproducibility of the results using factor analysis in four prestigious educational research journals headquartered in Turkey. The authors examined 448 different analyses reported in 275 articles published between 2010 and 2017 based on sampling method and data analysis procedures which were considered as two of the main dimensions of reproducible research.

Factor analyses were generally employed with a purpose of either scale development or scale adaptation in 75.3% of the 275 articles and they were used with different purposes for the remaining 24.7% of the articles. A clear definition of the population was not found in 77.8% of the studies which can be evidence for the threat to the validity. The number of articles in which the sampling method could not be determined or determined as the convenience sampling was 140 (50.9%). In 76% of the studies, the information about how the missing data was handled could not be identified, and the ones where the missing data was reported used outdated techniques, such as listwise deletion and mean imputation. In 90 of 275 studies, both the EFA and CFA were utilized using the same sample. The results obtained by EFA and CFA using the same data have been a subject of debate (Erkuş, 2016; Van Prooijen & Van Der Kloot, 2001). Considering the importance of a clear definition of the population and the use of proper sampling method that can produce generalizable results, these findings were evaluated as the evidence of non-reproducible results in those articles. Handling of missing data is an important part in factor analysis (Allison, 2003; Çüm, & Gelbal, 2015), as for the social sciences in general (Schafer, 1997; Schlomer, Bauman, & Card, 2010). The fact that the missing data was not explicitly addressed in the examined studies increased the concern for non-reproducible results in those articles. The missing data issue in the educational research conducted in Turkey was also noticed by Demir and Parlak (2012). Çüm and Gelbal (2015) stated that in the case of misuse of missing data techniques, the results could be misleading, and this is directly related to the reproducibility of the results. It is not clear why the missing data or the missing data technique were not mentioned in three of the four examined studies, if there were no missing data at all and it was due to forced responses, this is also alarming in terms of reproducibility (Ray, 1990; Xiao, Liu & Li 2017).

In factor analysis, another important issue regarding reproducibility of results is to provide adequate sample size (Wolf, Harrington, Clark, & Miller, 2013). Selecting an adequate sample size depends on the complexity of the model and the magnitude of the factor loadings. Monte Carlo simulations are powerful techniques that can be used to determine the appropriate size, but in the literature, there are recommendations for the ratio of the number of participants to the number of items, for example, 1 to 20 and 1 to 10 (Hogarty, Hines, Kromrey, Ferron, & Mumford, 2005). In the articles examined in this study, the median value of this ratio was found to be approximately 15, and in general, it was evaluated that the importance of sample size was recognized. The average number of items per factor was fewer than seven in half of the studies. In theory, if there are multiple factors in the model, a factor can be defined with two items, but it is recommended to have at least 3, 4, or 5 items per factor (Kline, 2015). Increasing the number of items can allow for a strong definition of the structure, thus enhance the reproducibility. In general, it was evaluated that the importance of the number of items per factor was not recognized in the articles examined for this study.

The model-data fit information used in factor analysis is a clue for the reproducibility of the results. Fit values would be low if there were unexplained variance sources or the model was not correctly specified, and this poses a risk for reproducibility. For the model-data fit information, what should be the cut-off values is the subject of several studies (Kline, 2015; Marsh, Balla, & McDonald 1988; O'Boyle & Williams 2011; Prudon, 2015), assuming RMSEA<0.06, SRMR <0.08, CFI, and TLI (NNFI, NFI, GFI, and AGFI> 0.95 indicate a good fit, nearly half of the studies examined were found to have difficulty in meeting these criteria. The ratio of the chi-square to the degrees of freedom was not taken into consideration in our evaluation, given that it should not be used (Kline, 2015). Furthermore, the fact that the estimator information was not identified in most of the analyses prevented us to determine whether the characteristics of the items were taken into consideration during the analysis process and this is another concern, as when the normality assumption is not met, treating categorical (e.g., Likert) variables as continuous is likely to harm reproducibility (Li, 2016).

Overall the majority of the investigated factor analysis results were evaluated to be non-reproducible in practice. This non-reproducibility issue seems to be more evident for the scale development and adaptation studies compared to studies with other quantitative purposes given that the later has better rates of a clear definition of the population and missing data, along with relatively larger sample sizes and decreasing number of convenience sampling utilization. This study has its limitations. One of them is that the scope is broad; however, as the title indicates, this is a preliminary study to show an alarming issue, namely, a possible reproducibility crisis of educational research studies published by Turkish Journals. Researchers are invited to conduct more in-depth reproducibility studies for example with a focus on particular scales, EFA and rotation options (e.g., Kline, 2015; Osborne, 2015; Saracli, 2011), CFA and modification issues (e.g., Asparouhov & Muthen, 2009; Mueller & Hancock, 2008). The second limitation is that model-fit information is affected at least by the sample size, estimator, and model specification; hence, the model-fit information was not considered as a main indicator of reproducibility, but rather considered as clues. The third limitation is that no guideline was provided for the practitioners. However, it was made clear that any attempt to recreate conditions to reproduce a practitioner`s results will fail if the population, sampling method, and the analyses procedures were not represented transparently. When these reproducibility basics are fulfilled, practitioners should take advantage of already published guidelines, for example, Büyüköztürk (2002), Erkuş (2016), Kline (2015), Öztürk, Eroğlu and Kelecioğlu (2015), Prudon (2015), Worthington and Whittaker (2006), and Wright (2017). It is also strongly recommended for practitioners to share their data-set and data analysis syntax whenever possible. The list of 275 articles investigated in this preliminary study and the data set including information from 448 analyses are provided as supplementary files.

REFERENCES

Acar, T. (2014). Ölçek geliştirmede geçerlik kanıtları: Çapraz geçerlik, sınıflama ve sıralama geçerliği uygulaması. *Kuram ve Uygulamada Eğitim Bilimleri*, *14*(2), 1-11. DOI: 10.12738/estp.2014.3.2107

Akbaş, U., & Tavşancıl, E. (2015). Farklı Örneklem büyüklüklerinde ve kayıp veri örüntülerinde ölçeklerin psikometrik özelliklerinin kayıp veri baş etme teknikleri ile incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *6*(1) 38-57.

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, *112*, 545-557.

Anderson, C. J., Bahník, Š., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., ... & Della, N. P. (2016). Response to Comment on" Estimating the reproducibility of psychological science". *Science (New York, NY)*, *351*(6277), 1037-1037.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural equation modeling: a multidisciplinary journal*, *16*(3), 397-438.

Atılgan, H., Kan, A., & Aydın, B. (2017). *Eğitimde Ölçme ve Değerlendirme*. Ankara: Anı Yayıncılık.

Balcı, A. (2015) *Sosyal Bilimlerde Araştırma: Yöntem, Teknik ve İlkeler*. Ankara: Pegem Yayınları.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, *13*(2), 186-203. DOI: 10.1207/s15328007sem1302_2

Bornstein, M. H., Jager, J., & Putnick, D. L. (2013). Sampling in developmental science: Situations, shortcomings, solutions, and standards. *Developmental Review*, *33*(4), 357-370. DOI: 10.1016/j.dr.2013.08.003

Bryant, F. B., & Yarnold, P. R. (1995). Principal-components analysis and exploratory and confirmatory factor analysis. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 99-136). Washington, DC, US: American Psychological Association.

Büyüköztürk, Ş. (2002). Faktör analizi: Temel kavramlar ve ölçek geliştirmede kullanımı. *Kuram ve Uygulamada Eğitim Yönetimi*, *32*, 470-483.

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

Çüm, S., & Gelbal, S. (2015). Kayıp veriler yerine yaklaşık değer atamada kullanılan farklı yöntemlerin model veri uyumu Üzerindeki etkisi. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi*, *1*(35), 87-111.

Çüm, S., & Koç, N. (2013). Türkiye'de psikoloji ve eğitim bilimleri dergilerinde yayımlanan ölçek geliştirme ve uyarlama çalışmalarının incelenmesi. *Journal of Educational Sciences & Practices*, *12*(24), 115-135.

Delice, A. (2010). Nicel araştırmalarda örneklem sorunu. *Kuram ve Uygulamada Egitim Bilimleri*, *10*(4), 1969-2018.

Demir, E., & Parlak, B. (2012). Türkiye'de eğitim araştırmalarında kayıp veri sorunu. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *3*(1), 230-241.

Erkuş, A. (2016). Ölçek geliştirme ve uyarlama çalışmalarındaki sorunlar ile yazım ve değerlendirilmesi. *Pegem Atıf İndeksi*, 1211-1224.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction*. White Plains, NY: Longman USA.

Gaskin, C. J., Orellana, L., Bowe, S. J., & Lambert, S. D. (2017). Why sample selection matters in exploratory factor analysis: Implications for the 12-item world health organization disability assessment schedule 2.0. *BMC Medical Research Methodology*, *17*(1), 40.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on estimating the reproducibility of psychological science. *Science*, *351*(6277), 1037-1037.

Gökmen, Ö. F., Uysal, M., Yaşar, H., Kirksekiz, A., Güvendi, G. M., & Horzum, M. B. (2017). Türkiye'de 2005-2014 yılları arasında yayınlanan uzaktan eğitim tezlerindeki yöntemsel eğilimler: Bir İçerik analizi. *Eğitim ve Bilim*, *42*(189), 1-25.

Göktaş, Y., Küçük, S., Aydemir, M., Telli, E., Arpacık, Ö., Yıldırım, G., & Reisoğlu, I. (2012). Türkiye'de eğitim teknolojileri araştırmalarındaki eğilimler: 2000-2009 dönemi makalelerinin içerik analizi. *Kuram ve Uygulamada Eğitim Bilimleri Dergisi*, *12*(1), 177-199.

Gül, Ş., & Sözbilir, M. (2015). Fen ve matematik eğitimi alanında gerçekleştirilen Ölçek geliştirme araştırmalarına yönelik tematik içerik analizi. *Eğitim ve Bilim*, *40*(178), 85-102.

Güvendir, M. A., & Özkan, Y. Ö. (2015). Türkiye'deki eğitim alanında yayımlanan bilimsel dergilerde ölçek geliştirme ve uyarlama konulu makalelerin incelenmesi. *Elektronik Sosyal Bilimler Dergisi*, *14*(52), 23-33.

Hogarty, K. Y., Hines, C. V., Kromrey, J. D., Ferron, J. M., & Mumford, K. R. (2005). The quality of factor solutions in exploratory factor analysis: The influence of sample size, communality, and overdetermination. *Educational and Psychological Measurement*, *65*(2), 202-226. DOI: 10.1177/0013164404267287

Hox, J. J. (1995). Amos, EQS, and Lisrel for windows: A comparative review. *Structural Equation Modeling: A Multidisciplinary Journal*, *2*(1), 79-91.

Karadağ, E. (2011). Eğitim bilimleri doktora tezlerinde kullanılan Ölçme araçlar: Nitelik düzeyleri ve analitik hata tipleri. *Kuram ve Uygulamada Eğitim Bilimleri*, *11*(1), 311-334.

Kish, L. (1965). *Survey sampling*. Oxford. England:John Wiley & Sons.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. New York, NY, US: Guilford publications.

Kozikoğlu, I., & Senemoğlu, N. (2016). Eğitim programları ve Öğretim alanında yapılan doktora tezlerinin içerik analizi (2009-2014). *Eğitim ve Bilim*, *40*(182), 29-41.

Kürşad, M. Ş., & Nartgün, Z. (2015). Kayıp veri sorununun çözümünde kullanılan farklı yöntemlerin Ölçeklerin geçerlik ve güvenirliği bağlamında karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *6*(2), 254-267. DOI: 10.21031/epod.95917

Landers, R. N., & Behrend, T. S. (2015). An inconvenient truth: Arbitrary distinctions between organizational, mechanical turk, and other convenience samples. *Industrial and Organizational Psychology*, *8*(2), 142-164.

Lavrakas, P. J. (2008). *Encyclopedia of survey research methods*. Thousand Oaks: Sage Publications.

Levy, P. S., & Lemeshow, S. (2013). *Sampling of populations: Methods and applications*. John Wiley & Sons.

Lewis-Beck, M. S., Bryman, A. , & Liao, T. F. (2004). *The Sage Encyclopedia of social science research methods.* Thousand Oaks: Sage.

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936-949. DOI: 10.3758/s13428-015-0619-7.

Lord, F. M., and Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, *103*(3), 391-410. DOI: 10.1037/0033-2909.103.3.391

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., ... & Castille, C. M. (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*,*1*(4),501-515. DOI: 10.1177/2515245918797607

Mueller, R. O., & Hancock, G. R. (2008). Best practices in structural equation modeling. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 488–508). Thousand Oaks: Sage Publications Inc.

Neuman, W. L. (2013). *Social research methods: Qualitative and quantitative approaches*.UK: Pearson education.

Nunnally, J., & Bernstein, I. (1994). *Psychometric Theory (3rd Edition).* New York: McGraw-Hill, Inc.

O'Boyle, E. H., Jr., & Williams, L. J. (2011). Decomposing model fit: Measurement vs. theory in organizational research using latent variables. *Journal of Applied Psychology, 96*(1), 1-12.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). doi: 10.1126/science.aac4716

Osborne, J. W. (2015). What is rotating in exploratory factor analysis. *Practical assessment, research & evaluation, 20*(2), 1-7.

Öztürk, N. B., Eroğlu, M. G., & Kelecioğlu, H. (2015). Eğitim alanında yapılan ölçek uyarlama makalelerinin incelenmesi. *Eğitim ve Bilim*, *40*(178), 123-137.

Peterson, R. A., & Merunka, D. R. (2014). Convenience samples of college students and research reproducibility. *Journal of Business Research*, *67*(5), 1035-1041. DOI: 10.1016/j.jbusres.2013.08.010

Prudon, P. (2015). Confirmatory factor analysis as a tool in research using questionnaires: A critique. *Comprehensive Psychology*, *4*, 1-19. DOI: 10.2466/03.CP.4.10

Ray, J. J. (1990). Acquiescence and problems with forced-choice scales. *The Journal of Social Psychology*, *130*(3), 397-399. DOI: 10.1080/00224545.1990.9924595

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354-373. DOI: 10.1037/a0029315

Saracli, S. (2011). Faktör analizinde yer alan döndürme metotlarının karşılaştırmalı incelenmesi üzerine bir uygulama. *Düzce Üniversitesi Sağlık Bilimleri Enstitüsü Dergisi*, *1*(3), 22-26.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman and Hall: CRC.

Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best practices for missing data management in counseling psychology. *Journal of Counseling Psychology*, *57*(1), 1-10.

Selçuk, Z., Palancı, M., Kandemir, M., & Dündar, H. (2014). Eğitim ve bilim dergisinde yayınlanan araştırmaların eğilimleri: İçerik analizi. *Eğitim ve Bilim*, *39*(173), 428-449.

Simon, A. (1979). Effects of selective sampling on a factor analysis. *The Journal of General Psychology*, *101*(2), 259-264. DOI: 10.1080/00221309.1979.9920079

Smith, T. (1983). On the validity of inferences from non-random sample. *Journal of the Royal Statistical Society. Series A (General), 146*(4), 394-403. DOI: 10.2307/2981454

Sözbilir, M., Güler, G., & Çiltaş, A. (2012). Türkiye'de matematik eğitimi araştırmaları: Bir içerik analizi Çalışması. *Kuram ve Uygulamada Eğitim Bilimleri*, *12*(1), 565-580.

Stapleton, L. M., & Leite, W. L. (2005). Teacher's corner: A review of syllabi for a sample of structural equation modeling courses. *Structural Equation Modeling*, *12*(4), 642-664.

Sterba, S. K. (2009). Alternative model-based and design-based frameworks for inference from samples to populations: From polarization to integration. *Multivariate Behavioral Research, 44*(6), 711-740.

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

Tyrer, S., & Heyman, B. (2016). Sampling in epidemiological research: Issues, hazards and pitfalls. *BJPsych Bulletin, 40*(2), 57-60.

Van Prooijen, J. W., & Van Der Kloot, W. A. (2001). Confirmatory analysis of exploratively obtained factor structures. *Educational and Psychological Measurement*, *61*(5), 777-792. DOI: 10.1177/00131640121971518

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement, 73*(6), 913-934.

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, *34*(6), 806-838.

Wright, A. G. (2017). The current state and future of factor analysis in personality disorder research. *Personality Disorders: Theory, Research, and Treatment*, *8*(1), 14-25.

Xiao, Y., Liu, H., & Li, H. (2017). Integration of the Forced-Choice Questionnaire and the Likert Scale: A Simulation Study. *Frontiers in Psychology*, *8*, 806. DOI: 10.3389/fpsyg.2017.00806

Yalçın, S., Yavuz, H. C., & Dibek, M. I. (2016). En yüksek etki faktörüne sahip eğitim dergilerindeki makalelerin İçerik analizi. *Eğitim ve Bilim*, *40*(182), 1-28.

Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling*, *17*(3), 392-423.

Yılmaz, K., & Altınkurt, Y. (2012). An examination of articles published on preschool education in Turkey. *Educational Sciences: Theory and Practice*, *12*(4), 3227-3241.

Yurdugül, H., & Bayrak, F. (2012). Ölçek geliştirme Çalışmalarında kapsam geçerlik Ölçüleri: Kapsam geçerlik indeksi ve kappa istatistiğinin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, Özel Sayı*, *2*, 264-271.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(1);12-29

A Statistical Comparison of Norm-Referenced Assessment Systems Used in Higher Education in Turkey*

Erkan Hasan ATALMIŞ **

Abstract

The purpose of this study is to identify different norm-referenced assessment systems used in Turkish higher education, and to compare them empirically. Norm-referenced assessment regulations of 70 universities in Turkey was primarily analyzed, and universities were divided into four different groups depending on their norm-referenced assessment systems (only applying T-score conversion, the most commonly used method; applying T-score conversion and quantiles together; applying T-score conversion, quantiles and standard deviation together; applying standard deviation based norm-referenced assessment system). After the algorithms of two universities applying T-score conversion and three universities applying other norm-referenced assessment system were selected, they were used to convert end-of-year grade for each course of 19,574 students in a state university into letter grades and 4-point system. To test the differences of the norm-referenced assessment systems used in these universities, the norm-referenced system of a university were compared with the criterion-referenced system of the same university as well as norm-referenced systems of other universities. The paired t-test was used to identify the difference between norm-referenced and criterion-referenced assessment, while the differences between norm-referenced assessment systems were analyzed through one-way analysis of variance. The findings revealed that the letter grades calculated through the norm-referenced assessment were statistically different than the ones calculated with criterion-referenced; besides, a statistically significant difference was identified between the letter grades obtained using the norm-referenced assessment systems of universities. At the end of the study, the findings were discussed in term of students and instructors.

Key Words: Norm-referenced assessment, criterion-referenced assessment, assessment in higher education, grading system.

INTRODUCTION

The aim of education and training is the disclosure of the cognitive, affective and psychomotor skills to the students in a planned and programmed way. To ensure this, the curriculum consisting of four basic components should be primarily determined. These components include (a) determining the behavioral objectives, (b) constructing the content in accordance with these objectives and the readiness of the students, (c) creating learning and teaching activities with the idea that each student learns differently; and (d) performing meaningful assessment and evaluation (Tan, 2015). In particular, the significance of measurement and assessment cannot be underestimated in terms of determining the extent to which the behavioral objectives within the program reflect the readiness of the students and identifying as to what extent learning and teaching activities are appropriate to the objectives and behaviors.

The concepts of measurement and assessment are different and even complementary concepts. Specifically, measurement refers to a variable or an object with numbers or symbols, while assessment provides a meaningful interpretation of the results obtained from the measurement by comparing them through a frame of reference. Previous studies have revealed that a frame of reference will vary across teacher notions, student success distribution in the class, student ability and their achievement scores related to the program (learning difference at the beginning and end of the program) as well as the

* A part of the study was presented at 2018 EDUCCON Education Conference (Ankara University, Ankara, Turkey.
** Assistant professor, Kahramanmaraş Sütçü Imam University, Faculty of Education, Kahramanmaraş-Türkiye, eatalmis@ksu.edu.tr, ORCID ID: orcid.org/0000-001-9610-491X

To cite this article:
Atalmış, E., H. (2019). A statistical comparison of norm-referenced assessment systems used in higher education in Turkey. *Journal of Measurement and Evaluation in Education and Psychology, 10*(1), 12-29. DOI: 10.21031/epod.487335

Received: 25.11.2018
Accepted: 28.01.2019

objectives of the program (Martin & Jolly, 2002; Turgut & Baykul, 2015; Yorke, 2011). Considering these situations, one of the most essential factors of an accurate assessment is the selection of an appropriate reference. The assessment is already categorized depending on the reference in use. Criterion-referenced in which absolute criterion is used is defined as an assessment which is accepted by everyone in the same way without reference to the group and group characteristics. Norm-referenced is a type of assessment which yields for the relative criteria and the assessment made depending on the criteria selected according to the predefined group and especially the success of the group.

Thorndike (2005) argues that the criterion-referenced assessment plays a significant role in directing learning and teaching activities since the use of this type of assessment is more relevant to what extent people achieve the level of targeted knowledge. Sadler (2005) suggests that the criterion-referenced assessment provides students with the grades they deserve due to the fact that the grades based on this assessment are calculated regardless of each student's achievement. In contrast to the criterion-referenced assessment, it is recommended that norm-referenced assessment be used for sorting, placement and in distinguishing the achievement sequence of the students (American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME), 2014). Nartgün (2007) has noted that the use of norm-referenced assessment is particularly relevant in the large-scale and national-scale examinations that have upper levels and that require placement.

Although there does not exist a definite line in terms of the use of criterion-referenced and norm-referenced assessment, the exams that the criterion-referenced method is applied may be prepared in accordance with the exam preparation guidelines (Thorndike, 2005). Haladyna and Rodriguez (2013) have indicated that susceptibility should be displayed for the construction of exam questions based on the objectives, item-writing guidelines and the cognitive taxonomy levels. Otherwise, the scores obtained as a result of the examination will tend to be less homogeneous; the variation of the measured feature will not be explained at the maximum level, and in this case, ranking or placement of the students according to the scores will not be reasonable in terms of measurement and assessment (Hambleton et al., 1978).

In particular, the question related to what kind of assessment type is applied is the subject of hot debate in higher education today. The current relevant studies have put forward that the use of norm-referenced assessment in universities with high competition and success motivation will be much more effective, and that criterion-referenced assessment will be more appropriate in schools with low achievement motivation since norm-referenced assessment will cause grading inflation (Başol, 2013; Selvi, 1998). In one of the most comprehensive studies on this subject, Johnson (2003) has argued that grading inflation is a serious problem in universities and that the method used for grading can vary across universities and faculties. Unfortunately, a limited number of studies have been conducted to compare criterion-referenced and norm-referenced assessment.

Having analyzed the guidelines of two different state universities using criterion-referenced and norm-referenced assessment, Nartgün (2007) has compared the grades with different distributions and suggested that the grades calculated with the norm-referenced assessment provide more than those of the students deserve. In other words, when compared with the criterion-referenced assessment, it is evident that norm-referenced assessment leads to grading inflation. However, criterion-referenced assessment method offers more effective results than the norm-referenced assessment in cases when the raw achievement grades are close to one another, that is, the scores are similar (standard deviation is low). Atılgan, Yurdakul, and Öğretmen (2012) have found different results compared to the findings of Nartgün (2007). 3,120 grades obtained from the students studying at the faculty of education and calculated through use of norm-referenced assessment have been converted into criterion-referenced assessment, and it has been determined that 42% of the grades are free from any change. Twenty-two percent of these grades increase in favor of the norm-referenced assessment, while the rest increases concerning criterion-referenced assessment. In addition, the results of the interviews conducted with the faculty members in the same study have shown that the grading system calculated through the

norm-referenced assessment has a negative impact on the interaction among students. This situation results in negative competition among the students and grouping; moreover, it also decreases the sense of trust towards each other and damages the value education of the students.

Besides, Duman (2011) has emphasized the positive side of the norm-referenced assessment method. In a study conducted with primary school prospective teachers, the norm-referenced assessment has been identified to partially compensate the grading deficiencies emerging due to the faculty members and the exam questions. Similar results have emerged in previous studies. In particular, both classroom assessment questions prepared by teachers and questions in question banks are of low quality (Demir & Atalmış, 2017; Downing, 2005; Masters et al., 2001; Mehrens & Lehmann, 1991; Tarrant et al., 2006). As a result of the analysis on measurement and assessment, both the reliability and validity of the test scores with low-quality questions will decrease and a false assessment mechanism (decision) will be formed depending on the test scores that are calculated incorrectly (Çelen & Aybek, 2013).

Besides, Nartgün (2007) has argued that the norm-referenced assessment would provide an advantage to students after graduation. The use of norm-referenced assessment system in higher education in the countries, especially in Turkey, where competition is common in terms of both university entrance exam and after graduation provides an opportunity to compare the grades of the students graduating from different universities on the same scale. To illustrate, the student who graduated with a score of 60 from a university that requires high score is considered to have near or equal achievement score with the student who has graduated with the score of 80 from a university requiring low achievement score.

In a more explicit way, just as the person who graduated with a lower GPA from University A that admits those with high scores cannot be perceived in the same way as the person who graduated with a higher GPA from University B that accepts those with low scores, the person graduating from University B can be perceived as more successful at first sight. However, the person graduating from University A may have received far more comprehensive, well-equipped and innovative education. Therefore, s/he may have experienced a difficult process and graduated with a low GPA. For this reason, a norm-referenced assessment system may enable to compare two students graduating from these two different universities through using the same scale.

The Assessment Systems Used at Universities in Turkey

Upon reviewing the state universities in Turkey, the norm-referenced assessment system is used as a grading system in the majority of the universities. Considering the universities that use criterion-referenced assessment, they have been identified to hold different pass/fail cutoff score and their letter grades are determined variously.

The letter grade of CC equals to 60 in most of these universities (e.g. Selçuk University, Şırnak University, Uludağ University), one of which is 54 (Amasya University), one is 64 (Abdullah Gül University), while 50 in others (e.g. Recep Tayyip Erdoğan University, Gümüşhane University, and Bayburt University); moreover, the grade of CC corresponds to 65 in some of the universities (e.g. Hacettepe University and Bolu Abant İzzet Baysal University) and 70 in only a small number of them (METU, Boğaziçi University, and Gaziantep University). In particular, the universities that keep CC score range high were observed to have high university entrance grades (e.g. Hacettepe University, Gebze Technical University, Boğaziçi University, and METU) or they have been separated from such universities in the following years. For instance, Gaziantep University may be considered a university separated from METU but still able to protect METU traditions.

Given the instructions of the universities using norm-reference assessment system in Turkey, they use different methods and algorithms depending on the number of students in the class, classroom grade point average, percentiles and standard deviations of grades. While some universities determine the grade of CC, they allow the instructors to intervene in addition to the classroom grade point average (Ankara University and Istanbul Technical University). Besides, criterion-referenced assessment or norm-referenced assessment systems is implemented in many universities according to the grade

interval by calculating T score (Akdeniz University, Aksaray University, Bartın University, Bitlis Eren University, Bursa Technical University, Bülent Ecevit University, Ege University, Fırat University, Hitit University, Karamanoğlu Mehmet Bey University, Kırıkkale University, Kilis 7 Aralık University, Mehmet Akif Ersoy University, Muğla Sıtkı Koçman University, Muş Alparslan University, Tunceli University, Uşak University).

Likewise, the letter grades of the students are given with the percentages of the students in the class in some of the universities that apply T score calculation method (Artvin Çoruh University, Atatürk University, Balıkesir University, Celal Bayar University, Cumhuriyet University, Çankırı Karatekin University, İzmir Katip Çelebi University, Kafkas University, Karadeniz Technical University, Marmara University, Namık Kemal University, Niğde University, Ondokuz Mayıs University, Süleyman Demirel University, Trakya University), while some of the universities determine the letter grades by taking the standard deviation of the class in addition to those mentioned above (Selçuk University, Yalova University).

In some universities that do not use the T score method, a criterion-referenced or norm-referenced assessment system is applied depending on the standard deviation of the grade distribution in the class as well as the number of students and classroom grade point average (İstanbul University, Çukurova University, Harran University, İnönü University, Kahramanmaraş Sütçü İmam University). Similarly, the upper and lower limits of the letter range in these universities are determined by the university administration. These different norm-referenced systems are presented in detail as follows:

Norm-Referenced Assessment System Only Applied According to T Score Conversion

This assessment system initially calculates students' raw success grades (RSG) through using midterm and final exam grades. RSG is calculated by taking 40% of midterm and 60% of the final exam. While determining students who will be included in the norm-referenced assessment system, RSG lower limit and the lower limit of number of students are used. Table 1 displays RSG lower limit and the number of students in universities applying norm-referenced assessment.

Table 1. The Lower Limit of RSG (TLLORSG) and the Number of Students (TLLOTNOS) in Universities Applying Norm-Referenced Assessment

University Name	TLLORSG	TLLOTNOS	University Name	TLLORSG	TLLOTNOS
Adana Sci. and Tech. Uni.	15	11	Iğdır Uni.	20	11
Adıyaman Uni.	20	20	İnönü Uni.	40	11
Ağrı İbrahim Çeçen Uni.	30	10	İskenderun Tech. Uni.	15	30
Akdeniz Uni.	20	15	İstanbul Tech. Uni.	-	-
Aksaray Uni.	35	11	İstanbul Uni.	35	20
Alanya Alaaddin Key. Uni.	20	16	İzmir Katip Çel. Uni.	30	11
Anadolu Uni.	25	30	Kafkas Uni.	40	11
Ankara Uni.	-	30	Kahraman. S. I. Uni.	25	15
Ardahan Uni.	20	10	Karadeniz Tech. Uni.	15	11
Artvin Çoruh Uni.	15	11	Karamanoğlu M. Uni.	20	10
Atatürk Uni.	-	10	Kırıkkale Uni.	15	30
Balıkesir Uni.	15	10	Kırklareli Uni.	20	-
Bandırma Onyedi Ey. Uni.	15	11	Kilis 7 Aralık Uni.	20	11
Bartın Uni.	15	10	Marmara Uni.	20	10
Bilecik Şeyh Edebali Uni.	45	10	Mehmet A. E. Uni.	15	20
Bitlis Eren University	20	20	Mimar Sinan F. A. Uni.	-	30
Bozok Uni.	-	-	Muğla S. K. Uni.	10	30
Bursa Technical Uni.	20	20	Muş Alparslan Uni.	15	10
Bülent Ecevit Uni.	35	25	Namık Kemal Uni.	15	11
Celal Bayar Uni.	20	20	Nevşehir H. B. V. Uni.	-	-
Cumhuriyet Uni.	15	11	Niğde Ö. H. Uni.	10	11

Çankırı Karatekin Uni.	25	10	Ondokuz Mayıs Uni.	20	11
Çukurova Uni.	35	20	Osmaniye K. A. Uni.	20	1
Dokuz Eylül Uni.	-	-	Sakarya Uni.	-	-
Dumlupınar Uni.	15	-	Selçuk Uni.	15	20
Ege Uni.	15	30	Süleyman Demirel Uni.	15	11
Erciyes Uni.	-	-	Trakya Uni.	15	11
Erzincan B. Y. Uni.	-	10	Tunceli Uni.	15	10
Eskişehir Osmangazi Uni.	-	-	Uludağ Uni.	20	-
Fırat University	10	15	Uşak Uni.	20	10
Gazi University	-	-	Yalova Uni.	20	20
Harran University	40	20	Yıldız Tech. Uni.	-	-
Hitit University	30	20	Yüzüncü Yıl Uni.	15	30

As can be seen in Table 1, the lower limit of RSG and the number of students are determined as 15 and 11 in 65 state universities using norm-referenced assessment system in Turkey. Specifically, 11 students with RSG scores greater than 15 are required to use a norm-referenced assessment system. Otherwise, criterion-referenced assessment system is supposed to be used rather than norm-referenced assessment.

Following this stage, students' scores will be converted into T scores by using the following formulas in the norm-referenced assessment system, which requires only "T-score conversion" (Güler, 2017).

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} \quad , \quad \sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}} \quad , \quad T = [(\frac{X_i - \mu}{\sigma}) \times 10] + 50$$

Here, N refers to the number of students participating in the assessment, X_i signifies students' RSG, μ, represents the students' RSG average, σ is the standard deviation of the students' RSG and T is the score converted from students' RSGs. After each student's T score is obtained, the letter grades are given to the students by using the value in Table 2 depending on the RSG average of the class.

Table 2. Calculation of Letter Grades in terms of T Score

Class Level	RSG average of the class	AA (4)	BA (3.5)	BB (3)	CB (2.5)	CC (2)	DC (1.5)	DD (1)	FF (0)
Outstanding	$80<\mu\leq100$	≥ 57	52-56.99	47–51.99	42–46.99	37–41.99	32–36.99	27–31.99	< 27
Excellent	$70<\mu\leq80$	≥ 59	54-58.99	49–53.99	44–48.99	39–43.99	34–38.99	29–33.99	< 29
Very Good	$62.5<\mu\leq70$	≥ 61	56–60.99	51–55.99	46–50.99	41–45.99	36–40.99	31–35.99	< 31
Good	$57.5<\mu\leq62.5$	≥ 63	58–62.99	53–57.99	48–52.99	43–47.99	38–42.99	33–37.99	< 33
Satisfactory	$52.5<\mu\leq57.5$	≥ 65	60–64.99	55–59.99	50–54.99	45–49.99	40–44.99	35–39.99	< 35
Sufficient	$47.5<\mu\leq52.5$	≥ 67	62–66.99	57–61.99	52–56.99	47–51.99	42–46.99	37–41.99	< 37
Poor	$42.5<\mu\leq47.5$	≥ 69	64–68.99	59–63.99	54–58.99	49–53.99	44–48.99	39–43.99	< 39
Fail	$\mu<42.5$	≥ 71	66–70.99	61–65.99	56–60.99	51–55.99	46–50.99	41–45.99	< 41

Table 2 suggests that a class in which the norm-referenced assessment system is applied is in one of eight different levels according to RSG average of the class. To illustrate, the class whose RSG average is between 80 and 100 is considered "outstanding", while the class whose RSG average varies between 57.5 and 62.5 is regarded as "Good". Taking the students' letter grades into account, the letter grade of a student whose T score is 58 and who is in the class with 55 RSG average (Good) is BB; whereas the student with the same T score but in the class with 65 RSG average (Very Good) has BA letter grade.

When the norm-referenced assessment regulations of the universities are examined, most of the universities use the chart as in Table 2 while calculating T score (Celal Bayar University, Kafkas University, Marmara University), while others use criterion-referenced assessment system for upper-level classes. In no uncertain terms, some of the universities with 60 and over (Akdeniz University), 70 and over (Bartın University, İzmir Katip Çelebi University, Muğla Sıtkı Koçman University), 80 and over RSG average (Karadeniz Technical University, Ondokuz Mayıs University, Uşak University) use criterion-referenced assessment system. Table 3 depicts the letter ranges in the criterion-referenced assessment applied in some universities. Besides, some of the universities use norm-referenced assessment system by decreasing the number of class levels (below 8) and enlarging the class level intervals in Table 2 (Bülent Ecevit University, Çankırı Karatekin University, Niğde University, Süleyman Demirel University).

Table 3. Letter Intervals of the Universities Regarding Criterion-Referenced Assessment System

4-point Grading System	Letter Grade	Aksaray Uni. RSG Intervals	Akdeniz Uni. RSG Intervals	Karadeniz Technical Uni. RSG Intervals	Selçuk Uni. RSG Intervals	İstanbul Uni. RSG Intervals
4.00	AA	90 – 100	87.5 – 100	90 – 100	88 – 100	88 – 100
3.50	BA	85 – 89	80.5 – 87.4	80 – 89	80 – 87	80 – 87
3.00	BB	80 – 84	73.5 – 80.4	75 – 79	73 – 79	73 – 79
2.50	CB	70 – 79	66.5 – 73.4	70 – 74	66 – 72	66 – 72
2.00	CC	60 – 69	59.5 – 66.4	60 – 69	60 – 65	60 – 65
1.50	DC	55 – 59	52.5 – 59.4	50 – 59	55 – 59	55 – 59
1.00	DD	50 – 54	45.5 – 52.4	40 – 49	50 – 54	50 – 54
0.50	FD	40 – 49	34.5 – 45.4	30 – 39	–	–
0.00	FF	0 – 45	0 – 34.4	0 – 29	0 – 39	0 – 49

Norm-Referenced Assessment System Applying T Score Conversion and Quantiles

This assessment system is based on the number of students participating in the norm-referenced assessment. Upon examining the norm-referenced assessment regulations of the universities applying this assessment system, only T score conversion is conducted in cases when the number of students participating in the norm-referenced assessment is 30 or over just as in Table 2, while the letter grades are based on quantiles as in Table 4 when the number of the students is between 10 (11 in some of the universities) and 29 (30 in some of the universities).

Table 4. The Calculation of Letter Grades Depending on Quantiles

Class Level	RSG average of the class	AA (4)	BA (3.5)	BB (3)	CB (2.5)	CC (2)	DC (1.5)	DD (1)	FF (0)
Outstanding	$70<\mu\leq100$	24(24)	15.2(39.2)	22.8(62)	11.6(73.6)	17.4(91)	4.8(95.8)	3.2(99)	1(100)
Excellent	$62.5<\mu\leq70$	18(18)	14.4(32.4)	21.6 (54)	12.8 (66.8)	19.2 (86)	7.2(93.2)	4.8(98)	2(100)
Very Good	$57.5<\mu\leq62.5$	14(14)	12.8 (26.8)	19.2 (46)	14.4 (60.4)	21.6 (82)	9(91)	6(97)	3(100)
Good	$52.5<\mu\leq57.5$	10(10)	11.6(21.6)	17.4(39)	14.8(53.8)	22.2 (76)	12(88)	8(96)	4(100)
Satisfactory	$47.5<\mu\leq52.5$	7(7)	9.6 (16.6)	14.4 (31)	15.2 (46.2)	22.8 (69)	14.4 (83.4)	9.6(93)	7(100)
Sufficient	$42.5<\mu\leq47.5$	4(4)	8(12)	12(24)	14.8 (38.8)	22.2(61)	17.4 (78.4)	11.6(90)	10(100)
Poor	$\mu<42.5$	3(3)	6 (9)	9(18)	14.4 (32.4)	21.6 (54)	19.2 (73.2)	12.8(86)	14(100)

* The values in parentheses indicate the percentage of the cumulative percentages.

First, the percentage of the students participating in the norm-referenced assessment is calculated while determining their letter grades and then their letter grades are identified through using Table 4. For instance, in a class where the RSG average is 60, the letter grade of a student in the top 10% is AA, while that of a student in the top 30% is identified as BB.

Norm-Referenced Assessment System Implementing T-Score Conversion, Quantiles, and Standard Deviation

Within this system, the RSGs' standard deviation of some students participating in the norm-referenced assessment in some universities is calculated in addition to the T-score conversion and the percentile method. If the standard deviation is below a certain value, a criterion-referenced assessment system is used. This value ranges between 4 (such as Yalova University) and 8 (Selçuk University) based on the regulations of the universities.

Standard Deviation Based Norm-Referenced Assessment System

This assessment system holds criterion or norm-referenced assessment systems by focusing on the standard deviation of the grade distribution in the classroom as well as the average of the student group (class) participating in the norm-referenced assessment. As indicated in the study conducted by Nartgün (2007), the effectiveness of the criterion or norm-referenced assessment systems depends on the standard deviation. In this system, moreover, criterion or norm-referenced assessment is applied depending upon the lower limit of the number of students participating in the norm-referenced assessment. This varies between 15 and 20 according to the regulations of the universities. Nevertheless, in this assessment system, the criterion-referenced assessment system is a prerequisite when the standard deviation of the grades in the class is below 8 (grade distributions are close to each other). In applying this system, RSG is calculated as the sum of 40% of the student's mid-term scores and 60% of their final scores.

Table 5 presents the letter grade determination table of the İstanbul University which first applied this norm-referenced assessment system.

Table 5. İstanbul University Letter Grade Calculation through Norm-Referenced Assessment System

Letter Grade	Very Poor: μ<44	Poor: 44≤μ<50	Below: 50≤μ<56	Average: 56≤μ<63
AA	[μ+1.881σ, 100]	[μ+1.645σ, 100]	[μ+1.476σ, 100]	[μ+1.227σ, 100]
BA	[μ+1.405σ, μ+1.881σ)	[μ+1.175σ, μ+1.645σ)	[μ+0.994σ, μ+1.476σ)	[μ+0.739σ, μ+1.227σ)
BB	[μ+0.706σ, μ+1.405σ)	[μ+0.524σ, μ+1.175σ)	[μ+0.358σ, μ+0.995σ)	[μ+0.126σ, μ+0.739σ)
CB	[μ+0.332σ, μ+0.706σ)	[μ+0.126σ, μ+0.524σ)	[μ-0.075σ, μ+0.358σ)	[μ-0.358σ, μ+0.126σ)
CC	[μ-0.176σ, μ+0.332σ)	[μ-0.468σ, μ+0.126σ)	[μ-0.772σ, μ-0.075σ)	[μ-0.878σ, μ-0.358σ)
DC	[μ-0.643σ, μ-0.176σ)	[μ-0.878σ, μ-0.468σ)	[μ-1.126σ, μ-0.772σ)	[μ-1.227σ, μ-0.878σ)
DD	[μ-1.175σ, μ-0.643σ)	[μ-1.405σ, μ-0.878σ)	[μ-1.645σ, μ-1.126σ)	[μ-1.751σ, μ-1.227σ)
FF	[35, μ-1.175σ)	[35, μ-1.405σ)	[35,μ-1.645σ)	[35,μ-1.751σ)
Letter Grade	Above Average: 63≤μ<71	Good: 71≤μ<80	Very Good: μ≥80	*For the absence of conflicts, the intervals are shown as"[" indicating "included" and closed from the left side, and ")" referring to "excluded" and open from the right side. In the table, μ suggests the average of the RSG values and σ shows the standard deviation of these values.
AA	[μ+0.915σ, 100]	[μ+0.583σ, 100]	[μ+0.440σ, 100]	
BA	[μ+0.385σ, μ+0.915σ)	[μ+0.100σ, μ+0.583σ)	[μ-0.100σ, μ+0.440σ)	
BB	[μ-0.075σ, μ+0.385σ)	[μ-0.305σ, μ+0.100σ)	[μ-0.496σ, μ-0.100σ)	
CB	[μ-0.524σ, μ-0.075σ)	[μ-0.739σ, μ-0.305σ)	[μ-0.915σ, μ-0.496σ)	
CC	[μ-0.994σ, μ-0.524σ)	[μ-1.126σ, μ-0.739σ)	[μ-1.282σ, μ-0.915σ)	
DC	[μ-1.341σ, μ-0.994σ)	[μ-1.476σ, μ-1.126σ)	[μ-1.645σ, μ-1.282σ)	
DD	[μ-1.881σ, μ-1.341σ)	[μ-2.054σ, μ-1.476σ)	[μ-2.326σ, μ-1.645σ)	
FF	[35,μ-1.881σ)	[35,μ-2.054σ)	[35, μ-2.326σ)	

Here μ represents classroom average and σ signifies the standard deviation of the distribution in the class. Formulas of μ and σ are presented as follows (Field, 2009).

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}, \qquad \sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}}$$

In this formula, X_i represents the RSG of a student participating in the norm-referenced assessment, while *N* shows the total number of the students participating in this assessment.

Table 3 indicates the criterion-referenced assessment letter intervals in İstanbul University when the norm-referenced assessment system is not applied. Criterion-referenced assessment is applied in cases where the number of students in the class is less than 10, and/or the standard deviation of the grade distribution in the class is below 8.

The Significance and Aim of Research

Unlike previous studies focusing a particular norm-referenced system in Turkey (Atılgan et al., 2012; Duman, 2011; Nartgün, 2007), this research has compared aforementioned four different norm-referenced assessment systems that are widely used in Turkey. In this regard, Aksaray University, which uses norm-referenced assessment for each grade level among the universities that apply only T-score conversion, and the norm-referenced assessment system of Akdeniz University (for the classes with over 60 RSG mean) that applies criterion-referenced assessment systems for the upper-level classes have been chosen. Among the universities that use T-score conversion and quantile method, the sample of the norm-referenced assessment system of Karadeniz Technical University has been chosen; Selçuk University's norm-referenced assessment system has been preferred as an example for the university using T score conversion, quantile, and standard deviation. Finally, İstanbul University norm-referenced assessment system has been chosen as it does not use T-score conversion and only uses a standard deviation-based conversion.

In order to achieve this goal, answers to the following questions have been sought:

1. Is there a statistically significant difference between students' letter grades calculated through norm-referenced assessment and criterion-referenced assessment?
2. Is there a statistically significant difference between students' letter grades calculated by different norm-referenced assessment systems?

METHOD

This study can be considered a causal-comparative research approach, seeking to determine differences between groups by examining differences in the experiences of group members (Lodico, Spaulding, & Voegtle, 2010). In this study, the groups of individuals are students whose letter grades calculated through different norm-referenced assessment and criterion-referenced assessment. This section holds information regarding the research sample, process, and data analysis.

Research Sample

The research data have been collected through the midterm and final grades of the students studying at different faculties and vocational colleges in a state university during the fall semester of 2014-2015, and the total of 19,574 students' RSGs have been considered during data analysis.

Process

Students' RSGs have been calculated by taking into account 40% of the midterm score and 60% of the final score. Afterward, the RSGs of the students have been calculated as letter grades by adapting them depending on the above-mentioned assessment systems of Aksaray University, Akdeniz University, Karadeniz Technical University, Selçuk University, and İstanbul University. Besides, the RSGs of the students have been converted into letter grades considering the absolute assessment table of these universities as displayed in Table 2. Thus, each student's letter grade calculated by means of both norm-referenced and criterion-referenced assessment has been determined and then converted to the

grade equivalent to 4- point grading system. Given the students took more than one course in the 2014-2015 fall semester, the same process has been performed for each course's RSG taken by each student. In other words, the analysis in this research has been conducted for 19,574 students' 157,983 letter grade.

Data Analysis

In order to identify the difference between norm-referenced and criterion-referenced assessment, which is the first research question, the paired t-test was used to analyze the difference between the two groups. For the second research question, the differences between the letter grades obtained by the norm-referenced assessment system of each of the universities mentioned above have been analyzed through one-way analysis of variance (ANOVA).

RESULTS

Figure 1 depicts the distribution of students' RSGs, are not converted scores through norm-referenced/criterion-referenced assessment.

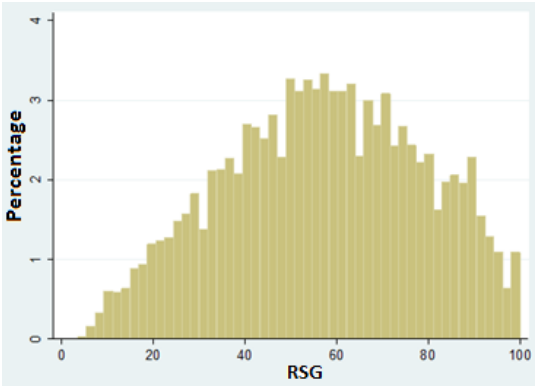


Figure 1. Students' End-of-Term Raw Success Grade (RSG) Distribution

Upon examining Figure 1, the skewness coefficient of the students' RSGs distribution was identified to be -0.109 and this value was considered normal because it is between -1 and +1 (Büyüköztürk, 2009).

In the next stage, norm-referenced and criterion-referenced assessment systems used by Aksaray University, Akdeniz University, Karadeniz Technical University, Selçuk University, and İstanbul University have been applied for these RSGs.

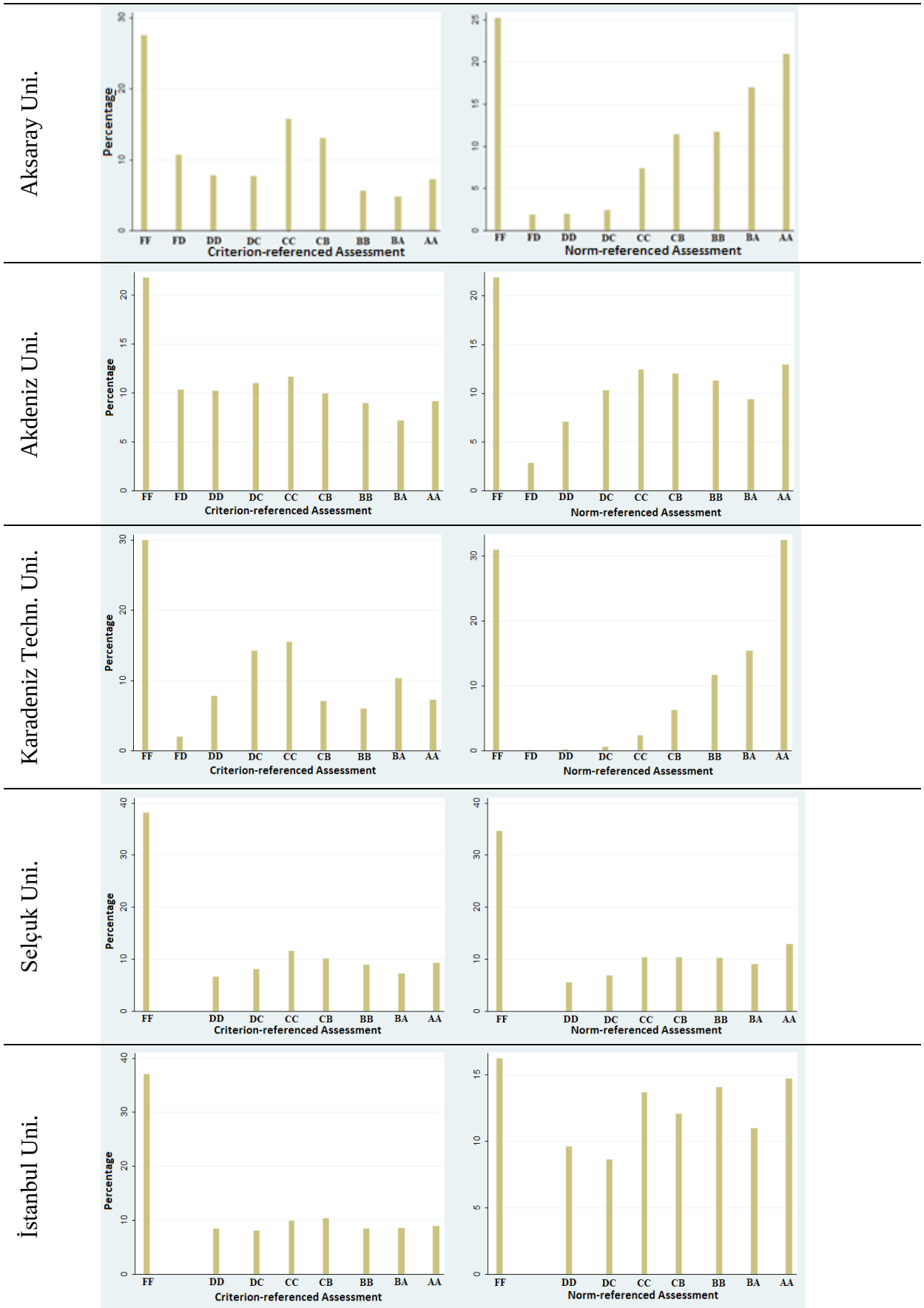


Figure 2. Criterion-Referenced Norm-Referenced Assessment Grade Distribution of the Universities

Figure 2 suggests that norm-referenced assessment generally increases letter grades of the students contrary to criterion-referenced assessment. As a result of the paired t-test, the norm-referenced assessment mean score of the students ($\bar{x}$=2.21, SD=1.51) has been noted to statistically differ from that of the criterion-referenced assessment ($\bar{x}$=1.59, SD=1.37) (t(789914)=-590.92, p<.05). Table 6 displays t-test results conducted for each university.

Table 6. t-Test Results Regarding Criterion-Referenced and Norm-Referenced Assessment for Different Universities

Universities	N	Criterion-referenced assessment		Norm-referenced assessment		df	*T*
		$\bar{x}$	SS	$\bar{x}$	SS		
Aksaray Uni.	157983	1.51	1.30	2.28	1.54	157982	-298.60*
Akdeniz Uni.	157983	1.68	1.33	1.97	1.37	157982	-271.74*
Karadeniz Technical Uni.	157983	1.62	1.34	2.41	1.69	157982	-335.21*
Selçuk Uni.	157983	1.57	1.44	1.77	1.50	157982	-129.00*
İstanbul Uni.	157983	1.58	1.44	2.20	1.32	157982	-374.76*

*p<.05

As shown in Table 6, the differences between the criterion-referenced and norm-referenced assessment scores for each university have been determined to be statistically significant and the scores increase in direction of norm-referenced assessment. ANOVA has been applied to explore the difference between the scores obtained from different norm-referenced assessment systems, which is related to other research question. Accordingly, a statistically significant difference has been determined ($F_{(4, 789910)}$=4632.88, p<.05). As a result of the post-hoc test (LSD), the scores calculated with the norm-referenced assessment score of each university are statistically different from those of the other university. The results are presented in Table 7.

Table 7. Comparison of Norm-Referenced Difference Scores among Universities

University (I)	University (J)	Difference (I-J)
Aksaray Uni.	Akdeniz Uni.	.312*
	Karadeniz Tech. Uni.	-.123*
	Selçuk Uni.	.516*
	İstanbul Uni.	.087*
Akdeniz Uni.	Karadeniz Tech. Uni.	-.434*
	Selçuk Uni.	.205*
	İstanbul Uni.	-.225*
Karadeniz Tech. Uni.	Selçuk Uni.	.634*
	İstanbul Uni.	.209*
Selçuk Uni.	İstanbul Uni.	-.430*

*p<.05

According to Table 7, the university which applies the most advantageous norm-referenced assessment system has been determined to be Karadeniz Technical University, which is followed by Aksaray University, Istanbul University, Akdeniz University, and Selçuk University, respectively.

RESULTS AND DISCUSSION

This research aims to explore whether students' letter grades differ across norm-referenced and criterion-referenced assessment methods and how this difference varies across universities. In this regard, the end-of-term raw achievement scores of 19,574 students who study at a state university during the fall term of 2014-2015 academic year have been converted into letter grades and 4-point

grading systems through use of both norm-referenced and criterion-referenced assessment regulations of the universities. After applying the paired samples t-test, the letter grades calculated via the norm-referenced assessment have been identified to be statistically significant and high compared to those calculated through the criterion-referenced assessment. In the following stage, ANOVA has been used in order to determine the difference between the letter grades obtained by using the norm-referenced assessment systems of the universities and the result has been found to be statistically significant. Specifically, students with the same RSG appear to have different letter grades in different universities.

The research findings have notably shown that the students' letters grades decrease as standard deviation in the norm-referenced assessment systems and low cut-off scores in criterion-referenced assessment systems are used. To exemplify, considering the lower grades obtained through Selçuk University norm-referenced assessment system, criterion-referenced assessment system are used for classes with RSG mean of over 70 and/or standard deviation below 8. Likewise, criterion-referenced assessment is used for the classes whose RSG mean is over 60 in Akdeniz University norm-referenced assessment system, while the same system is used for classes whose RSG mean is over 90 and/or standard deviation is below 8 in İstanbul University. In the norm-referenced assessment system of Aksaray University, criterion-referenced assessment is not applied depending on the RSG mean, whereas criterion-referenced assessment is used for the classes with 80 and over RSG meaning the norm-referenced assessment system applied by Karadeniz Technical University. However, 80-89 scores in the system applied by this university refer to BA and the scores above 90 signify AA letter grade, which increases students' letter grades.

When the research results are considered in general terms, the norm-referenced assessment has been determined to be much more in favor of students' letter grades compared to the criterion-referenced assessment. A similar result has emerged in the study conducted by Sayın (2016); however, different results have been found by Atılgan et al. (2012). This may result from the different use of the norm-referenced assessment algorithms and the small size sample group.

Based on the results of this research, two different evaluations can be made in terms of students and academic staff. On the basis of the student's perspective, students have higher grade point averages with the norm-referenced assessment than criterion-referenced assessment. This paves the way for the fact that norm-referenced assessment will lead to grading inflation as indicated by Başol (2015) in the related studies. Besides, different norm-referenced assessment systems reveal that the same RSG has been converted into different letter grades, meaning that the universities applying norm-referenced assessment system are more advantageous compared to the others. Both the difference between the norm-referenced and criterion-referenced grading system and the difference between the norm-referenced assessment systems may cause injustice. The concrete indicator of this situation occurs when the students apply to graduate programs. The degree effect of the GPA is used up to 40% in some universities is the evidence of the injustice of assessment systems used in the universities.

Upon analyzing the research results in terms of academic staff, the norm-referenced assessment system will be able to reduce the grading errors that will arise from the structure of the tests prepared by the academic staff. Considering that the academic staff may be lacking preparing a sufficient test or question technique, errors emerging due to the structure of the test, such as misinterpretation of the question or not being included in the current program, will cause students to get poor grades. However, since the findings of the previous studies (Nartgün, 2007; Sayın, 2016) and this study provide the conclusion that the norm-referenced assessment increases the grades of students, the norm-referenced assessment is likely to convert these lower grades, especially those of academic staff, into higher letter grades. The point to be noted here is that some students pass through the class without deserving it or being in high letter ranges. Thus, it is preferable to have an adequate level of the students included in the norm-referenced assessment and to determine the letter intervals rigorously.

What is more, it is of high significance to decide whether to use norm-referenced or criterion-referenced assessment depending on the purpose, structure, and results of the exams in terms of efficient measurement and assessment. Turgut and Baykul (2015) have noted that the scores will be

distributed symmetrically when there is a normal distribution; otherwise, the grades of the students will be largely affected by the other people in the class. Because the extreme values in both the right skewed and left skewed grades will change the mean of the distribution as well as increasing the standard deviation (Turgut & Baykul, 2015).

Besides, the use of criterion-referenced assessment will lead to different results in the case of the skewness of the raw score distributions. As the grade point averages will be high in the left-skewed distributions, many people in the class will pass with high letter grades, while in a class with a right-skewed distribution, the grades will be lower, thus the letter grades will be low and many students may fail. Given the skewness of the grades derives from the fact that the exam questions are too difficult or too easy, it is probable that the academic staff does not prepare qualified questions in terms of measurement and assessment. Thorndike (2005) draws attention to the fact that while preparing a qualified test, 25% of the questions should be difficult, 50% of them are at a medium level and 25% easy. Thus, the distribution of the grades will be closer to the normality.

Yücel (2015) has stated that the national exams in Turkey such as university entrance for which norm-referenced assessment is used measure the objectives at the level of remembering, understanding and applying levels. The use of blueprint in the exams prepared by the academic staff may affect the distribution of grades. In particular, writing the questions that will measure the cognitive gains of the students in all levels will determine how much the student has learned. The most commonly used type of question, open-ended questions and project-based assignments, which are the most commonly used type of questions, will provide the students with the objectives that need to be gained in the upper levels, namely, analysis, synthesis, and evaluation. It is expected that the number of questions will be higher in the exams prepared in this direction, as a result of which both the scope validity of the exam will increase and the grades obtained as a result of the application of the questions will be expected to be distributed normally. In other words, the exam consisting of questions related to all cognitive levels (knowledge, comprehension, application, analysis, synthesis, and evaluation) will undoubtfully affect the difficulty of the questions, and as Thorndike (2005) stated, it will cause the questions to be distributed normally in terms of the degree of difficulty. In short, the preparation of an exam within the framework of measurement and evaluation will affect the grade distribution, and consequently, exams can be evaluated through criterion-referenced assessment without the need for norm-referenced assessment.

REFERENCES

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Atılgan, H., Yurdakul, B., & Oğretmen, T. (2012). A research on the relative and absolute evaluation for determination of students achievement. *Inonu University Journal of the Faculty of Education, 13*(2), 79-98.

Başol, G. (2013). *Eğitimde ölçme ve değerlendirme*. Ankara: Pegem Akademi Yayınları.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Büyüköztürk, Ş. (2009). Sosyal bilimler için veri analizi el kitabı: İstatistik, araştırma deseni, SPSS uygulamaları ve yorum (9. baskı). Ankara: Pegem Yayınları.

Çelen, Ü., & Aybek, E. C. (2013). Öğrenci başarısının öğretmen yapımı bir testle Klasik Test Kuramı ve Madde Tepki Kuramı yöntemleriyle elde edilen puanlara göre karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *4*(2), 64-75.

Demir, P. & Atalmış, E.A. (2017). Öğretmen Yazılı Sınav Sorularının Hess Bilişsel Zorluk Matrisine Göre İncelenmesi. Köse, Selçuk, & Atalmış (Eds.), *Sosyo Ekonomik Stratejiler III – Eğitim* içinde (s. 43-73). Londra: IJOPEC Publication.

Duman, B. (2011). The views of classroom teachers related to norm-referenced assessment. *Education Sciences, 6*(1), 536-548.

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education, 10*(2), 133-143. DOI 10.1007/s10459-004-4019-5.

Field, A. (2009). *Discovering statistics using SPSS*. Thousand Oaks, CA: Sage publications.

Güler, N. (2017). *Eğitimde ölçme ve değerlendirme*. Ankara: Pegem Akademi Yayınları.
Haladyna, T.M. & Rodriguez, M.C (2013). *Developing and validating test items*. New York: Routledge.
Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research, 48*(1), 1-47.
Johnson, V. E. (2003). Grade inflation: a crisis in college education. New York, NY: Springer.
Lodico, M. G., Spaulding, D. T., & Voegtle, K. H. (2010). *Methods in educational research: From theory to practice*. San Francisco, CA: John Wiley & Sons.
Martin, I. G., & Jolly, B. (2002). Predictive validity and estimated cut score of an objective structured clinical examination (OSCE) used as an assessment of clinical skills at the end of the first clinical year. *Medical education, 36*(5), 418-425. https://doi.org/10.1046/j.1365-2923.2002.01207.x
Masters, J. C., Hulsmeyer, B. S., Pike, M. E., Leichty, K., Miller, M. T., & Verst, A. L. (2001). Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. *Journal of Nursing Education, 40*(1), 25-32. https://doi.org/10.3928/0148-4834-20010101-07
Mehrens, W.A. & Lehmann, I.J. (1991*). Measurement and evaluation in education and psychology.* New York: Harcourt Brace.
Nartgün, Z., (2007). Aynı puanlar üzerinden yapılan mutlak ve bağıl değerlendirme uygulamalarının notlarda farklılık oluşturup oluşturmadığına ilişkin bir inceleme. *Ege Eğitim Dergisi, 8* (1), 19- 40.
Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education, 30*(2), 175-194. https://doi.org/10.1080/0260293042000264262
Sayın, A. (2016). The Effect of using relative and absolute criteria to decide students' passing or failing a Course. *Journal of Education and Training Studies*, *4*(9), 1-9. http://dx.doi.org/10.11114/jets.v4i9.1571
Selvi, K. (1998). Üniversitelerde uygulanan başarı değerlendirme yaklaşımları. *Kurgu Dergisi*, *15*, 336-345
Tan, Ş. (2015). Öğretim hedeflerinin belirlenmesi. Şeref Tan (Ed.), *Öğretim ilke ve yöntemleri* içinde *(s.38-76)*. Ankara: Pegem Akademi Yayınları.
Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse education in practice*, *6*(6), 354-363. DOI:10.1016/j.nedt.2006.07.006
Thorndike, R.M. (2005). *Measurement and Evaluation in Psychology and* Education (7th Ed.). Upper Saddle River, NJ: Pearson Education.
Turgut, M. F., & Baykul, Y. (2015). *Eğitimde ölçme ve değerlendirme*. Pegem Akademi Yayıları.
Yorke, M. (2011). Summative assessment: dealing with the 'measurement fallacy'. *Studies in Higher Education, 36*(3), 251-273. https://doi.org/10.1080/03075070903545082
Yücel, C. (2015). Sınıf İçi Değerlendirme ve Not Verme. Emin Karip (Ed.), *Ölçme ve değerlendirme* içinde (s. 324-361). Ankara: Pegem Akademi Yayınları.

Türkiye'de Yükseköğretimde Kullanılan Bağıl Değerlendirme Sistemlerinin İstatistiksel Olarak Karşılaştırılması

Giriş

Eğitim ve öğretimin amacı, kazandırılmak istenen bilişsel, duyuşsal ve psikomotor becerileri öğrenciye planlı ve programlı bir şekilde sunmaktır. Bunu sağlamak için öncelikle dört temel öğeden oluşan öğretim programının belirlenmesi gerekmektedir. Öğretim programının bu öğeleri (a) hedef ve davranışların belirlemek, (b) içeriği bu hedef ve davranışlara tutarlı ve öğrencilerin hazırbulunuşluklarına uygun olarak yapılandırmak, (c) her öğrenci farklı öğrenir düşüncesiyle öğrenme ve öğretme aktiviteleri oluşturmak ve (d) anlamlı bir ölçme ve değerlendirme yapmaktır (Tan, 2015). Özellikle programda belirlenen hedef ve davranışların öğrencilerin hazırbulunuşluklarına ne derece sahip olduğu ve yine öğrenme ve öğretme aktivitelerinin hedef ve davranışlara ne derece uygun olduğunu belirlemede ölçme ve değerlendirmenin önemi göz ardı edilememektedir.

Ölçme ve değerlendirme kavramları sürekli beraber kullanılmasına rağmen, birbirinden farklı ve hatta birbirini tamamlayan kavramlardır. Özellikle ölçme bir değişkeni veya bir nesneyi sembollerle ifade ederken, değerlendirme ise ölçmeden elde edilen sonuçları bir ölçütle kıyaslayarak bu sonuçların anlamlaştırılmasını sağlamaktadır. Özellikle değerlendirmenin önemli bir öğesi olan ölçütü belirlemek oldukça karmaşık ve problemli bir süreç olduğu kabul edilmektedir. Ölçüt belirlemenin öğretmen kanısına, sınıftaki öğrenci başarı dağılımına, öğrenci yeteneğine, öğrencinin programdaki erişisine

(programın başındaki ve sonundaki öğrenme farkı) ve programın hedeflerine göre değişeceği önceki çalışmalarda belirtilmektedir (Turgut & Baykul, 2015; Martin & Jolly, 2002; Yorke, 2011). Bu durumlar göz önüne alındığında, doğru bir değerlendirme yapabilmenin en önemli unsurlarından bir tanesi uygun bir ölçütün seçilmesidir. Zaten değerlendirme kullanılan ölçüte göre sınıflandırılmaktadır. Mutlak ölçütün kullanıldığı diğer bir ifade ile grup ve grup özelliklerine bakılmaksızın herkes tarafından aynı şekilde kabul edilen değerlendirmeye mutlak değerlendirme (ölçüt dayanaklı) adı verilmektedir. Grup ve özellikle grubun başarı ortalamasına bağlı olarak seçilen ölçüte bağıl ölçüt ve yapılan değerlendirmeye de bağıl değerlendirme (norm dayanaklı) denilmektedir.

Türkiye'de devlet üniversitelerine bakıldığında yarıdan fazla üniversitede not verme sistemi olarak bağıl değerlendirme kullanıldığı görülmektedir. Mutlak değerlendirme kullanan üniversitelere bakıldığında farklı geçme ve kalma notlarına sahip olduğu ve harf aralıklarının farklı şekilde belirlendiği görülmektedir. Türkiye' deki bağıl değerlendirme sistemi kullanan üniversitelerin yönergeleri incelendiğinde sınıftaki öğrenci sayısı, sınıf not ortalaması, yüzdelik dilimleri ve notların standart sapmasına göre farklı yöntem ve algoritmalar kullandığı görülmektedir. Bazı üniversiteler CC notunu belirlerken sınıf ortalamasının yanında öğretim elemanlarının müdahale etmesine de müsaade etmektedir (Ankara Üniversitesi ve İstanbul Teknik Üniversitesi). Bunun yanı sıra bağıl değerlendirme kullanan birçok üniversitede bağıl değerlendirme sistemi uygulanırken öğrencilerin T puanının hesaplanarak sınıftaki not aralığına göre mutlak ya da değerlendirme sistemi kullanılmaktadır (Akdeniz Üniversitesi, Aksaray Üniversitesi, Bartın Üniversitesi, Bitlis Eren Üniversitesi, Bursa Teknik Üniversitesi, Bülent Ecevit Üniversitesi, Ege Üniversitesi, Fırat Üniversitesi, Hitit Üniversitesi, Karamanoğlu Mehmet Bey Üniversitesi, Kırıkkale Üniversitesi, Kilis 7 Aralık Üniversitesi, Mehmet Akif Ersoy Üniversitesi, Muğla Sıtkı Koçman Üniversitesi, Muş Alparslan Üniversitesi, Tunceli Üniversitesi, Uşak Üniversitesi gibi). Yine bu T puanı hesaplama yöntemini uygulayan üniversitelerin bazılarında ise öğrencilere verilen harf notları öğrencilerin sınıf içerisindeki yüzdelik dilimleri ile beraber düşünülerek verilirken (Artvin Çoruh Üniversitesi, Atatürk Üniversitesi, Balıkesir Üniversitesi, Celal Bayar Üniversitesi, Cumhuriyet Üniversitesi, Çankırı Karatekin Üniversitesi, İzmir Katip Çelebi Üniversitesi, Kafkas Üniversitesi, Karadeniz Teknik Üniversitesi, Marmara Üniversitesi, Namık Kemal Üniversitesi, Niğde Üniversitesi, Ondokuz Mayıs Üniversitesi, Süleyman Demirel Üniversitesi, Trakya Üniversitesi gibi) bir kısmında ise bunlara ilave olarak sınıfın standart sapması göz önünde bulundurularak verilmektedir (Selçuk Üniversitesi, Yalova Üniversitesi gibi). T puan yöntemini kullanmayan bazı üniversitelerde ise sınıftaki öğrenci sayısı ve sınıf not ortalamasının yanında sınıftaki not dağılımının standart sapmasına göre mutlak ya da bağıl değerlendirme sistemi uygulanmaktadır (İstanbul Üniversitesi, Çukurova Üniversitesi, Harran Üniversitesi, İnönü Üniversitesi, Kahramanmaraş Sütçü İmam Üniversitesi gibi). Yine bu üniversitelerde harf aralıklarının alt ve üst sınırları üniversite yönetimi tarafından belirlenmektedir.

Önceki çalışmalardan farklı olarak mevcut çalışmada Türkiye'de yaygın olarak kullanılan ve yukarıda bahsedilen 4 farklı bağıl değerlendirme sistemi karşılaştırılmıştır. Bu bağlamda sadece T puan dönüşümü uygulayan üniversiteler arasından her sınıf düzeyi için bağıl değerlendirme kullanan Aksaray Üniversitesi ve üst düzey sınıflar için mutlak değerlendirme sistemi uygulayan Akdeniz Üniversitesi'nin bağıl değerlendirme sistemi (ham başarı puanların (HBP) ortalaması 60 üstü sınıflar için) örneği seçilmiştir. T puan dönüşümü ve yüzdelik dilim yöntemi kullanılan üniversiteler arasından Karadeniz Teknik Üniversitesi'nin bağıl değerlendirme sistemi örneği seçilirken; T puan puan dönüşümü, yüzdelik dilim ve standart sapma kullanan üniversiteye örnek olarak ise Selçuk üniversitesi bağıl değerlendirme sistemi örneği seçilmiştir. Son olarak T puan dönüşümü kullanmayıp sadece standart sapma tabanlı bir dönüşüm kullanan üniversite olarak ise İstanbul bağıl değerlendirme sistemi örneği seçilmiştir.

Bu bağlamda aşağıdaki sorulara cevap aranacaktır:

1. Bağıl değerlendirme ve mutlak değerlendirme ile hesaplanan öğrenci harf notları arasında istatistiksel olarak fark var mıdır?
2. Farklı bağıl değerlendirme sistemleri ile hesaplanan öğrenci harf notları arasında istatistiksel olarak anlamlı bir fark var mıdır?

Yöntem

Araştırmada veri seti olarak 2014-2015 güz dönemindeki bir devlet üniversitesindeki farklı fakülte ve yüksekokullarındaki tüm öğrencilerin vize ve final notları kullanılmış, toplam 19,574 öğrencilerin HBP'leri veri analizi için göze önüne alınmıştır.

Öğrencilerin HBP'leri ara sınavın %40' ı final sınavının %60' ı alınarak hesaplanmıştır. Ardından öğrencilerin HBP'leri yukarıda bahsedilen Aksaray Üniversitesi, Akdeniz Üniversitesi, Karadeniz Teknik Üniversitesi, Selçuk Üniversitesi ve İstanbul Üniversitesi'nin uygulandığı bağıl değerlendirme sistemine göre uyarlanarak harf notu olarak hesaplanmıştır. Ayrıca bu çalışmada öğrencilerin HBP'leri yine bu üniversitelerin mutlak değerlendirme yönergelerine göre yeniden harf notuna dönüştürülmüştür. Her bir öğrencinin hem bağıl değerlendirmeyle hesaplanan hem de mutlak değerlendirme ile hesaplanmış harf notu ardından 4'lük sistemdeki not karşılığına çevrilmiştir. Öğrencilerin 2014-2015 güz döneminde birden fazla ders aldığı düşünüldüğünde aynı işlem her bir öğrencinin aldığı tüm derslerin HBP'ları için yapılmıştır. Diğer bir ifade ile bu süreç 19,574 öğrencinin 157,983 adet harf notu için yapılmıştır.

Öğrencilerin mutlak ve bağıl harf notlarının 4'lük sistemindeki karşılıkları hesaplandıktan sonra, ilk araştırma sorusu olan bağıl ve mutlak değerlendirme arasındaki farkı bulmak için tekrarlı ölçümlerindeki değişimi araştıran eşleştirilmiş iki grup arasındaki farkların testi (paired t-test) yöntemi kullanılmıştır. İkinci araştırma sorusu için yukarıda adı geçen herbir üniversitenin bağıl değerlendirme sistemiyle elde edilen harf notları arasındaki farklar ise tek yönlü varyans analizi (ANOVA) yöntemi ile bulunmuştur.

Sonuç ve Tartışma

Bu çalışmanın amacı öğrencilerin harf notlarının bağıl ve mutlak değerlendirme kullanılarak farklılık gösterip göstermediğini ve farklı bağıl değerlendirme sistemleri ile hesaplanan öğrenci harf notları arasında istatistiksel olarak anlamlı farklılık olup olmadığını bulmaktır. Bu bağlamda bir devlet üniversitesinde okuyan 19,574 öğrencinin her bir ders için dönem sonu ham başarı puanları farklı bağıl ve mutlak değerlendirme sistemine göre harf notuna ve ardından 4'lük sistemdeki not karşılığına çevrilmiştir. Araştırma sonunda, bağıl değerlendirme ile hesaplanan harf notların mutlak değerlendirme ile hesaplanan notlara göre istatistiksel olarak anlamlı ve yüksek olduğu elde edilmiştir. Ayrıca farklı bağıl değerlendirme sistemleri ile hesaplanan öğrenci harf notları arasında istatistiksel olarak anlamlı bir farklılık olduğu bulunmuştur.

Özellikle çalışmanın bulgularından üniversitelerin bağıl değerlendirme sistemlerinde standart sapma kullanımının yanında mutlak değerlendirmenin kullanılacağı kesme puanları düştükçe öğrenci harf notlarının düştüğü görülmektedir. Örneğin, Selçuk üniversitesi bağıl değerlendirme sistemi ile elde edilen notların düşük olması göz önüne alındığında, bu sistemin HBP ortalaması 70 üstü ve/veya standart sapması 8'in altında olan sınıflar için mutlak değerlendirme kullanıldığı görülmektedir. Yine Akdeniz üniversitesi bağıl değerlendirme siteminde HBP ortalaması 60 üstü sınıflar içinde mutlak değerlendirme kullanılırken, İstanbul üniversitesinde ise HBP ortalaması 90 üstü ve/veya standart sapması 8'in altında olan sınıflar için mutlak değerlendirme kullanıldığı görülmektedir. Aksaray üniversitesi bağıl değerlendirme sisteminde ise mutlak değerlendirme uygulaması HBP ortalamasına göre uygulanmamakta, Karadeniz Teknik üniversitesinin uyguladığı bağıl değerlendirme sisteminde ise HBP ortalaması 80 üstü sınıflar içinde mutlak değerlendirme kullanılmaktadır. Ancak bu üniversitenin uyguladığı sistemde 80-89 puanlar BA ve 90 üstü puanlar ise AA harf notuna dönüşmekte ve bu durum öğrencilerin harf notlarını artırmaktadır.

Genel olarak elde edilen sonuçlar düşünüldüğünde, bağıl değerlendirmenin mutlak değerlendirmeye göre harf notu olarak öğrenci lehine çalıştığını ortaya çıkmaktadır. Aynı sonuç yakın zamanda Sayın (2016) yılında yapılan çalışma ile desteklenmesine rağmen Atılgan ve diğerlerinin (2012) bulduğu sonuç ile farklılık göstermektedir. Bu durum Atılgan ve diğerlerinin (2012) çalışmasında kullandığı bağıl değerlendirme algoritması ile mevcut çalışmadaki kullanılan bağıl değerlendirme

algoritmalarından farklı ve daha az bir örneklem ile yapılmasından kaynaklanabileceği şeklinde açıklanabilir.

Bu çalışmanın sonuçlarından hareket ederek biri öğrenci açısından diğeri ise üniversitedeki öğretim elemanı açısından iki farklı yorum yapılabilir. Öğrenci açısından bakıldığında mutlak değerlendirmeye göre bağıl değerlendirme ile öğrenciler daha yüksek not ortalamasına sahip olurlar. Bu durum daha önceki çalışmalarda Başol (2015)' un ifade ettiği gibi bağıl değerlendirmenin not enflasyonuna sebep olacağı görüşünü desteklemektedir. Yine bu çalışmanın bulgularından hareketle farklı bağıl değerlendirme sistemleri aynı HBP'yi farklı harf notuna dönüştüğünü ortaya çıkarmakta, bu durum bağıl değerlendirme sistemi uygulayan bazı üniversitelerin diğer üniversitelere göre daha avantaj sağladığı düşünülebilir. Gerek mutlak ve bağıl değerlendirme not sistemi farklılığı, gerekse kullanılan bağıl değerlendirme sistemleri arasındaki farkın adaletsizliğe sebep olabilmektedir. Bu durumun somut göstergesi özellikle üniversite mezuniyet sonrasında öğrencilerin lisansüstü eğitime başvurularda ortaya çıkmaktadır. Bu başvurularda lisans mezuniyet ortalamasının etki derecesi bazı üniversitelerde %40'ı bulduğu düşünüldüğünde mezun olunan üniversitede kullanılan değerlendirme sistemlerinin adaletsizliğe ne derece sebep olduğu görülmektedir.

Çalışmanın sonuçlarına öğretim elemanları açısından bakıldığında bağıl değerlendirme sistemi öğretim elemanlarının hazırladıkları testlerin yapısından kaynaklanacak not hatalarını azaltabilecektir (Duman, 2011). Özellikle öğretim elemanlarının yeteri düzeyde test ya da soru hazırlama tekniğinden yoksun olabileceği düşünülürse, sorunun yanlış anlamlandırılması ya da mevcut programda olmaması gibi testin yapısından kaynaklanan hatalar her bir öğrencinin notunu düşürecektir. Ancak gerek önceki çalışmaların (Nartgün, 2007; Sayın, 2016) gerekse bu çalışmanın bulguları bağıl değerlendirmenin öğrenci notlarını artırdığı sonucunu desteklediğinden, bağıl değerlendirme özellikle öğretim elemanlarından kaynaklanan bu düşük notları daha yüksek harf notlarına dönüştürebilmesi muhtemeldir. Ancak bağıl değerlendirme ile bu düşük notların harf aralıkları öğrenciler lehine olacaktır. Burada dikkat edilmesi gereken nokta, bazı öğrencilerin hak etmedikleri halde dersten geçmeleri ya da yüksek harf aralığına düşmeleridir. Bu sebepten dolayı bağıl değerlendirmeye giren öğrencilerin yeteri düzeyde olması ve harf aralıklarının titizlikle belirlenmesi tercih edilmektedir.

Bunların yanı sıra herşeyden önce yapılan sınavın amacı, yapısı ve sonuçlarına bağlı olarak mutlak değerlendirme mi yoksa bağıl değerlendirme mi kullanılmasına karar vermek doğru bir ölçme ve değerlendirme açısından oldukça önemlidir. Turgut ve Baykul (2015) özellikle bağıl değerlendirme kullanılan dağılımın normal dağılım olması durumunda verilecek notların da simetrik olarak dağılacağını, aksi durumda öğrencilerin aldığı notların sınıftaki diğer kişilerden fazla etkileneceğini ifade etmektedir. Çünkü gerek sağa çarpık gerekse sola çarpık notlardaki uç değerleri hem dağılımın ortalamasını değiştirecek hem de standart sapmayı artıracaktır.

Yine ham puan dağılımlarının çarpık olması durumunda mutlak değerlendirme kullanılması ise farklı sonuçları doğuracaktır. Sola çarpık dağılımlarda sınıf not ortalaması yüksek olacağından sınıftaki birçok kişi yüksek harf notları ile geçerken, sağa çarpık dağılıma sahip bir sınıfta ise sınıftaki notlar düşük olacağından verilen harf notları da düşük olacak hatta birçok kişi dersten başarısız olabilecektir. Notların çarpık dağılım göstermesi sınav sorularının çok zor ya da çok kolay sorular sorulmasından kaynaklanacağı göz önüne alındığında öğretim elemanının ölçme ve değerlendirme adına nitelikli soruların hazırlanmadığı düşünülebilir. Thorndike (2005) nitelikli bir test hazırlarken soruların güçlük derecelerin dengeli olmasına yani soruların %25' inin zor, %50' sinin orta zorlukta ve %25' inin kolay olmasına dikkat çekmektedir. Böylelikle notlar dağılımı normale daha da yaklaşabilecektır.

Yücel (2015)'in ifade ettiği Türkiye'de bağıl değerlendirmenin kullandığı üniversite giriş sınavı gibi ulusal sınavlara bakıldığında bilişsel düzey bakımından alt düzey yani bilgi, kavrama ve uygulama düzeyindeki hedefleri ölçtüğünü söylenebilir. Özellikle sınıf içinde değerlendirmelerde öğretim elemanların hazırladıkları sınavları belirtke tablosu kullanması sınav sonucunda oluşacak not dağılımını etkileyebilecektir. Özellikle öğrencilerin bilişsel kazanımları ölçecek soruların tüm basamakları kapsayacak şekilde yazılması öğrencinin hangi hedefi ne derece öğrendiğini diğer bir ifade ile verilen bilgide ne kadar derinleştiğini ölçecektir. Bunun içinde en çok kullanılan soru tipi olan çoktan seçmeli yerine açık uçlu sorular ve proje tabanlı ödevler verilerek öğrenciye kazandırılması gereken hedefler üst basamaklara diğer bir ifade ile analiz, sentez ve değerlendirme basamağına

çıkaracaktır. Bu doğrultuda hazırlanan sınavlardaki soru sayısı fazla olması beklenip, bunun sonuncunda sınavın hem kapsam geçerliliğinin artması hem de soruların uygulanması sonucunda elde edilen notların da normal olarak dağılması beklenecektir. Daha açık bir ifade ile sınavın öğrencilerin bilişsel kazanımlardaki tüm basamakları (bilgi, kavrama, uygulama, analiz, sentez ve değerlendirme) kapsayacak şekildeki sorulardan oluşması soruların güçlük dereceleri etkileyecek ve Thorndike (2005)' in ifade ettiği gibi soruların güçlük derecesi bakımından dengeli bir şekilde dağılmasına sebep olacaktır. Kısacası bir sınavın ölçme ve değerlendirme çerçevesinde hazırlanması not dağılımını etkileyecek ve bunun sonucunda bağıl değerlendirmeye ihtiyaç duyulmayarak mutlak değerlendirme ile sınavlar değerlendirilebilecektir.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(1);30-48

Calculation of Effect Size in Single-Subject Experimental Studies: Examination of Non-Regression-Based Methods*

Nihal ŞEN** Sedat ŞEN ***

Abstract
It is observed that meta-analysis studies have not been included in single-subject experimental studies as much as the experimental studies. In order to overcome this deficiency in the literature, regression-based and non-regression-based indexes that can be used as the effect size in single subject experimental studies have been recently developed. Since most of the regression-based indexes are affected by the serial dependency of single-subject experimental data, non-regression-based indexes that were less affected by this dependency and were preferred more than regression-based indexes were the main subject of this study. Although there are many indexes that are not based on regression in single-subject experimental studies, it is observed that most of the researchers prefer the percentage of non-overlapping data (PND) and percentage of zero data (PZD). There are many controversies in the literature especially on the use of the PND index. In the absence of a study describing the alternative indexes of PND and PZD in Turkey literature, the examination of non-regression-based indexes makes this study important. The aim of this study is to examine the non-regression methods used to calculate the effect size in single-subject experimental studies and to show how these methods will be applied in single-subject experimental research. In this aim, how to prepare the data, how to analyze it, how to synthesize it for more than one study and how to interpret the results are discussed. In this study, suggestions were made for the researchers based on 10 different indexes.

Key Words: Single-subject experimental research, meta-analysis, effect size, non-regression-based indexes.

INTRODUCTION

In recent years, new trends in studies in the field of education show that more emphasis is given to the synthesis of data from studies in a specific field. Hattie (2009) stated that practitioners and policy makers should summarize and compare various types of evidence obtained through meta-analysis in order to close the gap between research and practice in education (Kavale, 2001). Meta-analysis (Glass, 1976) is a method that provides a statistical summary of the studies conducted in a specific field. Meta-analysis mainly helps to calculate the overall mean value of the subject matter using the effect size values obtained from quantitative studies on a specific subject.

The meta-analysis method, which provides many benefits for researchers, can be applied in traditional meta-analysis studies with the effect size values such as standardized mean difference, correlation and risk ratio (Lipsey & Wilson, 2001). It is of great importance for the researchers to determine the effectiveness of the intervention in experimental studies. Besides the statistical significance, calculating the effect size value, which indicates the practical importance, gives the researcher information that can be interpreted (such as small effect, large effect) about the effectiveness of the intervention phase. Indeed, the American Psychological Association strongly recommends that every published work should report an effect size value (American Psychological Association, 2010).

Although single-subject experimental research (Kırcaali-İftar & Tekin, 1997) has a long-standing history, meta-analysis practices to synthesize studies in this area have begun in the late 1980s. Most of

* This study was presented at the 27th International Congress on Educational Sciences (ICES/UEBK-2018) (April 18-22, 2018, Antalya).

** Ph.D. student, Bolu Abant Izzet Baysal University, Institute of Educational Sciences, Special Education, Bolu, Turkey, e-mail: nihalllsennn@gmail.com, ORCID ID: orcid.org/0000-0002-9511-8401
*** Asst. Prof. Dr., Harran University, Faculty of Education, Educational Sciences, Şanlıurfa, Turkey, e-mail: sedatsen@harran.edu.tr, ORCID ID: orcid.org/0000-0001-6962-4960

To cite this article:
Şen, N., & Şen, S. (2019). Calculation of effect size in single-subject experimental studies: Examination of non-regression-based methods. *Journal of Measurement and Evaluation in Education and Psychology, 10*(1), 30-48. DOI: 10.21031/epod.419625

Received: 30.04.2018
Accepted: 16.12.2018

the researchers in this area failed to develop an acceptable effect size due to the problems arising from the data structure in single-subject experimental studies. Among these problems, single-subject experimental study data included repeated measurements on the same individual, hence the dependence on residual values (Huitema, 1985), the low number of measurements and the non-normal distribution of these measurements. These problems prevented the applicability of these statistics by providing some assumptions of parametric statistics in single subject experimental studies. In order to solve these problems, many indexes have been suggested for meta-analyses that can be applied in single-subject experimental studies (Beretvas & Chung, 2008). Among others, values based on percentage of non-overlapping data (Scruggs, Mastropieri & Casto, 1987) and the standardized average difference (Busk and Serlin, 1992) of the experimental and control groups (i.e., the baseline and intervention levels in single-subject experimental studies) are the major ones. In addition, a large number of indexes were proposed in the literature, mainly based on percentage calculations (Ma, 2006; Parker & Vannest, 2009; Parker, Vannest, & Brown, 2009; Parker, Vannest, & Davis, 2011). In addition to these nonparametric methods, many statistical methods based on regression analysis have been developed (Allison & Gorman, 1993; Huitema & McKean, 2000; Swaminathan, Rogers, Horner, Sugai, & Smolkowski, 2014; van den Noortgate & Onghena, 2003).

The methods proposed in the literature are suitable for different data designs to be obtained from experimental studies with single-subjects. Especially, the methods not based on regression have been widely accepted and still used in single-subject experimental studies (Aslan, Yalçın, & Özdemir, 2016; Aydın, 2017; Karasu, 2009a, 2009b, 2011; Korkmaz & Diken, 2010; Sönmez & Diken, 2010; Tavil & Karasu, 2013). It is also observed that the PND and PZD indexes are widely used by researchers in Turkey. Especially in order to find a solution to the criticisms of PND, many alternative effect size indexes which are not based on regression have been proposed in the literature. However, it is observed that other non-regression-based alternatives are rarely being used by researchers in Turkey (Bozkus-Genç, 2017; Kaya, 2015; Uysal, 2017). One of the reasons for little use of these methods by the researchers may be due to lack of methodological studies describing these indexes in Turkey. Many researchers abroad conducted studies describing and comparing the effect size values that can be used for single-subject experimental studies (Alresheed, Holt, & Bano, 2013; Campbell, 2004; Heyvaert, Saenen, Campbell, Maes, & Onghena, 2014; Maggin, O'Keeffe, & Johnson, 2011; Maggin, Swaminathan et al., 2011; Manolov & Solanas, 2008; Manolov, Solanas, & Leiva, 2010; Olive & Franco, 2008; Parker et al., 2011; Wolery, Busick, Reichow, & Barton, 2010). Most studies in the literature are based on the comparison of non-regression-based indexes. Parker et al. (2011) compared nine indexes, but others generally compared a few indexes. The explanation and comparison of 10 indexes used in single-subject experimental research makes this study different from other studies in terms of the number of indexes discussed in a single study. According to the extensive literature review, the number of studies comparing or explaining these indexes in Turkey is almost negligible. The only example that can be given in this sense is Karasu (2009a)'s study of the effects of a group study selected from the treatment studies based on natural approaches to improve communication and social skills of children with autism. In that study, four different methods, percentage of non-overlapping data, percentage of zero data, Swanson model and ITSACORR (interrupted time-series analysis procedure), were compared to determine the most useful methods for calculating the effect-size in single-subject experimental studies. Correlation analyses was performed to determine the relationships between the effect size values obtained from four different methods and mean values were compared among the results of these methods.

The lack of a comprehensive study describing the indexes developed in the Turkish literature after PND and PZD makes this study important in which non-regression indexes are examined. The purpose of this study is to examine the non-regression methods of the meta-analysis of single-subject experimental studies and to show how these methods will be applied in single-subject experimental studies on sample data. In this purpose, how to prepare the data, how to analyze it, how to synthesize it for more than one study and how to interpret the results will be discussed. In this study, it was attempted to answer how the results obtained from sample data which are suitable for single-subject experimental structure differ between these indexes. All data used in this study are generated. Sample analysis was performed on the same data for all methods except PZD as PZD index was used to determine the effectiveness in

diminishing a behavior. Therefore, this can be considered a comparison study. The methods examined in this study do not include all indexes proposed in the literature. The most studied indexes were determined and included in this study by examining the comparison studies in single-subject experimental research literature (Alresheed et al., 2013; Campbell, 2004; Heyvaert et al., 2014; Maggin, O'Keeffe, & Johnson, 2011; Magin, Swaminathan et al., 2011; Manolov & Solanas, 2008; Manolov et al., 2010; Olive & Franco, 2008; Parker et al., 2011; Wolery et al., 2010).

Percentage of Non-overlapping Data (PND)

The percentage of non-overlapping data (PND) proposed by Scruggs et al. (1987) is a nonparametric index based on comparison of baseline and intervention levels. It is called percentage of non-overlapping data as it is based on taking into account the non-overlapping data between the baseline and intervention levels.

The calculation of this index is made by determining the ratio of the intervention level values that exceed the maximum value of the baseline level to the number of values obtained at the intervention level. The following steps are followed when calculating the PND value for an AB design:

1. Determine the maximum baseline value in the graph,
2. Draw a horizontal line from this determined maximum value to the right (intervention level),
3. Determine the intervention level values above this horizontal line,
4. Divide the number of data points obtained in the third step by the total number of data points at the intervention level,
5. The value obtained in the fourth step is multiplied by 100 to calculate the PND value.

In the intervention level situations where the target behavior is expected to increase, the intervention level values that exceed the highest value of the baseline level are taken into account when calculating the value of the PND. In cases where the target behavior is expected to decrease, the intervention level values below the lowest level of the baseline level are determined. The number of values obtained in these two cases is divided by the total number of data points at the intervention level. If a study involves more than one intervention, the PND indexes obtained from different interventions are combined to determine the mean value and this value is used to evaluate the overall effectiveness of all interventions. The PND value can also be calculated for the ABAB design, which is frequently applied in experimental studies with single subjects. In the ABAB design, the PND indexes are calculated for both of the AB designs. The mean of these two PND percentages is then calculated to find the overall value of the PND of the ABAB design.

The values of the baseline level (A) and the intervention level (B) of an AB design obtained from an individual's dependent variable (minimum score = 0, maximum score = 20) are presented in Figure 1. Ten measurements were assumed to be collected at both levels. This generated data set was assumed to represent an expected increase in the targeted behavior. When we look at the baseline level measurements in Figure 1, the highest baseline level value is 8. When a horizontal line is drawn over this value to the right, values above 8 are determined at the intervention level. According to the data in Figure 1, it can be seen that six data points (10, 9, 11, 10, 10, and 12) at the intervention level are above this line. If the number of data points exceeding the maximum value is divided by the total number of data points at the intervention level (10), a value of 0.6 is obtained. If we multiply this value by 100, the PND value is calculated as 60 (i.e., 0.6 × 100) based on the data presented in Figure 1.

According to Scruggs and Mastropieri (1998), the PND values should be over 90 in order to say that the intervention is "very effective". The values between 71 and 90 indicate "effective" intervention, while values between 50 and 70 indicate "questionable" or "moderate" effects. It indicates that the intervention "ineffective" where the percentage value is less than 50 (Scruggs, Mastropieri, Cook, & Escobar, 1986; Strain, Kohler, & Gresham, 1998). According to these criteria, the intervention in Figure 1 can be considered moderately effective.

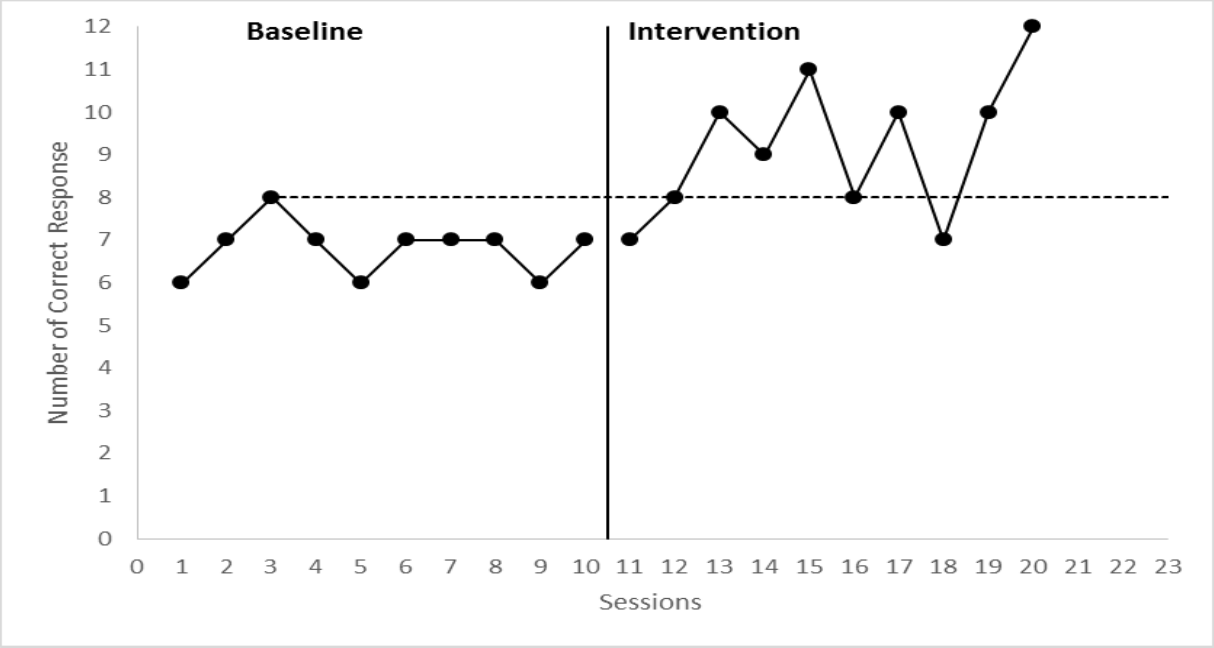


Figure 1. Calculation of Percentage of Non-overlapping Data in an AB Design.

Although PND index is not necessary to meet the assumptions of parametric statistics, it is often preferred by researchers as it is easy to compute and interpret (Ma, 2006). In addition to these advantages, the PND index has several limitations that have been criticized by researchers (Allison & Gorman, 1994; Strain et al., 1998). One of these criticisms is the failure of PND to behave as an effect size value. On the other hand, there are also explanations made by Scruggs and Mastropieri (2013) regarding the relationship between the size of the PND and the effect size. Another criticism of this value is that the percentage of non-overlapping data ignores all other values except for a single value in the baseline level. Another criticism directed at PND is that it does not take into account the trend stem from linear increase or decrease at the baseline level and does not detect the changes in slope. In the light of these criticisms, it should be noted that a reliable result cannot be obtained from the PND index alone (Scruggs & Mastropieri, 1998). This is because the PND index does not have a known sampling distribution and a *p* value cannot be calculated for this index (Parker, Hagan-Burke, & Vannest, 2007). Therefore, the fact that a *p* value cannot be calculated eliminates the possibility of making inferences based on this index. It is also criticized that the PND index value does not accurately reflect the effect of intervention level if 0 or 100 values are present at the baseline level (Ma, 2006). Scruggs and Mastropieri (1998) do not recommend calculating the PND in the case of a floor (0) or a ceiling (100) is present at among the baseline data points. Finally, as in other non-overlapping indexes, the increase in the value of the PND value as the number of intervention levels increases is seen as a limitation (Allison & Gorman, 1993).

Percentage of Zero Data (PZD)

Percentage of zero data (PZD) is an index developed by Scotti, Evans, Meyer, and Walker (1991) to show the effectiveness of the intervention level in single-subject experimental studies. PZD represents the degree of behavior suppression versus degree of behavior reduction and it is seen as a more stringent efficacy indicator (Campbell, 2004, p. 235). In other words, it is a measure that requires the targeted behavior to reach zero and remain at zero level.

The following steps are followed to obtain the PZD value in an AB design:

1. The first data point with zero at the intervention level is detected,
2. The number of intervention level data remaining at the zero point, including the zero point detected in the first step, is determined,

3. The number of zero values specified in the second step is divided by the number of data points after the first zero at the intervention level,
4. The value obtained in the third step is multiplied by 100 to obtain the PZD value.

The values of the baseline level (A) and the intervention level (B) of an AB design obtained from an individual's dependent variable (minimum score = 0, maximum score = 20) are presented in Figure 2. Ten measurements were assumed to be collected at both levels. This generated data set was assumed to represent an expected decrease in the target behavior.

When the intervention-level measurements are examined, it is the seventh data point where the participant gets first zero point. After this point, three more measurements were made on the participant and this participant scored two more zero points. Thus, the number of data points after this point, including the first zero, was four and the participant scored zero points in three of these four measurements. In order to calculate the PZD value, it is sufficient to divide the number of zero points by four. Thus, the PZD value is $3/4 \times 100 = 75$. In cases where there are multiple intervention levels (e.g., ABCD), some researchers (Reichle, 2007) have calculated the PZD value using only the last intervention level (e.g., D). In the multiple baseline designs across subjects, the PZD value is calculated for each subject separately (Schlosser & Koul, 2015). In their work on ABAB designs, Wehmeyer et al. (2006) stated that they calculated a PZD value for each pair of AB design and each calculated PZD was considered separate values.

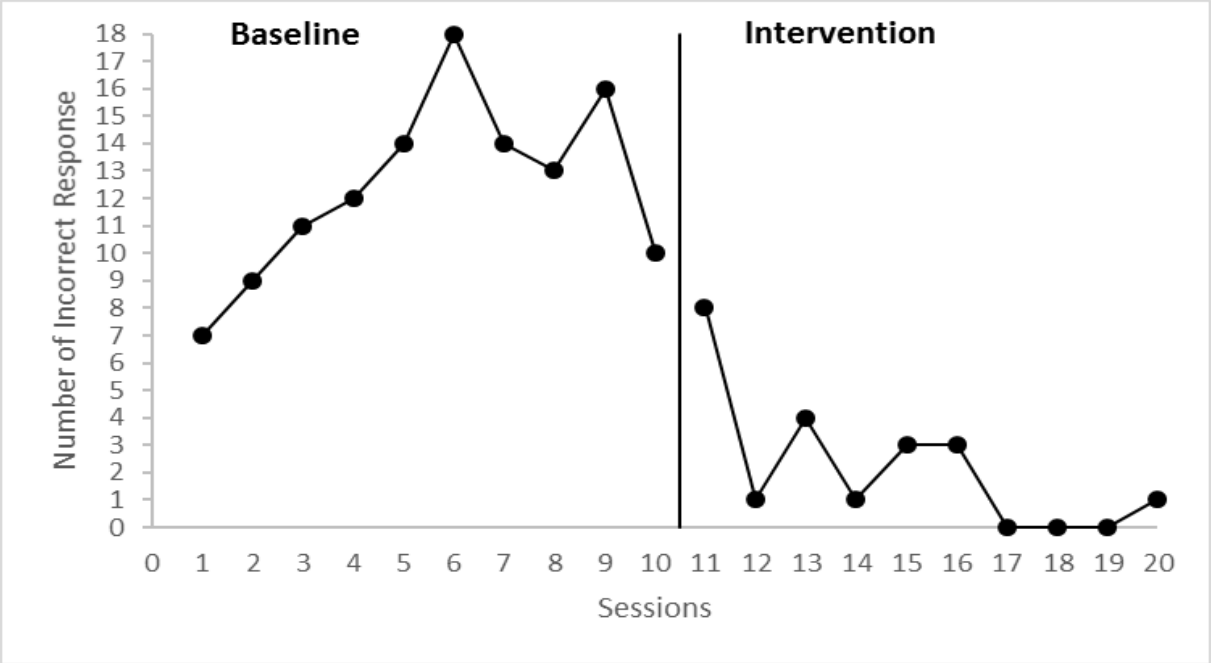


Figure 2. Calculation of Percentage of Zero Data in an AB Design.

The PZD index takes values from 0 to 100 and higher values indicate an effective intervention. The criteria suggested by Scotti et al. (1991) can be used in the interpretation of PZD. According to Scotti et al. (1991), values between 55-80% indicate "moderate" effect, while values above 80% indicate "high" effect. While PZD values below 18 percent imply "ineffective" intervention, values between 18-54% indicate "questionable" effect. According to the above criteria, the intervention in Figure 2 can be said to have "moderate" effect.

Since all data before the first zero point in the PZD index are ignored, data loss occurs in some cases when compared to PND. At the same time, non-zero values at the intervention level are not included in

the calculation of the PZD value. Besides, as in PND, the PZD index can also be affected easily by the trend arising from outliers and time (Allison & Gorman, 1993). Another disadvantage of the PZD index is that it only takes into account the data points at the intervention level. As it focuses only on the elimination of the target behavior, it is not suitable for every intervention situation.

Percentage of Data Points Exceeding the Median of Baseline Level (PEM)

As indicated by Ma (2006, p. 600), the calculation of PND index using the horizontal line in the presence of the 0 or 100 (ceiling and floor effect) value in the baseline level data makes it meaningless or zero. This is considered one of the weaknesses of this method. There may be a trend effect resulting from non-sudden decrease and increase in levels. This trend effect is ignored in the PND. For these reasons Ma (2006) argued that the risk of making Type II error was too high and, as an alternative to this method, the percentage of data points exceeding the median value of the baseline level (PEM) method was proposed.

When calculating the PEM value in an AB design, the following steps are followed:

1. The median value of the baseline level on the graph is determined,
2. A horizontal line is drawn from the median to the right.
3. The intervention level data points remaining above this horizontal line is determined,
4. The number of data points in the third step is divided by the total number of data points at the intervention level,
5. The value obtained in the fourth step is multiplied by 100 to calculate the PEM value.

If the intervention applied while calculating PEM value is expected to increase the target behavior, the intervention level data points above the median value of the baseline level are taken into account while the effect of the intervention in the intervention is expected to decrease the target behavior the intervention level data points below the median value are taken into account.

When we look at the baseline level data points in Figure 3, the median value of this level is 7. When a line is drawn over this value to the right side of the graph, values above 7 at the intervention level are determined. According to the data in Figure 3, eight data points (8, 10, 9, 11, 8, 10, 10, and 12) are seen above this line. At the intervention level, the value of the data exceeding the median value is divided by the total number of data points at the intervention level (10) to obtain a value of 0.8. If we multiply this value by 100, the PEM value in the data in Figure 3 is calculated as $0.8 \times 100 = 80$.

When the recommended criteria for the PEM index are considered, the values between 91% and 100% indicate a “very effective” intervention, while between 70% and 90% indicate “moderate” intervention effect. In studies with a PEM value below 70%, it can be said that the intervention effect is “questionable” or “ineffective” (Ma, 2006). According to Heyvaert et al. (2014), “questionable” intervention is between 50% and 70% and “ineffective” intervention occurs below 50%. In case of ineffective experiments, the data points show more or less continuous fluctuations around the median value. The PEM value obtained using the data in Figure 3 indicates a moderate effect according to the above criteria.

The PEM method has been found to be less affected by the autocorrelation in the data, and it has been found that it effectively differentiates the effective and ineffective interventions comparing to other indexes (Manolov, Solanas, & Leiva, 2010). The PEM index is used to calculate the change in the target behavior among AB levels, not the trend changes between the baseline and the intervention levels. In addition, this index does not take into account of the change and trend at the intervention level. The PEM value provides a partial solution to the inability to calculate the slope value in the PND when there is an orthogonal slope in the baseline- and intervention-level pairs after the first intervention level. According to Ma (2006), one of the limitations of PEM index is to ignore the magnitudes of the data points above the median in the calculation of this index, which means that this index is not sensitive to the data points above the median. In addition, how to calculate PEM in data cases other than AB designs was not mentioned in the original study.

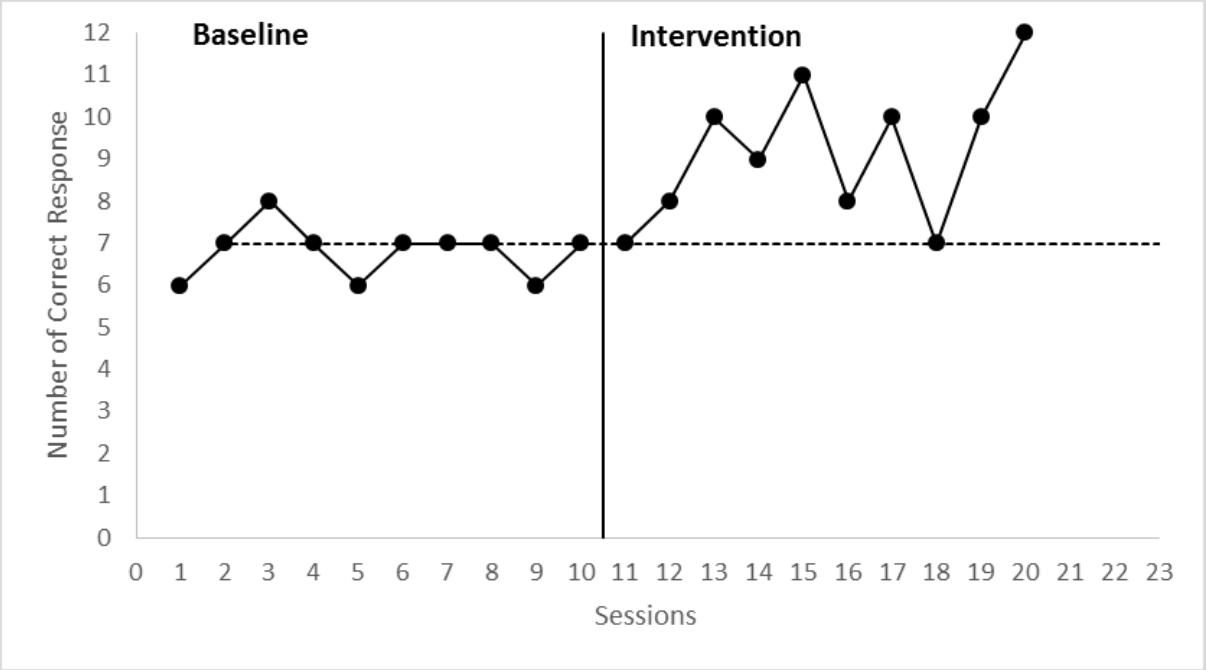


Figure 3. Calculation of Data Points Exceeding Median of Baseline Level in an AB Design.

Percentage of Data Exceeding a Median Trend of the Baseline Level (PEM-T)

Percentage of data exceeding the median trend of the baseline level (PEM-T) is an effect size index that takes into account the trend determined using the split-middle line technique (White & Haring, 1980) for single-subject experimental research. (Wolery et al., 2010). The PEM-T index, a version of PEM, is a non-parametric statistics used for the same purpose.

In the calculation of the PEM-T value in an AB design, the following steps are followed:

1. A split-middle line of trend is drawn on the graph from the baseline to the intervention level using the split-middle line technique as described in White and Haring (1980),
2. The intervention level data points above the median trend line are counted,
3. The number of data points in the second step is divided by the total number of data points at the intervention level,
4. The value obtained in the third step is multiplied by 100 to calculate the PEM-T value.

If we want to calculate the PEM-T value according to the baseline and intervention level data points presented in Figure 4, we need to create a median trend line using the data at the baseline level. To calculate PEM-T value in Figure 4, a median trend line was obtained using the R syntax created by Manolov, Sierra, Solanas and Botella (2014). As can be seen in Figure 4, eight data points at the intervention level remain above this trend line. Thus, we can calculate PEM-T as $8/10 \times 100 = 80$. In an experimental study with multiple AB designs, the PEM-T values are calculated for each AB design and the values obtained are averaged. The researchers who developed PEM-T did not make any suggestions on how to calculate this index for more complex designs.

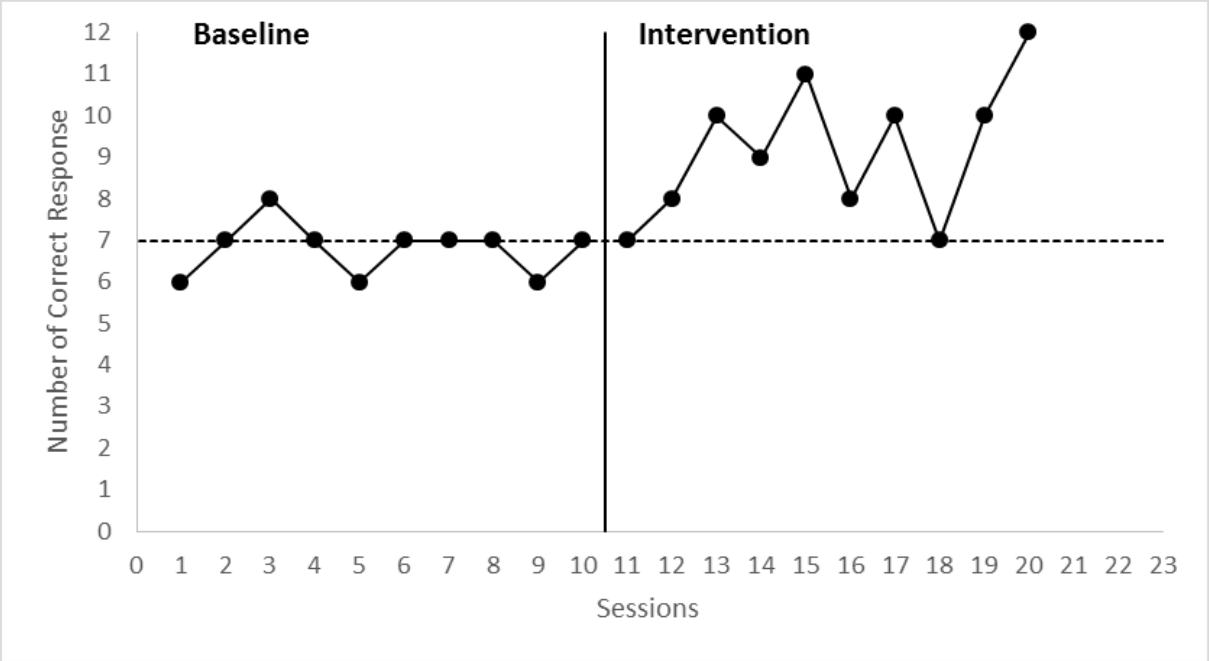


Figure 4. Calculation of the Percentage of Data Exceeding the Median Trend of the Baseline Level in an AB Design.

There are no clear criteria for interpreting PEM-T values in the literature. Some criteria presented in Ma (2006) can be used. According to these criteria, values of 90% and above indicate a "very effective" intervention and values between 70% and less than 90% indicate "moderate" intervention effect. Values between 50% and less than 70% refer to "questionable" or "mild" intervention. If PEM-T is below 50%, the intervention is considered "ineffective". These criteria need to be used more cautiously, as the effect size value may be smaller in graphs with a trend that begins at the baseline level. The PEM-T index also has some advantages and limitations. One of the main advantages of this index is that it is easy to interpret and can take account of the linear trend at the baseline level. The fact that the median trend line is affected by outlier values at the baseline level is also seen as a factor affecting the calculation of this index.

Percentage of All Non-Overlapping Data (PAND) and Phi Coefficient

The percentage of all non-overlapping data suggested by Parker et al., (2007) helps to find the percentage of non-overlapping data as in the other percentage of non-overlapping data indexes. The PAND value is defined as the percentage of remaining data points to total data points after eliminating the minimum number of data points that can remove the overlap between two levels (Parker et al., 2011).

In order to calculate the PAND value in an AB design, the following steps are followed:

1. Determine the data points that cause the overlap between the two levels,
2. Determine the minimum number of data points to eliminate the overlap between two levels,
3. Count the overlapping data points,
4. Calculate the percentage of overlapping data by dividing the number of overlapping data points determined in the third step by the total number of data points,
5. The percentage of overlapping data obtained in step 4 is subtracted from 100 to obtain the percentage of non-overlapping data (PAND).

Figure 5 shows the overlapping region and the data points that cause this overlapping area to calculate the PAND value according to the baseline and intervention level data. As shown in Figure 5, a line is drawn from the highest point of the baseline level and a line is drawn from the lowest point of the

intervention level and the overlapping region is determined between the two levels. A minimum number of data points are then decided to eliminate this overlap. If we remove a data point (8) at the baseline level and two data points at the intervention level (7 and 7) in the graph in Figure 5, we eliminate the overlap between these two levels and we get the number of all data points that do not overlap. Then, if we remove the number of data points (2 + 1 = 3) that overlap from the entire number of data points (20), we get the number of all the non-overlapping data points (17). Thus, we find the value of PAND (20-3)/20=0.85. If we multiply this value by 100 to represent it as a percentage, we get the percentage of all non-overlapping data.

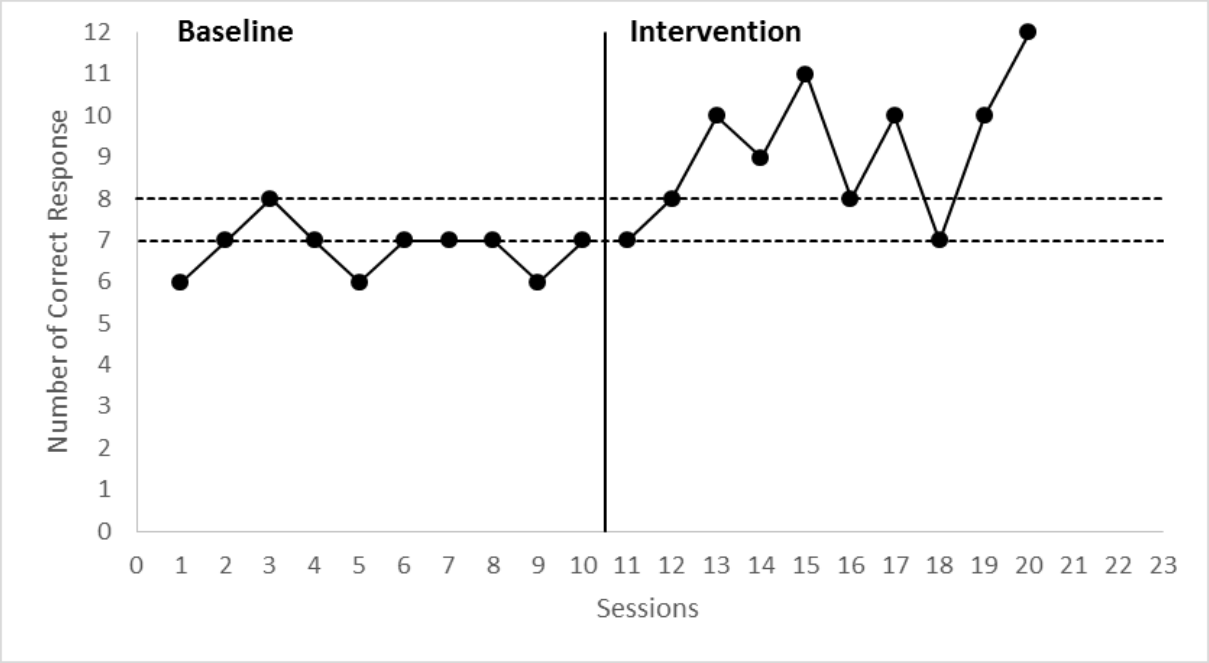


Figure 5. Calculation of the Percentage of All Non-Overlapping Data in an AB Design.

Parker et al. (2007) state that the PAND can be calculated only by looking at the graph, but it is difficult to find the correct results by inspecting the graph in cases where too many data points are present (complex single-subject experimental designs). PAND index, which is based on non-overlapping data values such as PND, is calculated using all data at the baseline level unlike the PND. Another advantage of the PAND index is that this index can be transformed into Pearson Phi correlation. Parker et al. (2007) state that the Phi value has a known sampling distribution, and the value of *p* for Phi can be obtained. It also allows for confidence intervals and power calculations. PAND index can be interpreted by converting it to Cohen's *d* value through the Phi value (Parker et al., 2007). Thanks to these features, the PAND value provides the opportunity to obtain a lot of information that cannot be obtained with the PND index. PAND, which can only measure average level changes, cannot control the positive baseline level trend (Parker et al., 2007, p. 196).

Nonoverlap of All Pairs (NAP)

The nonoverlap of all pairs (NAP) index developed by Parker and Vannest (2009) is a value calculated based on completely pair comparisons.

In an AB design, the following steps are followed to calculate the NAP value:

1. Each data point at the baseline level is compared with each data point at the intervention level. When we compare one of the baseline-level data points in an AB design to all of the intervention-level data points, there are three cases for each pair of data. The first of these cases

is the development of the data point taken at the baseline level towards the point of comparison at the intervention level (positive value). In other words, the data point at the intervention level is larger than that of baseline-level. Second, the level of baseline level data is no change (equal value or ties) from this level to the intervention level. Third, at the baseline level, the comparison of the data point at which we make the comparison is greater than the data points at the intervention level, i.e., it shows a decrease (negative).

2. A total score is calculated for each baseline level data point to be compared. When calculating this score, a value of 1 for positive ($A_n<B_n$) cases, a value of 0 for negative ($A_n> B_n$) and a value of 0.5 for cases with equal status ($A_n= B_n$) are given.
3. These values obtained in the second step are summed and the total development score is obtained for the baseline point subject to comparison.
4. We do the same for other data points at the baseline level, and we calculate the total development points.
5. By dividing this total development score obtained in the previous steps by the number of all data pairs that may be available (A × B), we obtain the NAP value.

The number of data pairs in an AB design is equal to product (A × B) of the number of data points at the baseline and intervention levels. The number of data pairs for the data in Figure 6 is 100 (i.e., 10 × 10). The data pairs that show increase (Pos), decline (Neg), and ties (Ties) are counted in Figure 6. Based on the data presented in Figure 6, development score is obtained as 10 (10 positive status) making the comparison of the first data point (6) at the baseline with each intervention level data point (6-7, 6-8, 6-10, 6-9, 6-11, 6-8, 6-10, 6-7, 6-10, 6-12). Similarly, for the second data point (i.e., 7), the development score is obtained as 9 (8 positive and 2 equal conditions). In the same way, the third data point (i.e., 8) of the baseline level is compared with each data point at the intervention level and shown in Figure 6 as an example. The development score for the third data point of the baseline is obtained as 7 (6 positive, 2 equal cases, 3 negative conditions) according to the comparison conditions in Figure 6. For each data point, the development scores are calculated as 10, 9, 7, 9, 10, 9, 9, 9, 10, and 9 for baseline sessions 1 to 10, respectively. When these values are summed, the total number of development data pairs equals to 91 and by dividing this number by the total number of data pairs (100), we calculate the NAP value as 91 (91%).

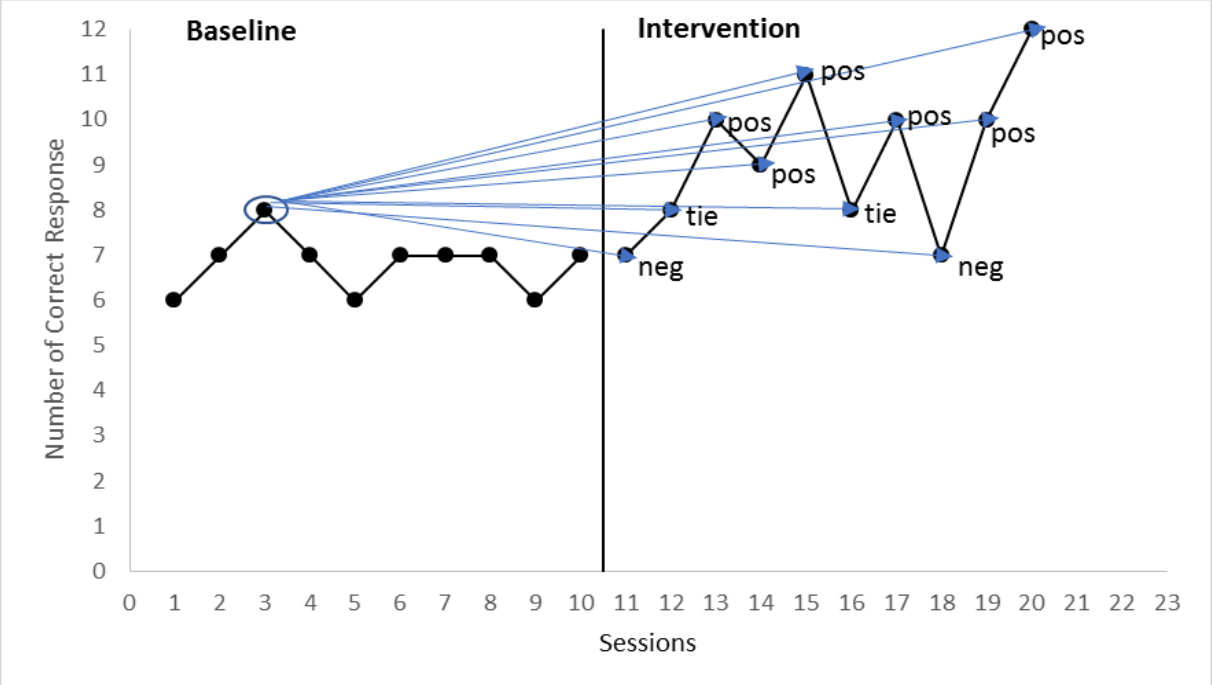


Figure 6. Showing the Data Pair Comparison Used for NAP Calculations for the 3rd Data Point (neg = negative; pos = positive; tie=equal).

The NAP index attempts to summarize the non-overlap between each baseline-level value and each intervention-level value. Briefly, it can be defined as the percentage of data showing development between A and B levels (Parker et al., 2011). Since the NAP takes into account all data pairs, it is a holistic non-overlapping data statistics and can be converted to the effect size Cohen's *d*. Like other non-overlapping data indexes, the NAP index can also be calculated using a graph, but it is not easy to compare all data pairs in large data sets. In addition to manual calculation from the graph, the NAP value can be obtained using receiver operating characteristics (ROC) or Mann Whitney U test analysis using familiar statistical programs such as SPSS. In the ROC analysis, Area Under the Curve (AUC) can be used to calculate the NAP index. If there are too many data points, or if there are multiple baseline and intervention levels, it may be difficult to calculate other non-overlapping data percentages from the graph. Even in these cases, the NAP value can be easily obtained by finding the area under the curve in the ROC analysis. In addition, the NAP value can be calculated by entering the data on a sheet via *http://www.singlecaseresearch.org/calculators/nap*. As mentioned in Parker and Vannest (2009, p. 359), the value under the curve (AUC value) is a probability value ranging from 0.5 to 1. According to the criteria given by Parker and Vannest (2009), the values less than 0.65 indicate "weak effect", the values between 0.66 and 0.92 indicate "moderate effect" and values above 0.92 indicate "very effective" intervention.

According to Parker and Vannest (2009), the NAP has five advantages compared to other non-overlapping data indexes. One of the main reasons for this is that they can better distinguish the results in the comparison of many published studies. The second advantage is that there is less error compared to other indexes which are calculated manually due to the fact that the scoring calculation can be performed with the help of commonly used computer programs such as SPSS. The third and fourth advantages are the high correlations between the data and the validity of the visual analysis and the conversion of the NAP to R^2 and thus to Cohen's *d* (Parker & Vannest, 2009). The fifth advantage given by Parker and Vannest (2009) is that the NAP value, which has low confidence intervals, offers more accurate results. Furthermore, in the case of autocorrelation in the longitudinal data, the NAP performs well (Manolov, Solanas, Sierra, & Evans, 2011).

Tau

Another effect size index developed by Parker, Vannest, Davis and Sauber (2011) is Tau (Kendall's Tau non-overlap). In the calculation of the tau index, non-overlapping data pairs are used as in the calculation of the NAP value. While the percentage of data pairs that do not overlap is found in the NAP, the percentage of overlap is excluded from the non-overlap percentage in the Tau index (Parker, Vannest, Davis, & Sauber, 2011). Given that PDP represents the number of data pairs that increases from the baseline level to the intervention level (positive) and the NDP represents the number of data pairs that are decreasing from the baseline level to the intervention level (negative) and TDP represents the total number of data pairs between the baseline and intervention levels. Tau index can be calculated as follows:

$$Tau = \frac{\text{PDP}-\text{NDP}}{\text{TDP}} \quad (1)$$

The number of all data pairs in an AB design, as in the PAND, is equal to the product number (A × B) of the baseline and intervention levels. Here, when we subtract the number of negative data pairs from the number of positive data pairs (the number of overlapping data pairs) and then dividing by the number of all data pairs, we get the Tau value.

In the data presented in Figure 6, by comparing each of the data pairs (100 data pairs) in a similar way to the calculations made in previous index, the number of positive pairs is 84, the number of negative pairs is 2 and the number of equal pairs is 14. Thus, the Tau value is calculated as (84-2) / 100 = 0.82. In order to convert this decimal value to a percentage scale, it is required to multiply by 100. The Tau value can also vary from 50% to 100%, as in the NAP index. As Parker et al. (2011) pointed out, the Tau value can be obtained using Kendall rank correlation or Mann-Whitney U test analysis using

conventional statistical programs (Parker, Vannest, Davis, & Sauber, 2011). The Tau value can be obtained by dividing the S value obtained in the Kendall rank correlation by the total number of data pairs (A × B).

In an AB design, the following steps are followed to obtain the Tau value from the Kendall rank correlation:

1. The Level variable is created: A Level variable is created by entering a value of 0 at the baseline level and a value of 1 at the intervention level (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1),
2. A Kendall rank correlation is calculated between the Level variable created in the first step and the raw data of baseline and intervention levels (6, 7, 8, 7, 6, 7, 7, 7, 6, 7, 7, 8, 10, 9, 11, 8, 10, 7, 10, 12),
3. S value is obtained from Kendall rank correlation (S = 82),
4. The S value in the third step is divided by the total number of data pairs (100) and the Tau value is obtained as .82. This value can be calculated by entering the raw data into a sheet via *http://www.singlecaseresearch.org/calculators/tau-u*.

Tau-U

Parker, Vannest, Davis, and Sauber (2011) proposed another Tau value (i.e., Tau-U) that can able to control undesirable positive baseline trend. The Tau-U index, as in Tau, does not only cover the non-overlapping data, but can also control the trend of the baseline level. In this way, the Tau-U index allows for the quantification of the increase in the intervention that occur only at the intervention stage, beyond the potential increase at baseline before the intervention (i.e., during the baseline level).

In an AB design, the following steps are followed to obtain the Tau-U value from the Kendall rank correlation:

1. A Level variable is created: Rank numbers are assigned to the data at the baseline level with the maximum value of 1 (3, 2, 1, 2, 3, 2, 2, 2, 3, 2) for (6, 7, 8, 7, 6, 7, 7, 7, 6, 7) and generated data are entered as Level A data. The last number in Level A is assigned to each element in Level B (7, 8, 10, 9, 11, 8, 10, 7, 10, 12) to continue (4, 4, 4, 4, 4, 4, 4, 4, 4, 4). A common Level variable is created by combining the data of these two levels (3, 2, 1, 2, 3, 2, 2, 2, 3, 2, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4),
2. A Kendall rank correlation is calculated between the Level variable created in the first step and the raw data of baseline and intervention levels (6, 7, 8, 7, 6, 7, 7, 7, 6, 7, 7, 8, 10, 9, 11, 8, 10, 7, 10, 12),
3. S value is obtained from Kendall rank correlation (S = 83),
4. The S value in the third step is divided by the total number of data pairs (100) and the Tau-U value is obtained as .83. We multiply by 100 to express this value on a scale of 100. This value can be calculated by entering the raw data into a sheet via *http://www.singlecaseresearch.org/calculators/tau-u*.

Parker et al. (2011) showed how the Tau-U value can be calculated for the AB and ABA designs. In this study, it was mentioned that only the first baseline (A) and intervention (B) levels are included in the calculation in more complex designs. Rakap (2015, p. 26) provided some explanations on how to use the Tau-U index, which is one of the most recent indexes developed in the literature, in other single-subject experimental designs. In designs with multiple baseline levels, Tau-U is calculated separately for each baseline-intervention (AB) comparison and the overall Tau-U value is obtained by taking the average of all Tau-U values obtained. When using the ABAB design, the Tau-U value is calculated for each baseline and intervention pairs (i.e., A_1B_1 and A_2B_2). The overall Tau-U value is also obtained by averaging the Tau-U values obtained from these pairs. In the ABCD design, according to Rakap (2015), Tau-U values should be obtained by comparing the intervention level with the baseline level for each level of intervention (e.g., AB, AC, AD given that A represents baseline level). The overall effect size of the full model should be calculated using the last intervention stage (AD comparison). In the case of

such design, the overall Tau-U value is the value obtained from the comparison of the baseline level and the final intervention level.

The criteria for the NAP suggested in Parker and Vannest (2009, p.364) can also be used for Tau-U. According to these criteria, percentages below 65 indicate "weak" or "small" effects, percentages between 66-92 indicate "moderate" effects, and percentages between 93 and 100 indicate "very effective" interventions. Since these are not specifically developed for Tau-U, caution should be noted in the use of these criteria. The Tau-U value of 82% using the data in Figure 6 indicates a moderate intervention.

The Tau-U has the flexibility to analyze trend and overlap separately or both at the same time. In addition to this flexibility, Tau-U index has more statistical power than other non-overlap indexes. The distinctiveness of the Tau-U index and its consistency with the visual analyses are among the strengths of this index. There is a decrease in the Tau-U index due to the attempt to control the trend. This decline is seen as a limitation of this index by Parker et al. (2011). The fact that the Tau-U index cannot be calculated in more complex single-subject experimental design than the AB and ABA designs is also seen as a limitation.

It should be noted that all of the non-overlap methods described so far are adapted from methods known as nonparametric dominance statistics in experimental studies. The dominance statistics can be defined as the probability of a randomly selected score from a group exceeding a score in a second group (Parker, Vannest, & Davis, 2011). The use of non-overlapping statistics, which are the equivalents of the dominance statistics in single-subject experimental studies has given more reliability to the publications in this area.

Mean Baseline Reduction (MBR)

Mean Baseline Reduction (MBR; Campbell, 2003) was developed by O'Brien and Repp (1990) to calculate the decrease in the level of intervention in the targeted behavior. The MBR index is also known as the percentage of decline from the baseline level and is applied to determine how much the behavior has decreased.

In order to calculate the MBR value, it is sufficient to calculate the means of the data points at the intervention level and baseline levels. The following steps can be followed when calculating MBR in an AB design:

1. Calculate the mean of the baseline level,
2. Calculate the mean of intervention level,
3. Subtract he mean of the intervention level from the mean of the baseline level,
4. Divide the value in the third step by the mean of the baseline level (Campbell, 2003),
5. By multiplying the value obtained in the fourth step by 100, the percentage of MBR is obtained. To put it another way; the MBR value is calculated with the following equation:

$$\text{MBR} = \frac{M_A - M_B}{M_A} \quad (2)$$

where M_A represents the mean of baseline level and M_B represents the mean of intervention level.

If the purpose of the intervention level is to increase rather than decrease a behavior, then the MBR value is calculated by changing the order of two terms (M_A and M_B) on the numerator part of Equation 2. In other words, the baseline level mean is subtracted from the intervention level mean and the resulting number should be divided by the mean of the baseline level. In order to be able to calculate the MBR over the data presented in Figure 6, we must first calculate the mean value for both levels. The mean values of the baseline level according to the data in Figure 6 is 6.8 and the mean value of the intervention level is calculated as 9.2. Thus, the percentage of MBR is calculated as ((9.2 - 6.8) /6.8) × 100 = 35.3.

When the MBR value was found to be 100%, the problematic behavior was completely eliminated and the 0% MBR value indicated that no change was observed according to the baseline level. Negative

MBR value shows that problematic behavior increases at intervention level. Some researchers recommend the use of the last three (Olive & Franco, 2008) or the last five (Lydon, Healy, O'Reilly, & McCoy, 2013) data points at the baseline level and intervention level when calculating the MBR value. If the values at the baseline level are zero, it is not possible to calculate the MBR value. This is a limitation of MBR. Although there are no clearly defined interpretation criteria for MBR, Bell, Skinner and Fisher (2009, p. 5) used small (.20), medium (.50) and high (.80) criteria for MBR values. In the literature, there are not many studies on how to calculate the MBR index in different single-subject experimental designs. Carr, Severtson, and Lepper (2009) used the last baseline and intervention-level data points to calculate the MBR for ABA designs. We also found Carr et al. (2009) used the baseline and intervention level data of each participant to calculate the MBR value for the multiple-baseline designs.

Standardized Mean Difference (SMD)

In the meta-analyses studies with multiple subjects, the standardized mean difference (Cohen's *d*, Hedges' *g* and Glass' delta) indexes are used to calculate the effect size. It has been suggested by researchers that a similar value to the standardized mean difference can be used in single-subject experimental studies (Busk & Serlin, 1992; Shadish, Hedges, & Pustejovsky, 2014). The following steps can be followed when calculating the SMD value (Busk & Serlin, 1992) to demonstrate the effectiveness of the experiment in single-subject experimental studies:

1. Calculate the mean of the baseline level,
2. Calculate the mean of the intervention level,
3. Calculate the standard deviation of the baseline level,
4. Subtract the mean of the baseline level from the mean of the intervention level,
5. The value in the fourth step is divided by the standard deviation of the baseline level (Campbell, 2003).

To put it another way; the SMD value is calculated with the following equation:

$$\text{SMD} = \frac{M_B - M_A}{S_A} \tag{3}$$

where M_A represents the mean of baseline level and M_B represents the mean of intervention level and S_A represents the standard deviation of the baseline level.

In order to calculate the SMD value using the data presented in Figure 6, we need to calculate mean values for both levels and only the standard deviation value for the baseline level. According to the data in Figure 6, the mean value of the baseline level is 6.8 and the mean value of the intervention level is calculated as 9.2. In addition, the standard deviation of the baseline level required to achieve the SMD value is 0.632. Thus, the SMD value is calculated as (9.2-6.8) /0.632 = 3.79. It is also shown by visual analysis and non-overlapping data-based indexes in which the effect of the intervention in Figure 6 is not very large. However, the result of the SMD causes the intervention to be seen as a very effective intervention. In the light of this example, it is likely that the SMD will provide misleading results in terms of its effectiveness.

In contrast to most other recommended indexes, the standardized mean difference (SMD) has a known sampling distribution, and allows statistics such as regression and ANOVA to be applied to the common scale. With the help of Binomial (Binomial) sign test, the confidence intervals for this index can be calculated (Busk & Serlin, 1992). According to Olive and Franco (2008, p. 8), there are many strengths of the SMD method. The first one is the use of mean value in the calculations and possibility to calculate this mean value in both increasing and decreasing intervention level effect situations. Unlike the approaches based on non-overlapping data percentage, no data value in the SMD method is ignored. Another strength of the SMD method is that it can be interpreted in the same way as the known effect size values (e.g., Cohen's *d*). According to Olive and Franco (2008, p. 7), the SMD value obtained in experimental studies with single subjects can be interpreted according to the criteria proposed by Cohen (1988). According to these criteria, d = 0.2, 0.5, and 0.8 are interpreted as small, medium, and large

effect sizes, respectively. Researchers are advised to be cautious when using these criteria. Olive and Smith (2005) state that there are many different single-subject experimental designs, and for each of these designs, the same way should be followed by using different data to calculate the SMD value. For example, the individual SMD values should be calculated for each subject in the multiple baseline designs that include different subjects. The original baseline level and the last applied intervention level should be used in the SMD calculation in the reversal designs. The SMD value calculation is also possible for an ABA'B' design. In these cases, the researcher should calculate the SMD values (SMD_1 and SMD_2) for both AB designs (AB and A'B') and obtain the mean of these two values. Among the criticized aspects of the SMD value is the possibility of a possible trend effect due to time and failure to account for the change in slope between levels. The standardized mean difference is considered to be a problematic effect size calculation method because of the effect of the autocorrelation and the resulting magnitude of the effect size is usually too large. In this context, many researchers in recent years are striving to develop this index. Although it is frequently used by many researchers, it is thought to be problematic to use them in the context of meta-analyses without overcoming the above-mentioned limitations. In addition, SMD statistics does not take into account the dependence on the longitudinal data structure as in most other non-regression approaches. This leads to incorrect estimation of standard error rates. Therefore, this may lead to an increase in the Type I or Type II error rates.

Reporting of Effect Size Indexes

The effect size values obtained from different studies must be gathered and interpreted to be used in a meta-analysis study. According to Maggin et al. (2011), the successful combination of effect sizes from different studies is useful to obtain generalizable results and to show to other researchers where the amount of effect at the intervention level is more or less. The summarization of the effect size values obtained from different studies is achieved by obtaining weighted average through traditional meta-analysis models in studies with multiple subjects. The proposed steps for traditional meta-analysis (literature review, selection of studies, calculation of effect sizes and summarization) also apply to the meta-analysis of single-subject experimental studies. The difference between multiple- and single-subject experimental studies is that how to summarize the effect size values obtained from different studies.

According to the research conducted by Maggin et al. (2011), researchers apply five different methods in experimental subjects. The most preferred of these methods is to summarize the effect size values obtained from all studies by calculating the mean effect size value. The second most commonly used summation method is to obtain an average value from all studies using a weighted average. In order to calculate the weighted average, it is necessary to obtain values such as inverse variance weights, confidence intervals or standard errors in the traditional meta-analysis. It is also possible to obtain weighted average with multilevel models. Other methods used to summarize effect sizes include reporting the median value of all studies, providing descriptive explanations, or showing the percentages of effective and ineffective studies.

In single-subject experimental studies, the researcher needs to pay attention to several points in obtaining the common effect size value using one of the methods described above (Vannest & Davis, 2013, pp. 107-108). First of all, the researcher who will make a meta-analysis should know the characteristics of single-subject experimental research. The researcher should be well aware of the levels in which a single subject will be compared between the levels of experimental research (Vannest & Davis, 2013). The calculation of the indexes presented above is shown for the AB design. Almost all of these indexes can also be applied to complex single-subject experimental designs. Even in these complex designs, most researchers obtain the index values described above by selecting a level A and a level B. What is important is to justify why the researcher has made such a decision. In addition, the researcher who is planning to make a meta-analysis in single-subject experimental studies should explain the method of calculating the effect size index of his/her choice and should provide a detailed description of the calculation method. WWC (Kratochwill et al., 2010) and many researchers recommend calculating and reporting the effect sizes using multiple indexes until a single effect size index is developed that is sensitive to all conditions specific to the data patterns obtained from single-subject experimental studies.

In addition to reporting the effect size value for each study included in the meta-analysis, the meta-analyst should also report details such as the number of participants in the research, the type and the number of behaviors examined, the number of studies and the type of weighting used (Vannest & Davis, 2013). Confidence intervals should also be provided when reporting effect size index values.

DISCUSSION and CONCLUSION

In this study, 10 indexes, which are used as effect size values in single subject experimental research, were examined. It was tried to explain how to calculate these indexes developed by different researchers on sample data. It is expected that this study will benefit the researchers with single-subject experimental studies. Firstly, PND and PZD indexes, which are the most preferred indexes in domestic and international literature, are explained and, the properties of other alternative indexes are presented. In this study, it was observed in the literature that the PND was preferred due to the fact that it is easily computable among the indexes compared, but the limitations were ignored. It has been underlined that some of the other indexes like PND (PAND, NAP, PEM and Tau) offer solutions to the limitations of PND but do not offer the advantages of traditional effect size values.

The fact that the effect size indexes are used in different behavior situations are among the distinguishing features of these indexes. While most of the indexes examined in our study are used in cases where target behavior is expected to increase or decrease, PZD index is used only in cases where the expected behavior is expected to disappear. If the researcher is doing an intervention on the elimination of a behavior in the participant, the index to be preferred may be the PZD. Another difference is the number of data points used in the calculation indexes. Some indexes only take into account data values at the intervention level (e.g., PZD), while some indexes (PND, PEM, PAND, etc.) take into account one or a few of the baseline data points. Therefore, it would be more accurate to use indexes that take into account all the data in both the baseline and intervention levels (NAP, Tau, Tau-U, SMD, and MBR). Statistics that take into account all of the values in the data are less influenced by outliers. In particular, the NAP, Tau and Tau-U indexes tend to produce more accurate results because they have more statistical power than other nonparametric indexes (Parker et al., 2011). Ignoring some data points can lead to misleading decisions about the effectiveness of the intervention level. In addition to taking into account all data points, indexes such as SMD and MBR are characterized by being able to be transformed into traditional effect size values or interpreted in the same way. In this context, it is observed that the values obtained in the SMD calculations are higher than the group design studies. Another index with this feature is the PAND index which can be converted to *d* value by obtaining Phi value. Most of the percentage indexes presented in this study do not produce values that can be interpreted as effect size values obtained from studies with multiple subjects. Another limitation of the indexes covered in this study is that most indexes cannot take into account the effect of autocorrelation and trend in single-subject experimental studies (as in time series). The autocorrelation effect is defined as the positive or negative correlation of repeated data collected from the same individual. This effect leads to incorrect calculation of the standard deviation values. The index values, which can take this negative effect into account (Tau-U), are found more ideal by the researchers. In cases of autocorrelation and data without trend, the use of PND and PEM-T may be recommended to practitioners to determine whether an intervention is effective in terms of ease. However, it would be more appropriate to use more advanced methods (e.g., Tau-U) in the context of meta-analyses.

REFERENCES

Alresheed, F., Hott, B. L., & Bano, C. (2013). Single subject research: A synthesis of analytic methods. *The Journal of Special Education Apprenticeship*, *2*(1), 1–18.

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case∗. *Behaviour Research and Therapy*, *31*(6), 621–631.

Allison, D. B., & Gorman, B. S. (1994). "Make things as simple as possible, but no simpler." A rejoinder to Scruggs and Mastropieri. *Behaviour Research and Therapy*, *32*(8), 885–890.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Aslan, C., Yalçın, G. & Özdemir, S. (Mayıs, 2016). *Sosyal öykü tekniğinin etkililiği: Betimsel değerlendirme ve meta-analiz çalışması*. Teacher Education in Special Education, Vocational Training and Sports Konferansı (ELMIS), Konya, Türkiye.

Aydın, O. (2017). *Otizm spektrum bozukluğu olan bireylere matematik becerilerinin öğretimi: tek-denekli araştırmalarda betimsel ve meta analiz* (Yayımlanmamış yüksek lisans tezi). Anadolu Üniversitesi, Eğitim Bilimleri Enstitüsü, Eskişehir.

Bell, R. J., Skinner, C. H., & Fisher, L. A. (2009). Decreasing putting yips in accomplished golfers via solution-focused guided imagery: A single-subject research design. *Journal of Applied Sport Psychology, 21*(1), 1–14.

Beretvas, S. N., & Chung, H. (2008). A review of meta-analyses of single-subject experimental designs: Methodological issues and practice. *Evidence-Based Communication Assessment and Intervention, 2*, 129–141.

Bozkuş-Genç., G. (2017). *Otizm spektrum bozukluğu olan çocuklara soru sorarak iletişim başlatmanın kazandırılmasında temel tepki öğretiminin etkileri* (Yayımlanmamış doktora tezi). Anadolu Üniversitesi, Eğitim Bilimleri Enstitüsü, Eskişehir.

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp.187–212). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Campbell, J. M. (2003). Efficacy of behavioral interventions for reducing problem behavior in persons with autism: A quantitative synthesis of single-subject research. *Research in Developmental Disabilities, 24*, 120–138.

Campbell, J. M. (2004). Statistical comparison of four effect sizes for single-subject designs. *Behavior Modification, 28*(2), 234–246.

Carr, J. E., Severtson, J. M., & Lepper, T. L. (2009). Noncontingent reinforcement is an empirically supported treatment for problem behavior exhibited by individuals with developmental disabilities. *Research in Developmental Disabilities, 30*(1), 44–57.

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

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3–8.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York, NY: Routledge.

Heyvaert, M., Saenen, L., Campbell, J. M., Maes, B., & Onghena, P. (2014). Efficacy of behavioral interventions for reducing problem behavior in persons with autism: An updated quantitative synthesis of single-subject research. *Research in Developmental Disabilities, 35*(10), 2463–2476.

Huitema, B. E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment, 7*, 107–118.

Huitema, B. E., & Mckean, J. W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement, 60*(1), 38–58.

Karasu, N. (2009a). Özel eğitimde delile dayalı yöntemlerin belirlenmesi: Tek denekli çalışma analizleri ve karşılaştırmaları. *Türk Eğitim Bilimleri Dergisi, 7*(1), 143–163.

Karasu, N. (2009b). Otizmden etkilenmiş bireylerde sosyal ve iletişim becerilerini arttıran yöntemlerin delile dayalı yöntem olarak belirlenmesi: Bir meta-analiz örneği. *Türk Eğitim Bilimleri Dergisi, 7*(3), 713–739.

Karasu, N. (2011). Otizmli bireylerin eğitiminde video ile model olma uygulamalarının değerlendirilmesi: Bir alanyazın derlemesi ve meta-analiz örneği. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Özel Eğitim Dergisi, 12*(2), 1–12.

Kavale, K. A. (2001). Decision making in special education: The function of meta-analysis. *Exceptionality, 9*, 245–268.

Kaya, F. (2015). *Otizm spektrum bozukluğu olan öğrencilere yiyecek-içecek hazırlama becerilerinin öğretiminde sesli anlatım içeren ve içermeyen video ipucunun karşılaştırılması* (Yayımlanmamış yüksek lisans tezi). Anadolu Üniversitesi, Eğitim Bilimleri Enstitüsü, Eskişehir.

Kırcaali-İftar, G., & Tekin, E. (1997). *Tek denekli araştırma yöntemleri*. Ankara: Türk Psikologlar Derneği Yayınları.

Korkmaz, Ö. T., & Diken, İ. H. (2010). Stereotipik davranışların azaltılmasında kullanılan yöntemlerin etkililiği: Betimsel ve meta analizi. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Özel Eğitim Dergisi, 11(2)*, 1–12.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). Single-case designs technical documentation. *What works clearinghouse*. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Lydon, S., Healy, O., O'Reilly, M., & McCoy, A. (2013). A systematic review and evaluation of response redirection as a treatment for challenging behavior in individuals with developmental disabilities. *Research in Developmental Disabilities, 34*(10), 3148–3158.

Ma, H. (2006). An alternative method for quantifying synthesis of single-subject research: Percent of data points exceeding the median. *Behavior Modification, 30*, 598–617.

Maggin, D. M., O'Keeffe, B. V., & Johnson, A. H. (2011). A quantitative synthesis of methodology in the meta-analysis of single-subject research for students with disabilities: 1985–2009. *Exceptionality, 19*(2), 109–135.

Maggin, D. M., Swaminathan, H., Rogers, H. J., O'keeffe, B. V., Sugai, G., & Horner, R. H. (2011). A generalized least squares regression approach for computing effect sizes in single-case research: Application examples. *Journal of School Psychology, 49*(3), 301–321.

Manolov, R., Sierra, V., Solanas, A., & Botella, J. (2014). Assessing functional relations in single-case designs: Quantitative proposals in the context of the evidence-based movement. *Behavior modification, 38*(6), 878–913.

Manolov, R., & Solanas, A. (2008). Comparing N=1 effect size indices in presence of autocorrelation. *Behavior Modification, 32*, 860–875.

Manolov, R., Solanas, A., & Leiva, D. (2010). Comparing "visual" effect size indices for single-case designs. *Methodology, 6*, 49–58.

Manolov, R., Solanas, A., Sierra, V., & Evans, J. J. (2011). Choosing among techniques for quantifying single-case intervention effectiveness. *Behavior Therapy, 42*(3), 533–545.

O'Brien, S., & Repp, A. C. (1990). Reinforcement-based reductive procedures: A review of 20 years of their use with persons with severe or profound retardation. *Journal of the Association for Persons with Severe Handicaps, 15*(3), 148–159.

Olive, M. L., & Franco, J. H. (2008). (Effect) size matters: And so does the calculation. *The Behavior Analyst Today, 9*(1), 5–10.

Olive, M. L., & Smith, B. W. (2005). Effect size calculations and single subject designs. *Educational Psychology, 25*(2-3), 313–324.

Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data (PAND): An alternative to PND. *Journal of Special Education, 40*, 194–204.

Parker, R. I., & Vannest, K. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy, 40*(4), 357–367.

Parker, R. I., Vannest, K. J., & Brown, L. (2009). The improvement rate difference for single-case research. *Exceptional Children, 75*(2), 135–150.

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification, 35*(4), 303–322.

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy, 42*(2), 284–299.

Rakap, S. (2015). Effect sizes as result interpretation aids in single-subject experimental research: description and application of four nonoverlap methods. *British Journal of Special Education, 42*(1), 11–33.

Reichle, J. (2007). Amongst methodologies of functional behavioral assessment, functional analysis yields more effective suppression outcomes. *Evidence-Based Communication Assessment and Intervention, 1*(4), 153–155.

Schlosser, R. W., & Koul, R. K. (2015). Speech output technologies in interventions for individuals with autism spectrum disorders: a scoping review. *Augmentative and Alternative Communication, 31*(4), 285–309.

Scotti, J. R., Evans, I. M., Meyer, L. H., & Walker, P. (1991). A meta-analysis of intervention research with problem behavior: Treatment validity and standards of practice. *American Journal on Mental Retardation, 96*, 233–256.

Scruggs, T. E., & Mastropieri, M. A. (1998). Summarizing single-subject research: Issues and applications. *Behavior Modification, 22*(3), 221–242.

Scruggs, T. E., & Mastropieri, M. A. (2013). PND at 25: Past, present, and future trends in summarizing single-subject research. *Remedial and Special Education, 34*(1), 9–19.

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education, 8*, 24–33.

Scruggs, T. E., Mastropieri, M. A., Cook, S. B., & Escobar, C. (1986). Early intervention for children with conduct disorders: A quantitative synthesis of single-subject research. *Behavioral Disorders, 11*, 260–71.

Shadish, W. R., Hedges, L. V., & Pustejovsky, J. E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology, 52*(2), 123–147.

Sönmez, M., & Diken, İ. H. (2010). Problem davranışların azaltılmasında işlevsel iletişim öğretiminin etkililiği: Betimsel ve meta-analiz çalışması. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Özel Eğitim Dergisi*, *11*(1), 1–16.

Strain, P. S., Kohler, F. W., & Gresham, F. (1998). Problems in logic and interpretation with quantitative syntheses of single-case research: Mathur and colleagues (1998) as a case in point. *Behavioral Disorders, 24*, 74–85.

Swaminathan, H., Rogers, H. J., Horner, R. H., Sugai, G., & Smolkowski, K. (2014). Regression models and effect size measures for single case designs. *Neuropsychological Rehabilitation, 24*(3-4), 554–571.

Tavil, Y. Z. ve Karasu, N. (2013). Aile eğitim çalışmaları: Bir gözden geçirme ve meta-analiz örneği. *Eğitim ve Bilim, 38*(168), 85–95.

Uysal, H. (2017). *Zihin yetersizliği olan öğrencilere temel toplama işlemlerinde akıcılık kazandırmada iki farklı uygulamanın karşılaştırılması* (Yayımlanmamış yüksek lisans tezi). Anadolu Üniversitesi, Eğitim Bilimleri Enstitüsü, Eskişehir.

Van den Noortgate, W., & Onghena, P. (2003). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments & Computers, 35,* 1–10.

Vannest, K.J., & Davis, H. S. (2013). Synthesizing single case research to identify evidence based treatments. In B. G. Cook, M.Tankersley, & T. J. Landrum (Eds.), *Evidence-Based practices* (pp. 93–119). UK: Emerald Group Publishing.

Wehmeyer, M. L., Palmer, S. B., Smith, S. J., Parent, W., Davies, D. K., & Stock, S. (2006). Technology use by people with intellectual and developmental disabilities to support employment activities: A single-subject design meta analysis. *Journal of Vocational Rehabilitation, 24*(2), 81–86.

White, O. R., & Haring, N. G. (1980). *Exceptional teaching* (2nd ed.). Columbus, OH: Merrill.

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *The Journal of Special Education, 44*(1), 18–28.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(1);49-67

Are Differentially Functioning Mathematics Items Reason of Low Achievement of Turkish Students in PISA 2015?*

Serkan ARIKAN**

Abstract
In PISA 2015 the average mathematics score of Turkey decreased dramatically. One of the reasons could be the psychometric properties of mathematics items of PISA 2015. Therefore, it is necessary to evaluate PISA mathematics items for language DIF. In the study, three different DIF detection methods were used: logistic regression (LR), Mantel-Haenszel (MH) and structural equation modeling (SEM). Eleven items were found to have DIF when Turkish and English speaking students were compared. The effect sizes of mathematics performance differences between Turkish and English speaking students before and after excluding DIF items did not change which indicated that DIF items did not cause Turkish students to perform lower than expected. All the DIF items were open response format in which answers were rated by experts and computers. The DIF items favoring Turkish students were mainly related to the basic cognitive process.

Key Words: PISA 2015, Mathematics Performance, DIF, Turkish Student, Low Achievement

INTRODUCTION

The Programme for International Student Assessment (PISA) aims to provide internationally comparable data for 15-year-old students' performance based on reading, mathematics and science. PISA is administered every 3 year which makes possible to monitor progress of educational systems. The results of PISA get great attention by educators, researchers and policy makers as PISA provides detailed information about more than 70 countries. PISA 2015 application had great coverage in which 35 OECD countries and 37 partner countries participated to the assessment. PISA has many additional important features that make it unique and different from other assessments. For instance, PISA links student performance results data with student level variables like students' background and attitudes towards learning and with school level variables like school characteristics. PISA aims to measure students' capacity to apply knowledge and skills in key subjects which is defined as "literacy". (OECD, 2016a).

Turkey, a member of OECD, participates PISA regularly since 2003. Turkey's performances on mathematics were below the average score of 500; 423 in PISA 2003, 424 in PISA 2006, 445 in PISA 2009, 448 in PISA 2012, and 420 in PISA 2015. Similarly the average science scores of Turkey were 434 in PISA 2003, 424 in PISA 2006, 454 in PISA 2009, 463 in PISA 2012, and 425 in PISA 2015; the average reading scores of Turkey were 441 in PISA 2003, 447 in PISA 2006, 464 in PISA 2009, 475 in PISA 2012, and 428 in PISA 2015 (MEB, 2015; MEB, 2016). Through PISA 2012, Turkey had an increasing trend in the scores, however, in PISA 2015 the average scores decreased dramatically. The reasons of this very low score on PISA 2015 are necessary to be investigated.

There might be several reasons of the low scores of Turkish students in PISA 2015. There might be a problem in psychometric properties of items that were used in the PISA 2015 assessment; there might be a problem in the comparability of the samples over years; the change in test administration method (computer based administration instead of paper and pencil test) might cause lower scores. It

*An early draft of this paper was presented at European Conference on Educational Research (ECER) in Bolzano, Italy in 2018.

**Asst. Prof. Dr., Mugla Sitki Kocman University, Faculty of Education, Mugla-Turkey, serkanarikan@mu.edu.tr, ORCID ID: 0000-0001-9610-5496

To cite this article:
Arıkan, S. (2019). Are differentially functioning mathematics items reason of low achievement of Turkish students in PISA 2015?. *Journal of Measurement and Evaluation in Education and Psychology, 10*(1), 49-67. DOI: 10.21031/epod.466860

Received: 03.10.2018
Accepted:12.02.2019

is also possible that the low scores might be as a result of the change in the curriculum, educational practices or country level educational policies in Turkey. This study focused on the psychometric properties of the PISA 2015 mathematics items as a source of low scores of Turkish students.

Comparative assessments should be fair to all groups of students. Psychometric properties of these assessments should be controlled to prevent any unintended bias. PISA is mainly developed in English first and then adapted to other languages including Turkish (OECD, 2017). Therefore, it is necessary to evaluate whether PISA mathematics items functioned differently for Turkish and English speaking students who answered adapted items and original items, respectively. Finding evidence for fairness of items in terms of psychometric properties could help us to eliminate one of the possible reasons of sharp decrease of Turkish students' mathematics performance in 2015.

Differential item functioning (DIF) detection methods are widely used to evaluate the fairness and equality of tests on item level in investigating the comparability of translated and/or adapted measures (Zumbo, 2007). DIF occurs and threatens the comparability of scores if students with the similar ability level on the underlying construct in different groups do not have the similar probability of getting the right answers for a specific item (van de Vijver & Leung, 1997; Zumbo, 2007). Evaluating items in terms of DIF is a necessary preliminary analysis before conducting any comparative study. Otherwise, if a test contains items having DIF, observed differences in scores could be related to the problematic items rather than true differences on the underlying trait or ability (He & van de Vijver, 2013). If an item is detected as having DIF statistically, the context of the item should be examined by experts to evaluate whether the item indeed biased against one group systematically (van de Vijver & Leung, 1997). However, judgmental expert evaluation alone might not be always successful to detect why DIF occurs. For example, Angoff (1993) reported that even item writers often had problems to understand why some perfectly reasonable items showed large amounts of DIF. Some scholars investigated whether student background variables could be potential explanations of sources of DIF (Joldersma & Bowen, 2010; Liu et al., 2016; Zumbo & Gelin, 2007).

PISA items are prepared very carefully under the guidance of the experts by international team of item developers. Translatability reviews are conducted considering translation, adaptation and cultural issues (OECD, 2017). However, many researchers reported that PISA mathematics items contained DIF (Demir & Kose, 2014; Kankaras & Moors, 2014; Lyons-Thomas, Sandilands, & Ercikan, 2014; Yildirim & Berberoglu, 2009). Yildirim and Berberoglu (2009) reported that 5 out of 21 mathematics items in PISA 2003 were flagged as having DIF in the comparison of Turkish and American students (3 of these items favored Turkish students). Lyons-Thomas et al. (2014) found that there were gender DIF in PISA 2009 mathematics items of students in Canada, Finland, Shanghai, and Turkey. Demir and Kose (2014) identified many DIF items in PISA 2009 mathematics assessment when they compare answers of Turkish students with German, Finish and American students. Therefore, there is a possibility that PISA 2015 mathematics items might contain DIF items that could cause a decline in Turkish students' mathematics scores. There is not any study that investigated whether PISA 2015 items contained DIF across Turkish and English speaking students.

Purpose of the Study

Having DIF items for a language group is a threat to comparability of test scores. In this study, PISA 2015 mathematics items were analyzed in terms of DIF for Turkish, English and American students. The main idea is that whether the low mathematics scores of Turkish students could be due to DIF items against Turkish students. Therefore, in order to test this claim, DIF analyses using answers of Turkish and English student, as well as Turkish and American students were conducted separately. The research questions of this study were

(1) Are there any items having DIF in PISA 2015 mathematics test in comparing Turkish and English students?

(2) Are there any items having DIF in PISA 2015 mathematics test in comparing Turkish and American students?

(3) Are there any changes in the effect sizes of mathematics performance differences among groups before and after excluding DIF items, if any?

METHOD

Participants

The data of this study were obtained from the PISA 2015 data set. In PISA, the target population is all 15-year-old students of participating countries. PISA has rotated booklet design in which each student answers linked portion of all items. Therefore, ability level of each student could be estimated from all items without requiring a student to answer all items (OECD, 2016b). This study used the data of all Turkish, English and American students who answered mathematics items in booklets 43, 45, and 47. These three booklets were selected because they included all the items and there were no overlap of items. The participants were 491 Turkish students, 1154 English students and 448 American students.

Instrument

In PISA 2015, a total of 69 mathematics items were used to collect information about students' mathematics performance and a student responded approximately 23 mathematics items. PISA aims to measure mathematical literacy level of students defined as the capacity of students to apply acquired knowledge and skills to different problems and challenges they encounter. The mathematical processes measured in PISA are formulate (formulating situations mathematically), employ (employing mathematical concepts, facts, procedures and reasoning), and interpret (interpreting, applying and evaluating mathematical outcomes) (OECD, 2016b). These mathematical processes have a hierarchical order in which interpret represents the highest cognitive process. In Table 1, Table 2 and Table 3, item number, item code, item label, item format, cognitive processes measured by each item and item-level percentage correct values for Turkish, English and American students in booklet 43, 45 and 47 were reported.

Data Analysis

In the study, three different DIF detection methods were used. These DIF detection methods were logistic regression (LR), Mantel-Haenszel (MH) and structural equation modeling (SEM). As each DIF method is based on different statistical procedures, and studies reported that there might be low to medium coherence among DIF detection methods (Atalay, Gok, Kelecioglu & Arsan, 2012), more than one method was used. In order to get more consistent findings, an item that showed DIF in at least two different methods was considered to contain DIF across language groups. Sixty-nine mathematics items were evaluated in terms of DIF for Turkish-English and Turkish-American students groups.

In the logistic regression method, as the first step, only total score (model 1), then total score and grouping variable (model 2), and finally total score, grouping variable and their interaction (model 3) were used as predictors. Significance of country and their interaction, and the change in R^2 value were taken as evidence for uniform bias and non-uniform bias, respectively (Zumbo, 1999). Zumbo and Thomas (1997) proposed that ΔR^2 (the difference between model 3 and model 1) higher than 0.130 indicates moderate DIF and higher than 0.260 indicates large DIF. Jodoin and Gierl (2001) proposed lower values to detect DIF; ΔR^2 higher than 0.035 indicates moderate DIF and higher than 0.070 indicates large DIF. In this study the criteria of Jodoin and Gierl was used to detect DIF items as it requires lower values which allows to detect more items. Therefore, the possibility to omit an

item that might have bias will be minimized. SPSS 22.0 programs were used to conduct logistic regression analysis.

Table 1. Item Descriptions for Booklet 43

Item No	Item Code	Item Label	Item Format	Cognitive Domain	Turkish p value	English p value	American p value
B43_1	CM033Q01S	A View Room-Q01	SMC	interpret	.56	.74	.75
B43_2	CM474Q01S	Running Time-Q01	SMC	employ	.44	.63	.64
B43_3	DM155Q02C	Population Pyramids-Q02	OR	interpret	.22	.57	.43
B43_4	CM155Q01S	Population Pyramids-Q01	CMC	employ	.46	.66	.63
B43_5	DM155Q03C	Population Pyramids-Q03	OR	employ	.07	.13	.14
B43_6	CM155Q04S	Population Pyramids-Q04	CMC	interpret	.32	.54	.43
B43_7	CM411Q01S	Diving-Q01	OR	employ	.25	.52	.43
B43_8	CM411Q02S	Diving-Q02	SMC	interpret	.29	.48	.51
B43_9	CM803Q01S	Labels-Q01	OR	formulate	.10	.28	.20
B43_10	CM442Q02S	Braille-Q02	CMC	interpret	.14	.20	.25
B43_11	DM462Q01C	Third Side-Q01	OR	employ	**.13**	**.01**	**.03**
B43_12	CM034Q01S	Bricks-Q01	OR	formulate	.17	.32	.23
B43_13	CM305Q01S	Map-Q01	SMC	employ	.31	.39	.42
B43_14	CM496Q01S	Cash Withdrawal-Q01	CMC	formulate	.23	.47	.41
B43_15	CM496Q02S	Cash Withdrawal-Q02	OR	employ	.47	.68	.59
B43_16	CM423Q01S	Tossing Coins-Q01	SMC	interpret	**.77**	.84	**.71**
B43_17	DM406Q01C	Running Tracks-Q01	OR	employ	**.09**	.24	**.07**
B43_18	DM406Q02C	Running Tracks-Q02	OR	formulate	.01	.08	.04
B43_19	CM603Q01S	Number Check-Q01	CMC	employ	.23	.32	.31
B43_20	CM571Q01S	Stop The Car-Q01	SMC	interpret	.22	.39	.34
B43_21	CM564Q01S	Chair Lift-Q01	SMC	formulate	**.39**	**.37**	.41
B43_22	CM564Q02S	Chair Lift-Q02	SMC	formulate	.33	.42	.35

Note: CMC: Complex Multiple Choice; OR: Open Response; SMC: Simple Multiple Choice

Table 2. Item Descriptions for Booklet 45

Item No	Item Code	Item Label	Item Format	Cognitive Domain	Turkish p value	English p value	American p value
B45_1	CM447Q01S	Tile Arrangement-Q01	SMC	employ	.52	.55	.53
B45_2	CM273Q01S	Pipelines-Q01	CMC	employ	**.37**	.37	**.32**
B45_3	CM408Q01S	Lotteries-Q01	CMC	interpret	.29	.40	.34
B45_4	CM420Q01S	Transport-Q01	CMC	interpret	.30	.54	.51
B45_5	CM446Q01S	Thermometer Cricket-Q01	OR	formulate	.65	.68	.67
B45_6	DM446Q02C	Thermometer Cricket-Q02	OR	formulate	.02	.08	.05
B45_7	CM559Q01S	Telephone Rates-Q01	SMC	interpret	**.54**	.59	**.49**
B45_8	DM828Q02C	Carbon Dioxide-Q02	OR	employ	.52	.66	.57
B45_9	CM828Q03S	Carbon Dioxide-Q03	OR	employ	.24	.29	.27
B45_10	CM464Q01S	Fence-Q01	OR	formulate	**.20**	**.19**	**.15**
B45_11	CM800Q01S	Computer Game-Q01	SMC	employ	**.88**	**.86**	**.78**
B45_12	CM982Q01S	Employment Data-Q01	OR	employ	.71	.84	.81
B45_13	CM982Q02S	Employment Data-Q02	OR	employ	.14	.40	.35
B45_14	CM982Q03S	Employment Data-Q03	CMC	interpret	.57	.63	.64
B45_15	CM982Q04S	Employment Data-Q04	SMC	formulate	.31	.49	.37
B45_16	CM992Q01S	Spacers-Q01	OR	formulate	.48	.70	.68
B45_17	CM992Q02S	Spacers-Q02	OR	formulate	.06	.11	.10
B45_18	DM992Q03C	Spacers-Q03	OR	formulate	**.05**	**.03**	.05
B45_19	CM915Q01S	Carbon Tax-Q01	SMC	employ	.31	.49	.39
B45_20	CM915Q02S	Carbon Tax-Q02	OR	employ	.54	.66	.61
B45_21	CM906Q01S	Crazy Ants-Q01	SMC	employ	.35	.61	.47
B45_22	DM906Q02C	Crazy Ants-Q02	OR	employ	.18	.39	.31
B45_23	DM00KQ02C	Wheelchair Basketball-Q02	OR	formulate	.02	.09	.05

Note: CMC: Complex Multiple Choice; OR: Open Response; SMC: Simple Multiple Choice

Table 3. Item Descriptions for Booklet 47

Item No	Item Code	Item Label	Item Format	Cognitive Domain	Turkish p value	English p value	American p value
B47_1	CM909Q01S	Speeding Fines-Q01	OR	interpret	.48	.90	.84
B47_2	CM909Q02S	Speeding Fines-Q02	SMC	employ	.20	.46	.51
B47_3	CM909Q03S	Speeding Fines-Q03	OR	interpret	.06	.24	.26
B47_4	CM949Q01S	Roof Truss Design-Q01	CMC	employ	.38	.67	.60
B47_5	CM949Q02S	Roof Truss Design-Q02	CMC	employ	.20	.33	.26
B47_6	DM949Q03C	Roof Truss Design-Q03	OR	formulate	.18	.24	.30
B47_7	CM00GQ01S	Advertising Column-Q01	OR	formulate	**.05**	.06	**.03**
B47_8	DM955Q01C	Migration-Q01	OR	interpret	.41	.79	.68
B47_9	DM955Q02C	Migration-Q02	OR	interpret	**.34**	**.30**	**.21**
B47_10	CM955Q03S	Migration-Q03	OR	employ	.01	.08	.05
B47_11	DM998Q02C	Bike Rental-Q02	OR	interpret	.52	.77	.84
B47_12	CM998Q04S	Bike Rental-Q04	CMC	employ	.28	.30	.28
B47_13	CM905Q01S	Tennis balls-Q01	CMC	interpret	.50	.70	.72
B47_14	DM905Q02C	Tennis balls-Q02	OR	interpret	.20	.41	.31
B47_15	CM919Q01S	Fan Merchandise-Q01	OR	employ	.69	.83	.75
B47_16	CM919Q02S	Fan Merchandise-Q02	OR	formulate	.21	.39	.40
B47_17	CM954Q01S	Medicine doses-Q01	OR	employ	.36	.64	.70
B47_18	DM954Q02C	Medicine doses-Q02	OR	employ	.13	.35	.33
B47_19	CM954Q04S	Medicine doses-Q04	OR	employ	.01	.29	.21
B47_20	CM943Q01S	Arches-Q01	SMC	formulate	.37	.45	.47
B47_21	CM943Q02S	Arches-Q02	OR	formulate	.00	.02	.01
B47_22	DM953Q02C	Flu test-Q02	OR	interpret	.11	.33	.31
B47_23	CM953Q03S	Flu test-Q03	OR	formulate	.12	.47	.38
B47_24	DM953Q04C	Flu test-Q04	OR	formulate	.00	.11	.07

Note: CMC: Complex Multiple Choice; OR: Open Response; SMC: Simple Multiple Choice

The Mantel-Haenszel DIF detection method is based on building of *K* two-by-two contingency tables, where K represents the number of discrete score categories that are used to match the comparison groups. For each matched score level, the expected and observed ratios are compared by chi-square method (Holland & Thayer, 1986). Then The MH D-DIF index is calculated using these comparisons with logarithmic transformations in which a negative value indicates the item favors reference group over the focal group (Holland & Thayer, 1988). Educational Testing Service (ETS) proposed a criterion to flag DIF items: The MH D-DIF index between 1 and 1.5 indicates moderate DIF and The MH D-DIF index higher than 1.5 indicated large DIF (Zieky, 1993). DIFAS 5.0 program was used for MH DIF detection analysis (Penfield, 2005).

In the SEM procedure, a Confirmatory Factor Analysis (unifactorial, with all items as indicators of the latent variable) is conducted to assess configural, metric and scalar invariance. The difference between incremental types of model fit is evaluated as the factor loadings and intercepts are forced to be equal for comparison groups (van de Vijver, 2017). If the difference in comparative fit index (CFI) and Tucker Lewis index (TLI) between configural, metric and the scalar invariance model is larger than 0.010, the modification indices are investigated to identify DIF items (Cheung and Rensvold, 2002). Mplus 7.4 program was used for SEM DIF detection procedure (Muthen & Muthen, 2015).

After detecting DIF items, the effect sizes of mathematics performance differences among student groups before and after excluding DIF items were calculated. The change in effect sizes was investigated. Effect size allows researchers to compare the difference between groups without being affected from sample size (Field, 2013). For comparing means of two groups, Cohen's *d* is frequently used as an indicator of effect size. Cohen's *d* is calculated as the difference between the group means divided by the pooled standard deviation. Cohens' *d* value around 0.2 is considered as a

small, around 0.5 represents a moderate and around 0.8 is considered as large effect size (Cohen, 1988).

RESULTS

Preliminary Analysis

Reliability Analysis of the Instrument

Cronbach's alpha reliability coefficients of the PISA 2015 mathematics tests for booklets 43, 45 and 47 were calculated as 0.78, 0.79, 0.76 for Turkish students, 0.81, 0.84, 0.85 for English students, and 0.80, 0.86, 0.86 for American students, respectively. These values indicated good internal consistency (Cicchetti, 1994).

DIF Results

In this section, results based on LR, MH and SEM DIF detection methods were presented. Overall results were compared at the end of this section.

Logistic Regression DIF Results

DIF results using LR method was presented in Table 4. In comparing answers of Turkish and English student, 10 out of 69 items (B43_11, B45_10, B45_13, B45_18, B47_1, B47_6, B47_7, B47_8, B47_9 and B47_19) were flagged as having DIF. When answers of Turkish and American student were compared, 14 out of 69 items (B43_11, B43_15, B43_16, B45_10, B45_11, B45_13, B45_18, B47_1, B47_6, B47_7, B47_9, B47_11, B47_14 and B47_19) were flagged as having DIF.

Table 4. Logistic Regression DIF Results

Item No	Booklet 43		Booklet 45		Booklet 47	
	TR-ENG ΔR2	TR-USA ΔR2	TR-ENG ΔR2	TR-USA ΔR2	TR-ENG ΔR2	TR-USA ΔR2
1	.012	.027	.014	.016	**.089****	**.064***
2	.024	.012	.014	.020	.001	.015
3	.033	.012	.008	.006	.000	.009
4	.008	.019	.014	.028	.003	.000
5	.002	.006	.011	.014	.013	.017
6	.007	.000	.003	.006	**.046***	**.039***
7	.004	.004	.003	.018	**.047***	**.057***
8	.010	.029	.005	.008	**.041***	.031
9	.004	.003	.016	.014	**.107****	**.194****
10	.017	.001	**.052***	**.059***	.009	.016
11	**.299****	**.147****	.030	**.094****	.005	**.043***
12	.001	.006	.005	.014	.006	.010
13	.005	.002	**.038***	**.053***	.000	.004
14	.003	.018	.005	.002	.014	**.048***
15	.003	**.036***	.002	.002	.011	.033
16	.011	**.045***	.013	.033	.009	.004
17	.012	.031	.009	.006	.000	.025
18	.011	.029	**.118****	**.121****	.001	.004
19	.003	.005	.003	.000	**.039***	**.056***
20	.010	.015	.018	.012	.005	.003
21	.021	.009	.015	.006	.001	.001
22	.008	.012	.013	.013	.005	.000
23	-	-	.013	.009	.019	.014
24	-	-	-	-	.014	.022

Note: * indicates the item shows moderate level of DIF; ** indicates the item shows large level of D

Mantel-Haenszel DIF Results

DIF results using MH method was presented in Table 5. In comparing answers of Turkish and English student, 10 out of 69 items (B43_11, B45_10, B45_13, B45_18, B47_1, B47_6, B47_7, B47_9, B47_10 and B47_19) were flagged as having DIF. When answers of Turkish and American student were compared, 10 out of 69 items (B43_11, B45_10, B45_13, B45_18, B47_1, B47_7, B47_9, B47_11, B47_14 and B47_19) were flagged as having DIF.

Table 5. Mantel-Haenszel DIF Results

Item No	Booklet 43		Booklet 45		Booklet 47	
	TR-ENG ΔMH	TR-USA ΔMH	TR-ENG ΔMH	TR-USA ΔMH	TR-ENG ΔMH	TR-USA ΔMH
1	-.215	-.566	.444	.306	**-1.634****	**-1.445***
2	-.212	-.390	.579	.630	-.166	-.648
3	-.969	-.504	.068	.312	-.2486	-.519
4	.168	-.072	-.551	-.951	-.0888	-.010
5	.124	-.263	.605	.198	.634	.539
6	-.306	-.041	-.642	-.573	**1.495***	.441
7	-.386	-.151	.142	.481	**1.196***	**1.850****
8	-.186	-.516	-.130	.074	-.945	-.403
9	-.561	-.124	.594	.273	**2.260****	**2.241****
10	.947	.102	**1.843****	**1.611****	**-1.057***	NA
11	**3.910****	**2.732****	.812	**1.107***	-.297	**-1.106***
12	-.030	.341	-.355	-.611	.095	.162
13	.109	-.213	**-1.078***	**-1.131***	-.036	-.212
14	-.262	-.275	.310	-.049	.755	**1.564****
15	.149	.259	-.181	.162	.468	.980
16	.168	1.040	-.593	-.893	.433	.219
17	-.306	.878	.591	.259	.108	-.617
18	-.802	-.616	**3.385****	NA	-.095	.207
19	.204	.018	-.215	-.160	**-3.060****	**-1.820****
20	-.132	-.184	-.024	.006	.318	.099
21	.678	.306	-.681	-.201	NA	NA
22	.112	.339	-.540	-.450	-.263	-.068
23	-	**-**	-.751	-.071	-.923	-.421
24	-	**-**	**-**	**-**	NA	NA

Note: * indicates the item shows moderate level of DIF; ** indicates the item shows high level of DIF; NA indicates calculation problem due to low correct response ratio

SEM DIF Results

SEM DIF results were presented in Table 6. In comparing answers of Turkish and English student, 4 out of 69 items (B45_2, B45_10, B45_13 and B45_18) were flagged as having DIF. When answers of Turkish and American student were compared, 2 out of 69 items (B45_13 and B47_9) were flagged as having DIF.

Table 6. SEM DIF Results

Booklet	Model	χ^2/df	RMSEA	CFI	ΔCFI	TLI	ΔTLI	DIF ITEMS
43 TR-UK	Configural	1.192**	.027	.971		.967		None
	Metric	1.222**	.029	.966	.005	.962	.005	
	Scalar	1.232**	.029	.963	.003	.961	-.001	
43 TR-USA	Configural	1.140*	.030	.962		.958		None
	Metric	1.159*	.032	.957	.005	.952	.006	
	Scalar	1.162*	.032	.954	.003	.951	.001	
45 TR-UK	Configural	1.221**	.028	.967		.963		
	Metric	1.220**	.028	.967	.000	.964	-.001	
	Scalar	1.342***	.035	.946	.021	.943	.021	2, 10, 13, 18
	Scalar-Items Removed	1.309***	.033	.957	.010	.954	.010	
45 TR-USA	Configural	1.159*	.032	.960		.957		13
	Metric	1.158*	.032	.961	-.001	.957	.000	
	Scalar	1.199**	.036	.948	.013	.945	.012	
	Scalar-Items Removed	1.180**	.034	.957	.004	.954	.003	
47 TR-UK	Configural	1.558***	.045	.940		.934		None
	Metric	1.539***	.044	.939	.001	.936	-.002	
	Scalar	1.635***	.048	.929	.010	.925	.011	
	Scalar-Items Removed	1.621***	.048	.930	.009	.926	.010	
47 TR-USA	Configural	1.511***	.057	.901		.891		9
	Metric	1.549***	.059	.889	.013	.883	.008	
	Scalar	1.577***	.061	.883	.006	.877	.006	
	Scalar-Items Removed	1.531***	.058	.893	-.004	.887	-.004	

*$p < .05$. **$p < .01$. ***$p < .001$.

Overview of DIF Results

Since each DIF detection method is based on different calculations, an item flagged by at least two method was considered as containing DIF (Table 7). In comparing answers of Turkish and English student, 9 out of 69 items (B43_11, B45_10, B45_13, B45_18, B47_1, B47_6, B47_7, B47_9 and B47_19) were flagged as having DIF by at least two methods. It is necessary to report which items favored Turkish students and which items favored English students. Among these 9 items, 6 of them favored Turkish students (B43_11, B45_10, B45_18, B47_6, B47_7, B47_9) whereas 3 of them favored English students (B45_13, B47_1, B47_19).

When answers of Turkish and American student were compared, 10 out of 69 items (B43_11, B45_10, B45_11, B45_13, B47_1, B47_7, B47_9, B47_11, B47_14 and B47_19) were flagged as having DIF by at least two methods. Among these 10 items, 5 of them favored Turkish students (B43_11, B45_10, B47_7, B47_9, B47_14) whereas 4 of them favored American students (B45_13, B47_1, B47_11, B47_19). LR results suggested that item B45_11 had non-uniform DIF. The related graphical percentages were given in Appendix A and B. The flagged items were generally consistent across Turkish-English and Turkish-American student comparisons. Items B43_11, B45_10, B47_7 and B47_9 favored Turkish students whereas items B45_13, B47_1, B47_19 favored English speaking students.

Table 7. Overall DIF Results

Booklet	LR	MH	SEM	Items Commonly Flagged
43 TR-UK	11	11	-	11^{TR}
43 TR-USA	11, 15, 16	11	-	11^{TR}
45 TR-UK	10, 13, 18	10, 13, 18	2, 10, 13, 18	10^{TR}, 13^{UK}, 18^{TR}
45 TR-USA	10, 11, 13, 18	10, 11, 13	13	10^{TR}, 11^{*}, 13^{USA}
47 TR-UK	1, 6, 7, 8, 9, 19	1, 6, 7, 9, 10, 19	-	1^{UK}, 6^{TR}, 7^{TR}, 9^{TR}, 19^{UK}
47 TR-USA	1, 6, 7, 9, 11, 14, 19	1, 7, 9, 11, 14, 19	9	1^{USA}, 7^{TR}, 9^{TR}, 11^{USA}, 14^{TR}, 19^{USA}

Note: TR: items favoring Turkish students; UK: items favoring English students; USA: items favoring American students; * non-uniform DIF

Table 8 showed item formats and cognitive domains measured by the DIF items. All the DIF items were open response format in which students constructed the answers and then the answers were rated. Also, among 7 items that favored Turkish students 4 of them were related to formulate cognitive process which is the lowest cognitive process in PISA mathematics assessment. There was no formulate items that favored English or American students.

Table 8. Item Characteristics of DIF Items

Item No	Favoring	Item Label	Item Format	Cognitive Domain
B43_11	Turkish	Third Side - Q01	OR	Employ
B45_10	Turkish	Fence - Q01	OR	Formulate
B45_18	Turkish	Spacers - Q03	OR	Formulate
B47_6	Turkish	Roof Truss Design - Q03	OR	Formulate
B47_7	Turkish	Advertising Column - Q01	OR	Formulate
B47_9	Turkish	Migration - Q02	OR	Interpret
B47_14	Turkish	Tennis balls - Q02	OR	Interpret
B45_13	English&American	Employment Data - Q02	OR	employ
B47_1	English&American	Speeding Fines - Q01	OR	interpret
B47_11	American	Bike Rental - Q02	OR	interpret
B47_19	English&American	Medicine doses - Q04	OR	employ

Note: CMC: Complex Multiple Choice; OR: Open Response; SMC: Simple Multiple Choice

Effects of DIF Items on Mathematics Performance Differences

There were mathematics performance differences between Turkish students and English speaking students. Effect size, the standardized mean-difference, allows us to compare the difference between groups without being affected from sample size (Field, 2013). In this part, the original effect sizes and the effect sizes excluding DIF items were reported (Table 9). Between Turkish and English students, there were .51 to .93 effect size differences originally in these booklets. According to Cohen (1988), these values represent moderate to large difference between students. When all DIF items were excluded, effect sizes did not change. Similarly, between Turkish and American students, the original effect sizes were calculated as .28 to .85. According to Cohen (1988), these values represent small to large difference between students. When all DIF items were excluded, effect sizes were very close. The evaluation of the effect size change implied that DIF items generally balanced out each other and did not create any disadvantageous results for Turkish students.

Table 9. Effect Size Change

Booklet	43 TR-UK	43 TR-USA	45 TR-UK	45 TR-USA	47 TR-UK	47 TR-USA
Effect Size	.74	.53	.51	.28	.93	.85
All Items						
Effect Size	.78	.57	.51	.29	.94	.84
Excluding all DIF Items						
Effect Size	.78 Item11	.57 Item11	.54 Item10	.30 Item10	.86 Item1	.80 Item1
Excluding a DIF Item						
Effect Size			.48 Item13	.30 Item11	.96 Item6	.86 Item7
Excluding a DIF Item						
Effect Size			.53 Item18	.24 Item13	.94 Item7	.93 Item9
Excluding a DIF Item						
Effect Size					.99 Item9	.80 Item11
Excluding a DIF Item						
Effect Size					.91 Item19	.88 Item14
Excluding a DIF Item						
Effect Size						.83 Item19
Excluding a DIF Item						

Note: Item numbers given in the table represents the eliminated items.

DISCUSSION

This study has a great importance as it aimed to shed a light on possible causes of low mathematics scores of Turkish students in PISA 2015. Through PISA 2012, Turkey had an increasing trend in their mathematics scores, however, in PISA 2015 the average mathematics score decreased dramatically. In the study, whether the low performance of Turkish students could be due to differentially functioning items was investigated. As PISA is mainly developed in English first and then adapted to other languages including Turkish, evaluating whether PISA mathematics items functioned differently for Turkish and English speaking students was the main focus of the study. In comparing responses of Turkish and English students, 9 items (out of 69) were detected as having DIF. Similarly, 10 items were found to have DIF when Turkish and American students were compared. The surprising finding was that among these DIF items, more items favored Turkish students than they favored English or American students. The standardized mathematics performance differences (measured by effect-size) between Turkish and English speaking students before and after excluding DIF items did not change. Therefore, it is concluded that DIF items did not cause Turkish students to perform lower. Therefore, there is no evidence that PISA items created a disadvantage for Turkish students. Therefore, among possible reasons of low achievement of Turkish students, a problem due to the psychometric properties of PISA items was eliminated. There is still a further need to investigate and focus on other possible reasons of low achievement of 15-year-old Turkish students by conducting new comparative studies.

The possible reasons of these lower scores in PISA 2015 could be the problem of comparability of the Turkish samples over years; the effects of change in test administration method (computer based administration instead of paper and pencil test); the change in the curriculum, educational practices or country level educational policies. One of the reasons of the decrease in the PISA scores could be the selected sample of Turkey. The sampling procedure and coverage rates were reported in PISA technical reports. The coverage rates are important as they give clues about the representativeness of the population. Turkey's coverage rates in PISA were increased over years. The coverage rates were 36% in 2003, 47% in 2006, 57% in 2009, 68% in 2012 and 70% in 2015. Spaull (2017) studied coverage rates and sample of Turkey and he concluded that there was a large change in the proportions of Turkish students that were not sampled in PISA, therefore the validity of the comparisons of the results could have some problems. There is a need to conduct further studies on these sampling issue of Turkey. The other reason could be the change in the administration method of PISA. There was a shift from paper-and-pencil tests in PISA 2012 to computer-based tests (CBT) in 2015. There is a debate over effect of CBT on test results (Jerrim, 2016; Jerrim, Micklewright,

Heine, Salzer, & McKeown, 2018; Komatsu & Rappleye, 2017). Investigating possible effects of CBT on Turkish students' scores would be an informative study about the decrease in scores. Another reason of the decrease in the scores could be related to curriculum change and educational policies. Students who took PISA 2015 in Turkey were mainly 9th or 10th graders. In Turkey, there are frequent changes in curriculum and educational policies in all level of educational system. For instance, in 2012, when students who join the PISA 2015 administration were in 6th or 7th grade, the K-12 education system in Turkey has undergone some major changes and students were allowed to continue their high school in the form of distant education (Gün & Baskan, 2014). The effects of these curriculum and system changes on PISA scores are worth to investigate. The last but not the least, the congruence between educational practices in Turkey and cognitive skills measured in PISA might create a low score for Turkish students. As PISA aims to measure students' capacity to apply knowledge and skills that are related to be successful in modern societies (OECD, 2016a), acquiring curriculum related knowledge might not be enough to be successful in PISA. However, in TIMSS 2015, another large scale assessment that focus more on curriculum, Turkish students increased their scores in both mathematics and science (Yıldırım et al., 2016). A study focuses on the increase of scores on the curriculum focused large scale assessment but the decrease of scores on capacity focused large scale assessment of Turkish students would be informative.

This study found DIF items in mathematics assessment, however the DIF items did not lead Turkish students to perform lower in PISA 2015. The DIF flagged items were generally consistent across Turkish-English and Turkish-American student comparisons. Among 9 items that were flagged as DIF in Turkish and English student comparison, 7 of them were also flagged in Turkish-American comparison. As these items were not released, it was not possible to evaluate the content of items to speculate why these items contained DIF consistently across different comparison groups. There is a need to identify possible sources of DIF, hopefully after items are released. The results of the study were consistent with the other researchers who found DIF items in PISA (Demir & Kose, 2014; Kankaras & Moors, 2014; Lyons-Thomas, Sandilands, & Ercikan, 2014; Yildirim & Berberoglu, 2009).

Although mathematics items were not released, there was an information about item format and cognitive processes measured by each item. There were relationship between DIF items and their format and cognitive processes. First of all, all the DIF items were open response items in which students' answers were rated by experts or computers (OECD, 2017). Among 69 items, 18 open response items were coded by experts and 22 open response items were coded by computers. Multiple coding design was used to monitor coder reliabilities within and across countries. The open-ended coding system was used to simplify the coding process. National Project Managers of each country were expected to investigate the systematic pattern of irregularities. For OECD countries, the median within-country agreement of raters was 97.5% and the median across-country agreement of raters was 97.9% in mathematics. For Turkey, within-country agreement of raters was 97.7% and across-country agreement of raters was 93.9% which was the second lowest (OECD, 2017). As all DIF items were open response items, and across-country agreement of Turkey was lower than OECD countries, it would be informative to know whether the coding could cause an advantage or disadvantage for Turkish students. Another issue is that the DIF items favoring Turkish students were mainly related to formulate cognitive process. Formulate cognitive process is defined as formulating situations mathematically which is the lowest cognitive process in PISA. In Turkish educational system there are problems that teachers do not give adequate emphasis to develop higher cognitive processes. Turkish students generally encounter with items that are related to basic skills as comprehension rather than higher order thinking skills as problem solving (Arikan, van de Vijver & Yagmur, 2016; Doganay & Bal, 2010; Temur, 2012). Therefore, Turkish students' high familiarity of basic cognitive skills could cause more formulate items to be detected as having DIF.

In the study three different DIF identification methods were applied. Logistic regression and Mantel-Haenszel DIF methods gave similar results compared to structural equation modeling DIF method. Structural equation modeling DIF results were more conservative in detecting items as DIF compared to the two other methods. Although logistic regression and Mantel-Haenszel methods

produced similar results, logistic regression method detected more items as having DIF compared to Mantel-Haenszel method. Except one item in booklet 47 (TR-UK comparison), all items flagged by Mantel-Haenszel were also flagged by logistic regression method in all booklets for all comparisons. Therefore, in this study, it was observed that logistic regression method flagged more items as having DIF. On the other hand, structural equation modeling DIF method flagged items having DIF very rarely compared to other two methods. Atalay et al. (2012) compared logistic regression and Mantel-Haenszel methods in their simulation study and concluded that Mantel-Haenszel method was more sensitive in detecting DIF items. On contrary to this study, Gok, Kelecioglu and Dogan (2010) found more gender and school type DIF using Mantel-Haenszel method compared to logistic regression method in high school entrance examination items of Turkey. These findings indicate that different conditions and different methods could lead to different results in detecting DIF. Therefore, using more than one DIF detection methods is also advised according to results of this study and current literature.

Limitations

There are limitations to mention about the study. The major limitation is that since the items were not released, it was not possible to identify sources of DIF by investigating the content. Identifying possible causes could give information to item developers to decrease the number of DIF items.

REFERENCES

Angoff, W. (1993). Perspective on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–24). Hillsdale, NJ: Lawrence Erlbaum Associates.

Arikan, S., van de Vijver, F., & Yagmur, K. (2016). Factors contributing to mathematics achievement differences of Turkish and Australian Students in TIMSS 2007 and 2011. *Eurasia Journal of Mathematics, Science and Technology Education*, *12*, 2039-2059. doi:10.12973/eurasia.2016.1268a

Atalay Kabasakal, K., Gok, B., Kelecioglu, H., & Arsan, N. (2012). Comparing different differential item functioning methods: A simulation study. *Hacettepe University Journal of Education*, *43*, 270-281.

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

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284–290. doi:10.1037/1040-3590.6.4.284.

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

Demir, S., & Köse, İ. A. (2014). An analysis of the differential item function through Mantel-Haenszel, SIBTEST and Logistic Regression Methods. *Journal of Human Sciences*, *11*(1), 700-714.

Doganay, A., & Bal, A. P. (2010). The measurement of students' achievement in teaching primary school fifth year mathematics classes. *Educational Sciences: Theory and Practice*, *10*, 199-215.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics*. Sage.

Gök, B., Kellecioğlu, H., & Doğan, N. (2010). Değişen madde fonksiyonunu belirlemede Mantel–Haenszel ve Lojistik Regresyon tekniklerinin karşılaştırılması. *Eğitim ve Bilim*, *35*(156), 3-16.

Gün, F., & Baskan, G. A. (2014). New education system in Turkey (4+ 4+ 4): A critical outlook. *Procedia-Social and Behavioral Sciences*, *131*, 229-235.

He, J., & Van de Vijver, F. J. R. (2013). Methodological issues in cross-cultural studies in educational psychology. In G. A. D. Liem & A. B. I. Bernardo (Eds.), *Advancing cross-cultural perspectives on educational psychology: A festschrift for Dennis McInerney* (pp. 39-56). Charlotte, NC: Information Age Publishing.

Holland, P. W., & Thayer, D. T. (1986). *Differential item functioning and the Mantel-Haenszel procedure* (ETS Research Report No. RR-86-31). Princeton, NJ: ETS.

Holland, P. W. and Thayer, D. T. (1988). Differential item performance and Mantel-Haenszel procedure. En H.Wainer & H. I. Braun (Eds.), *Test Validity*, pp. 129-145. Hillsdale, N.J.: Erlbaum.

Jerrim, J. (2016). How Shift to Computer-based Tests Could Shake up PISA Education Rankings. *The Conversation*. Retrieved from http://theconversation.com/how-shift-to-computer-based-tests-could-shake-up-pisa-education-rankings-54869

Jerrim, J., Micklewright, J., Heine, J. H., Salzer, C., & McKeown, C. (2018). PISA 2015: how big is the 'mode effect'and what has been done about it?. *Oxford Review of Education*, *44*(4), 476-493.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329-349.

Joldersma, K., & Bowen, D. (2010). *Application of Propensity Models in DIF Studies To Compensate For Unequal Ability Distributions*. Paper presented at the annual meeting of National Council on Measurement in Education, Denver, CO.

Kankaraš, M., & Moors, G. (2014). Analysis of cross-cultural comparability of PISA 2009 scores. *Journal of Cross-Cultural Psychology*, *45*(3), 381-399.

Komatsu, H., & Rappleye, J. (2017). Did the shift to computer-based testing in PISA 2015 affect reading scores? A view from East Asia. *Compare: A Journal of Comparative and International Education*, *47*(4), 616-623.

Liu, Y., Zumbo, B. D., Gustafson, P., Huang, Y., Kroc, E., & Wu, A. D. (2016). Investigating Causal DIF via Propensity Score Methods. *Practical Assessment, Research & Evaluation*, *21*(13), 1-24.

Lyons-Thomas, J., Sandilands, D. D., & Ercikan, K. (2014). Gender Differential Item Functioning in Mathematics in Four International Jurisdictions. *Education &Science*, *39*(172), 20-32.

MEB (2015). *PISA 2012 Araştırması Ulusal Nihai Raporu*. Ankara. Retrieved from https://drive.google.com/file/d/0B2wxMX5xMcnhaGtnV2x6YWsyY2c/view

MEB (2016). *PISA 2015 Ulusal Raporu*. Ankara. Retrieved from http://pisa.meb.gov.tr/wp-content/uploads/2016/12/PISA2015_Ulusal_Rapor1.pdf

Muthen, B. O., & Muthen, L. K. (2015). *Mplus (Version 7.4)*. California. Los Angeles.

OECD (2016a). *PISA 2015 Results (Volume I): Excellence and Equity in Education.* Paris: OECD Publishing. doi:10.1787/9789264266490-en

OECD (2016b). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic and FinancialLiteracy.* PISA, OECD Publishing, Paris.doi:10.1787/9789264255425-en

OECD (2017). *PISA 2015 Technical Report.* Paris: OECD Publishing. Retrieved from http://www.oecd.org/pisa/data/2015-technical-report/

Penfield, R. D. (2005). DIFAS: Differential Item Functioning Analysis System. *Applied Psychological Measurement, 29*, 150-151.

Spaull, N. (2017). *Who Makes It Into PISA?: Understanding the Impact of PISA Sample Eligibility Using Turkey as a Case Study (PISA 2003 - PISA 2012).* OECD Education Working Papers, No. 154, OECD Publishing, Paris. http://dx.doi.org/10.1787/41d175fc-en

Temur, Ö. D. (2012). Analysis of prospective classroom teachers' teaching of mathematical modeling and problem solving. *Eurasia Journal of Mathematics, Science & Technology Education, 8*(2), 83-93. doi:10.12973/eurasia.2012.822a

Van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis of comparative research*. Thousand Oaks, CA: Sage.

Van de Vijver, F. J. R. (2017). Capturing bias in structural equation modeling. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis. Methods and applications* (2nd, revised edition). New York, NY: Routledge.

Yıldırım, A., Özgürlük, B., Parlak, B., Gönen, E., & Polat, M. (2016). *TIMSS 2015 ulusal matematik ve fen bilimleri ön raporu 4. ve 8. sınıflar*. TC Milli Eğitim Bakanlığı Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü, Ankara.

Yildirim, H. H., & Berberoğlu, G. (2009). Judgmental and statistical DIF analyses of the PISA-2003 mathematics literacy items. *International Journal of Testing*, *9*(2), 108-121.

Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–364). Hillsdale, NJ: Lawrence Erlbaum.

Zumbo, B. D. (1999). Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language assessment quarterly*, *4*(2), 223-233.

Zumbo, B. D., & Gelin, M. N. (2005). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological/community moderated (or mediated) test and item bias. *Journal of Educational Research & Policy Studies*, *5*(1), 1-23.

Zumbo, B. D., & Thomas, D. R. (1997). *A measure of effect size for a model-based approach for studying DIF*. Prince George, Canada: Edgeworth Laboratory for Quantitative Behavioral Science, University of Northern British Columbia.

PISA 2015'de Türk Öğrencilerin Düşük Başarı Göstermelerinin Nedeni Değişen Madde Fonksiyonu (DMF) içeren maddeler midir?

GİRİŞ

Uluslararası Öğrenci Değerlendirme Programı (PISA) 15 yaşındaki öğrencilerin okuma, matematik ve fen okuryazarlığı alanlarındaki becerilerini uluslararası karşılaştırmalara olanak veren bir yapıda ölçmektedir. Katılan ülke sayısının giderek arttığı PISA'ya 70'in üzerinde ülke dahil olmaktadır. (OECD, 2016a). OECD üyesi olan Türkiye PISA'ya 2003 yılından beri düzenli olarak katılmaktadır. Ortalama puanın 500 olduğu PISA matematik okuryazarlık testinde, Türkiye PISA 2003'de 423, PISA 2006'da 424, PISA 2009'da 445, PISA 2012'de 448 ve PISA 2015'de 420 ortalama puan almıştır (MEB, 2015; MEB, 2016). Benzer bir değişim hem fen hem de okuma alanlarında da mevcuttur. PISA 2012'ye doğru artan yönde olumlu gelişmeler yaşanırken, 2015 yılında ciddi bir düşüşün yaşanması oldukça dikkat çekicidir. Bu düşüşün nedenlerinin araştırılması gerekmektedir. Nedenlerden bir tanesi ölçme aracında kullanılan maddelerin dil açısından yanlılık göstermeleri olabilir. Ölçme sonuçlarının sınavın uygulandığı dilden bağımsız olarak sonuçlar üretmesi beklenir. PISA soruları çoğunlukla İngilizce olarak geliştirilmekte, ardından diğer dillere adaptasyonu yapılmaktadır (OECD, 2017). Bu sebeple PISA matematik sorularının Türkçe ve İngilizce konuşan ülkelerdeki öğrenciler için değişen madde fonksiyonu (DMF) gösterip göstermediğinin incelenmesi gereklidir. Bu çalışmada Türkiye'deki öğrencilerin düşük puan alma nedeninin maddelerin DMF içermeleri olup olmadığı incelenecek, eğer neden bu değil ise de bu ihtimal elenerek, diğer ihtimallere odaklanılacaktır.

DMF tespit etme yöntemleri kullanılarak testlerin madde bazında yanlılık gösterip göstermediği ile ilgili ön inceleme yapılabilmektedir (Zumbo, 2007). DMF'nin ve sonrasında madde yanlılığının ortaya çıkması öğrenci gruplarının puanlarını doğru bir şekilde karşılaştırmayı engellemektedir. Aynı beceri düzeyine sahip iki öğrenci grubunun bir soruyu yanıtlama olasılıkları farklılaştığında DMF ortaya çıkmaktadır (van de Vijver & Leung, 1997; Zumbo, 2007). Bir maddede istatistiksel olarak DMF çıkarsa, uzmanlar o soruyu incelemeli ve neden DMF çıktığını yorumlayarak maddenin ilgili gruplar için yanlılık gösterip göstermediğine karar vermelidir (van de Vijver & Leung, 1997).

PISA soruları oldukça geniş bir uzman kadrosu tarafından titizlikle hazırlanmakta ve adaptasyon süreçleri gerçekleştirilmektedir (OECD, 2017). Ancak yine de, araştırmalar PISA matematik sorularında DMF içeren maddeler olduğunu raporlamışlardır (Demir & Kose, 2014; Kankaras & Moors, 2014; Lyons-Thomas, Sandilands, & Ercikan, 2014; Yildirim & Berberoglu, 2009). Bu sebeple PISA 2015 maddelerini de DMF içerip içermedikleri bakımından incelemek faydalı olacaktır. Alan yazında PISA 2015 maddelerini Türk öğrenciler ve İngilizce konuşan öğrenciler bakımından DMF için karşılaştıran bir çalışmaya rastlanmamıştır.

Bu amaçla bu çalışmada Türk, İngiliz ve Amerikan öğrencilerin matematik sorularına verdikleri yanıtlar DMF içerip içermedikleri yönünden incelenmiştir. Türk öğrencilerin düşük matematik performansı gösterme nedenlerinden birisi olarak DMF içeren maddelerin olup olmaması incelenmiştir. Araştırma soruları ise

(1) Türk ve İngiliz öğrencileri karşılaştırıldığında, DMF içeren PISA 2015 matematik sorusu var mıdır?

(2) Türk ve American öğrencileri karşılaştırıldığında, DMF içeren PISA 2015 matematik sorusu var mıdır?

(3) DMF içeren maddeler testten çıkarıldığında matematik performans farklarından ortaya çıkan etki büyüklükleri değişmekte midir?

YÖNTEM

Örneklem

PISA 15 yaşındaki öğrencilerin ilgili konu alanlarındaki performanslarını ölçerken eksik test deseni kullanmaktadır (OECD, 2016b). Farklı kitapçıklar testin farklı sorularını içermektedir. Kitapçık 43, 45 ve 47 bir araya gelince tüm soruları içermektedir. Bu sebeple 43, 45, 47 numaralı kitapçıklara yanıt veren öğrenciler bu çalışmanın örneklemini oluşturmaktadır. Bu çalışmada 491 Türk, 1154 İngiliz ve 448 Amerikan öğrenci yer almaktadır.

Ölçme Aracı

PISA 2015 kapsamında öğrencilerin matematik performanslarının değerlendirmesi için toplam 69 madde kullanılmıştır. Her bir öğrenci yaklaşık 23 soru yanıtlamıştır. PISA matematik testindeki bu sorular ölçtükleri beceriler bakımından hiyerarşik bir yapıda hazırlanmıştır. En temel beceri olarak formüle etme, ardından uygulama ve en üst düzey düşünme süreci olarak yorumlama becerisi yer almaktadır (OECD, 2016b).

Veri Analizi

Bu çalışmada 3 farklı DMF belirleme yöntemi kullanılmıştır. Bu yöntemler logistik regresyon (LR), Mantel-Haenszel (MH) ve yapısal eşitlik modelidir (SEM). Her metot farklı hesaplama yöntemlerine dayalı olduğu için (Atalay Kabasakal, Gok, Kelecioglu & Arsan, 2012) daha tutarlı sonuçlar için en az 2 yöntemde farklılık gösteren maddeler DMF içeriyor olarak kabul edilmiştir. Logistik regresyon analizinde ilk adım olarak toplam puan, ikinci adım olarak toplam puan ve grup değişkeni, üçüncü adım olarak da toplam puan, grup değişkeni ve toplam puan ile grup değişkeninin etkileşimi modellere eklenmektedir. ΔR^2 0.035'den büyük ise DMF olduğuna karar verilmiştir (Jodoin and Gierl, 2001). SPSS programı kullanılarak bu analizler gerçekleştirilmiştir. Mantel-Haenszel metodunda ise grupların toplam puanına göre K adet 2x2 çapraz tablolar baz alınarak ki-kare değerleri hesaplanmaktadır. Daha sonra ilgili dönüşümler yapılarak MH D-DIF indeksi oluşturulmaktadır (Holland & Thayer, 1986). Bu değer 1'den büyük ise DMF olduğuna karar verilmektedir (Zieky, 1993). DIFAS 5.0 programı ile hesaplamalar yapılmıştır (Penfield, 2005). SEM ile DMF belirleme yönteminde ise doğrulayıcı faktör analizinde ilgili parametrelerin eşit olmaya zorlanması sonucunda elde edilen fit değerlerine büyük etkisi olan maddeler DMF içeren madde olarak belirlenmektedir (van de Vijver, 2017). Comparative fit index (CFI) ve Tucker Lewis index (TLI) değerleri arasındaki fark 0.010'dan büyük ise modifikasyon indeksleri incelenerek DMF içeren maddeler tespit edilir (Cheung and Rensvold, 2002). Bu analizde Mplus 7.4 programı kullanılmıştır (Muthen & Muthen, 2015).

SONUÇ VE TARTIŞMA

İç Tutarlılık

PISA 2015 matematik sınavı için Cronbach's alpha iç tutarlılık katsayıları kitapçık 43, 45 ve 47 için Türk öğrenciler için sırasıyla 0.78, 0.79, 0.76; İngiliz öğrenciler için 0.81, 0.84, 0.85; ve Amerikan öğrenciler için 0.80, 0.86, 0.86 olarak hesaplanmıştır. Bu değerler testin iyi düzeyde iç tutarlılığa sahip olduğunu göstermektedir (Cicchetti, 1994).

DMF sonuçları

Bu kısımda LR, MH ve SEM yöntemleri kullanılarak elde edilen DMF sonuçları verilmektedir.

LR yöntemi ile elde edilen sonuçlar Tablo 4'de verilmektedir. Türk ve İngiliz öğrenciler karşılaştırıldığında, 69 maddeden 10 tanesi (B43_11, B45_10, B45_13, B45_18, B47_1, B47_6, B47_7, B47_8, B47_9 ve B47_19), Türk ve Amerikan öğrenciler karşılaştırıldığında, 69 maddeden 14 tanesi (B43_11, B43_15, B43_16, B45_10, B45_11, B45_13, B45_18, B47_1, B47_6, B47_7, B47_9, B47_11, B47_14 ve B47_19) DMF içermektedir. MH yöntemi ile elde edilen sonuçlar Tablo 5'de verilmektedir. Türk ve İngiliz öğrenciler karşılaştırıldığında, 69 maddeden 10 tanesi (B43_11, B45_10, B45_13, B45_18, B47_1, B47_6, B47_7, B47_9, B47_10 ve B47_19) Türk ve Amerikan öğrenciler karşılaştırıldığında, 69 maddeden 10 tanesi (B43_11, B45_10, B45_13, B45_18, B47_1, B47_7, B47_9, B47_11, B47_14 ve B47_19) DMF içermektedir. SEM yöntemi ile elde edilen sonuçlar Tablo 6'da verilmektedir. Türk ve İngiliz öğrenciler karşılaştırıldığında, 69 maddeden 4 tanesi (B45_2, B45_10, B45_13, B45_18) Türk ve Amerikan öğrenciler karşılaştırıldığında, 69 maddeden 2 tanesi (B45_13 ve B47_9) DMF içermektedir.

En az iki yöntem tarafından DMF içerdiği görülen maddeler burada listelenmiştir. Türk ve İngiliz öğrenciler karşılaştırıldığında, 69 maddeden 9 tanesi (B43_11, B45_10, B45_13, B45_18, B47_1, B47_6, B47_7, B47_9 ve B47_19) her iki yönteme göre DMF içermektedir. Ayrıca, hangi maddelerin hangi grubun lehine çalıştığının raporlanması da önem taşımaktadır. Bu 9 maddeden 3 tanesi Türk öğrenciler lehine (B43_11, B45_10, B45_18, B47_6, B47_7, B47_9) 3 madde ise İngiliz öğrencilerin lehine çalışmaktadır (B45_13, B47_1, B47_19). Türk ve Amerikan öğrenciler karşılaştırıldığında, 69 maddeden 10 tanesi (B43_11, B45_10, B45_11, B45_13, B47_1, B47_7, B47_9, B47_11, B47_14 ve B47_19) her iki yönteme göre DMF içermektedir. Bu 10 maddeden 5 tanesi Türk öğrenciler lehine (B43_11, B45_10, B47_7, B47_9, B47_14) 4 madde ise Amerikan öğrencilerin lehine çalışmaktadır (B45_13, B47_1, B47_11, B47_19). Bir madde (B45_11) kısmen Türk öğrencilerin lehine, kısmen ise Amerikan öğrencilerin lehine çalışmaktadır. Türk-İngiliz ve Türk-Amerikan karşılaştırmaları benzer sonuçlar vermiştir.

Tablo 8 incelendiğinde, DMF gösteren tüm maddelerin açık uçlu sorular olduğu görülmektedir. Ayrıca, Türk öğrencilere hem İngiliz hem de Amerikalı öğrencilere göre avantaj sağlayan 7 sorunun 4 tanesinin en alt düşünme sürecini ölçen formüle etme düşünme süreci ile ilgili olduğu görülmektedir. Formüle etme becerisini ölçen hiçbir soru İngiliz ve Amerikan öğrencilerin lehine çalışmamaktadır.

DMF Sonuçları ve Etki Büyüklüğü

Türk öğrenciler ile İngiliz ve Amerikalı öğrenciler arasında başarı farkı bulunmaktadır. Gruplar arası farkları örneklemdeki kişi sayısından bağımsız olarak değerlendirebilmek için etki büyüklüğünü kullanmak iyi bir yöntemdir (Field, 2013). Tablo 9'da öğrenci grupları arasındaki farkın etki büyüklüğü tüm maddeler kullanılarak ve DMF gösteren maddeler çıkarıldığında hesaplanmıştır. Türk ve İngiliz öğrenciler arasında başlangıçta .51 ile .93 arasında değişen etki büyüklüğü hesaplanmıştır. DMF içeren maddeler çıkarıldığında ise bir değişiklik gözlenmemiştir. Aynı şekilde Türk ve Amerikalı öğrenciler arasında .28 ile .85 arasında değişen etki büyüklüğü gözlenmiştir. DMF içeren maddeler çıkarıldığında yine farkın değişmediği görülmüştür.

Tartışma

Bu çalışma Türk öğrencilerin PISA 2015 matematik testinden çok düşük alma nedenlerinden birisi olabilecek olan DMF içeren maddeleri incelemesi bakımından oldukça önemlidir. Araştırmada önceki bölümlerde belirtildiği gibi DMF içeren maddeler tespit edilmiştir. Ancak, bu maddeler sadece Türk öğrencilerin aleyhinde çalışmamaktadır. DMF içeren maddelerin bir kısmı Türk öğrencilerin lehine çalışmaktadır. Ek olarak, etki büyüklükleri karşılaştırıldığında DMF içeren maddelerin toplam puanlarda herhangi bir gruba bir avantaj sağladığına dair kanıt bulunmamaktadır.

Puanlardaki düşüş için farklı nedenlere odaklanmak gerekmektedir. Türk öğrencilerin PISA 2015 ortalama matematik puanlarında neden düşüş yaşadıklarını tespit etmek için yıllar içerisinde seçilen örneklemlerin karşılaştırılabilirliği, sınavın kağıt kalem formatı yerine artık bilgisayar ortamında uygulanması ve ülke bazındaki eğitim sistemi, öğretim programları ve eğitim politikalarında yaşanan değişimler gibi farklı değişkenleri de incelemek gerekmektedir.

PISA'daki sorular yayınlanmadığı için DMF içeren maddelerin yanlılık gösterip göstermediğine dair uzman incelemesi yaptırılamamıştır. Ancak, soruların özellikleri incelendiğinde bazı önemli ipuçları elde edilmiştir. DMF içeren tüm maddelerin açık uçlu sorulardan oluşması bu soruların puanlanma süreçlerinin yeniden gözden geçirilmesi gerektiğini göstermektedir. Bu puanlama sırasında maddeler DMF içeriyor hale gelmiş olabilir. Diğer bir bulgu da, Türk öğrencilerin lehine çalışan maddelerin çoğunun en alt düzey düşünme sürecini içeren maddeler olmasıdır. Bu tip maddelerin hiçbiri İngilizce konuşan öğrencilere DMF göstermemiştir. Türkiye'deki eğitim genel olarak çok soru çözmeye dayandığı için, öğrenciler temel becerileri geliştirmiş ve bu tip sorularla daha fazla karşılaşmış olabilir (Arikan, van de Vijver & Yagmur, 2016; Doganay & Bal, 2010; Temur, 2012). Bu durum da bu tip maddelerin Türk öğrenciler lehine DMF göstermiş olabileceği anlamına gelmektedir. Son olarak, kullanılan DMF belirleme yöntemleri karşılaştırıldığında logistik regresyon ve Mantel-Haenszel yöntemlerinin yapısal eşitlik modeline göre birbirine daha yakın sonuçlar verdiği görülmüştür.

Appendix A.

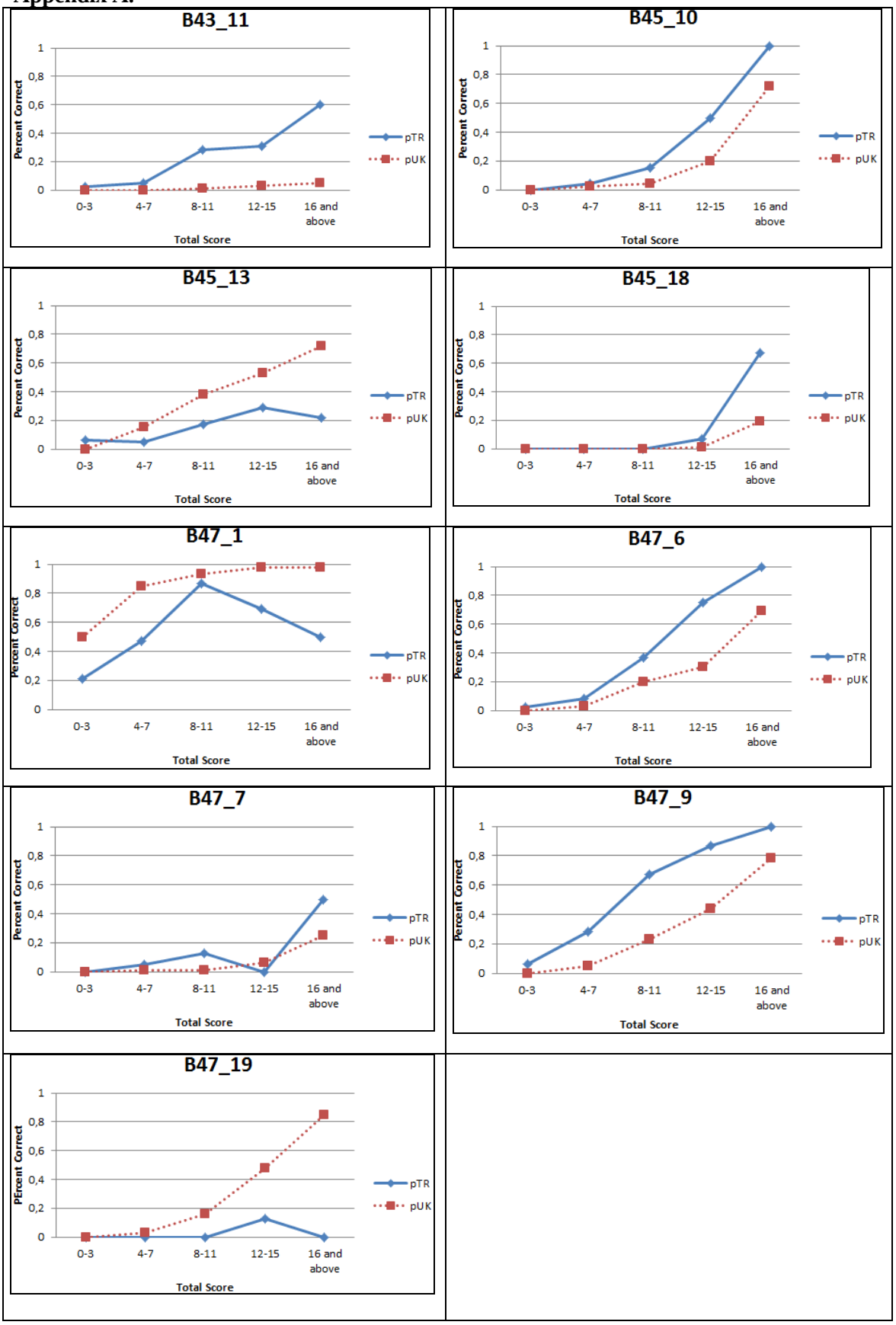


Figure 1. Graphical Representation of DIF Items for TR and UK Students

Appendix B.

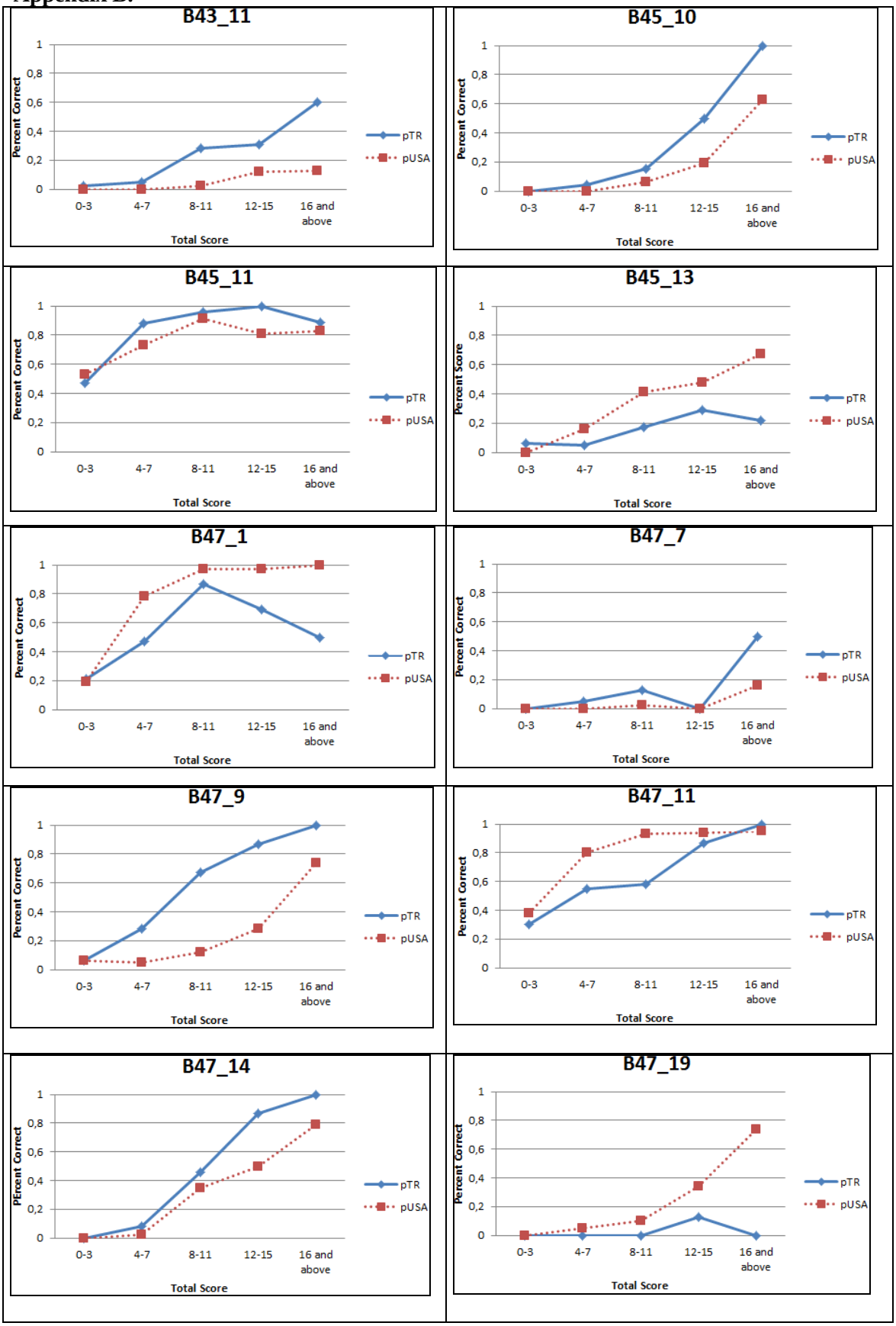


Figure 2. Graphical Representation of DIF Items for TR and USA Students

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(1);68-79

Opinions on the Impacts of the TEOG System from Teachers Whose Courses are not Included in the TEOG Exam

Seher ULUTAŞ** R. Nükhet ÇIKRIKÇI ***

Abstract
The object of this study is to determine of teachers' opinions on the impacts of exam for the Transition from Basic Education to Secondary Education TEOG system on their teaching activities in visual arts, technology and design, music and physical education courses that are not covered by the TEOG exam. This research was conducted as a survey study and the study group was determined by the purpose of sampling strategy, data analysis plan and easy accessibility approach. This research was carried out with 35 teachers who teach visual arts, physical education, technology and design, and music courses in public schools in Ankara. Teachers' opinions were obtained with a form containing five open-ended questions and demographic and occupational characteristics of the teachers. A descriptive analysis approach was applied to the written opinions of the teachers. As a result of this research according to the opinions of teachers; visual arts, technology and design, music and physical education courses which are not included by the TEOG exams were considered to be insignificant by students, school administrators and parents. In addition, teachers stated that the TEOG system affected their teaching and evaluation activities negatively in the classroom, and the teachers were unable to evaluate their students objectively due to the TEOG system, and students, administrators and parents expected or demanded the teacher to give higher grades. Because of these situations, the relations between the teachers, students, school administrators and parents were affected negatively.

Key Words: TEOG system, common exam, teachers' opinions.

INTRODUCTION

One of the most important wealth of a country is manpower; i.e., human capital. It is accepted all over the world that the quality of manpower affects the development and prosperity of countries, and what determines the quality of manpower is the quality of education. For this reason, both politicians and parents desire that children and young people receive education in quality schools as part of the process of developing skilled future generations. The level of the education in a school is affected by various factors, such as the location of the school, the characteristics of the students, the support of the families, the socio-economic characteristics of the families, and the quality of the teachers. Since these factors are not at the same level in each school, the standard of education and level of students' achievement are not the same in every school. National and international evaluation studies in Turkey have shown that there are significant achievement differences between school types in Turkey (MEB, 2010; MEB, 2015; MEB, 2016b; ERG, 2017). In addition, there are significant differences between high schools in terms of university admission rates in Turkey (MEB, 2012). For this reason, parents want their children to be educated in secondary schools that have a record of higher student success. However, selecting which children attend a public school is based on the result of the central common exams in Turkey.

In Turkey from 1997 onwards, the transition from primary education to secondary education was carried out by the Ministry of National Education (MoNE) every year under the name of Exam for

* An early draft of this paper was presented in "IV International Eurasion Educational Research Congress 2017", Pamukkale University Denizli

** Dr., Ministry of National Education, Ankara, Turkey, e-mail: seherulutas@yahoo.com.tr, ORCID ID: orcid.org/0000-0002-4124-2140

*** Prof. Dr. Istanbul Aydin University, Faculty of Arts and Humanities, Istanbul, Turkey, e-mail: nukhet0405@gmail.com, ORCID ID: orcid.org/0000-0001-8853-4733

To cite this article:
Ulutaş, S., & Çıkrıkçı, N., R. (2019). opinions on the impacts of the TEOG system from teachers whose courses are not included in the TEOG exam. *Journal of Measurement and Evaluation in Education and Psychology, 10*(1), 68-79. DOI: 10.21031/epod.342086

Received: 06.10.2017
Accepted:24.10.2018

Secondary Education (ESE/OKS) for eighth-grade students. In 2008, OKS was abolished and SBS (Exam for Proficiency Level-EPL) was implemented centrally by MoNE at the end of the sixth, seventh and eighth grades, limited to the acquisition/objectives in the curricula of that year (MEB, 2012, pp. 1-2).

SBS has been gradually eliminated since 2010 and the system of Transition from Basic Education to Secondary Education (TEOG) began to be implemented in the 2013–2014 academic year. In this system, placement in secondary education is based on the achievement scores of the student at the end of the year and the scores obtained from central common exams in the eighth grade of schools. In this system, the central common exams which comprise Turkish, mathematics, science, religious culture and ethics, the history of the Turkish Republic, revolution and Ataturkism, and foreign language courses are given once in each semester. The TEOG central exam has certain common features, such as the opportunity for make-up exams, being implemented in two school days, covering the subjects studied until the exam day, duration of the exams being similar to that of an ordinary exam, false answers not affecting the scores of correct answers, and each student taking the exams in their own schools. According to MoNE, the aim of the TEOG system is to strengthen the student, teacher and school relationship, make teachers and school's role more effective in the educational process, ensure simultaneous implementation of national curricula across the country, and reduce students' test anxiety. Other objectives include increasing teacher's professional performance, reducing the need for out-of-school education institutions, monitoring and evaluating implementation of curriculum and student achievements objectively, and remove the negative effects of a single session test MEB, 2013; MEB, 2014).

In the TEOG system, the calculation of the points for placement in secondary education institutions is as follows: The weighted common exam scores (AOSP) obtained from the common exams held every semester are calculated by multiplying the score of each course by four coefficients for Turkish, mathematics, and science courses, and two coefficients for the history of the Turkish Republic, revolution and Ataturkism, foreign language, and religious culture and ethics courses. The total possible score is 700 points. In the calculation of the points for secondary school placement, student's achievement scores at the end of the sixth, seventh and eighth grades and AOSP of the eighth grade are collected, and the total score obtained is divided into two to obtain the "basis score for placement in secondary education institutions". This score is based on a possible total of 500 points (MEB, 2013, MEB, 2014, MEB, 2016a).

In Turkey, there are several studies about this exam system. According to the results of one of these studies, teachers have positive and negative opinions about the TEOG system. The positive aspects of the system are that students take their exams in their own schools, the exam is applied in two semesters, a compensatory exam is offered, and correction formula is not used in scoring. In addition, in this research, teachers stated that the standard of the school will increase because of the importance of the assessment of the courses in the TEOG system. Teachers also stated that this exam positively affected the student's interest and motivation toward courses and reduced student absenteeism. The negative aspects of the system are that it does not reduce the necessity tutoring and attending private cram courses and that the reliability of exams are inadequate (Şad & Şahiner, 2016). Another study that surveyed the opinions of teachers and students found that the TEOG system was objectionable in terms of equality of opportunity in education, but not using correction formula against the chance factor in the calculation of test scores and the test being conducted in students' own schools increased their motivation. It was also found that the year-end achievement score had positive impact on school success because of participation in the evaluation in the TEOG system. It was also determined that the TEOG system affecting the achievement scores of the year-end in secondary education entrance was the reason for the increase in interest in lessons except for the six basic courses (Özkan & Özdemir, 2014). In another study about the TEOG system, the positive aspects expressed by the students were application of the central common exams in a modular way, having a break between the exams, the test consisting of multiple choice questions, and wrong answers not negating correct answers. However, the students generally regarded central exams as stressful (Öztürk & Aksoy, 2014). In research undertaken with science teachers (Atila & Özeken, 2015), the participants stated that the

common test of the TEOG system affected the functioning of the teaching, and they tended to focus on solutions for test questions. A significant number of the teachers stated that this situation caused the problem of objectivity, and thus they hesitated about grading their students. Some of the teacher participants stated that the school administrators wanted the students to be seen as successful in their institutions, and since the parents knew that the grades given by the teachers affected the secondary school placement, they wanted to meet with the teachers to discuss the grades, which put pressure on the teachers. Demirtaşlı (2016) gathered the opinions of primary school teachers about the effects of the TEOG system on teaching and evaluation activities). The teachers explained that the preparation process for the TEOG system common exams had a negative effect on the teaching and evaluation activities in the classroom because the students wanted to engage in test practice in most of the class time. Furthermore, the students' anxiety about the test and spending time on preparing for it meant that they did not participate in social, cultural and sports activities in the school.

It is seen that central exams, such as the TEOG system have positive and negative effects on students and educational activities in most countries. In a study conducted in the United States, more than 80% of the students stated that they needed to increase their studying time due to this type of exams, and this reduced the time spent on extracurricular activities. It was found that nearly half of the students stated that there was a decrease in extracurricular activities. In addition, over 80% of the students stated that they felt depressed, anxious and embarrassed after the exams (Cornell, Krosnick & Chang, 2006). In a study concerning all primary and secondary school teachers in the city center of Virginia on the effects of a central exam on learning and teaching, more than 80% of teachers believed that teaching had changed due to the implementation of these exams. Teachers expressed that they ignored lessons that were not within the scope of the test and focused more on lessons within the scope of testing. More than one-third of these teachers stated that they were concerned that the excessive focus on the lessons in the exam limited or eliminated the teaching time for lessons or activities that were not assessed. The responses of more than 50% of teachers who participated in the survey showed that the main purpose of the teaching was to improve students' performance in these tests and complained about "teaching to the test" (Sullivan, 2006). According to interviews with participating teachers in a survey of the National Assessment Program (NAPLAN), a centralized test in Australia, it was stated that the effects of the test on the students and curriculum were not entirely negative. The test supported the increasing focus on students' high-level education, reaching full aims in the curriculum, and critical reading-writing and arithmetic skills to become participant citizens after completing secondary school. In addition, the results of the test help teachers improved their professional development and teaching practices. However, concerns were reported by teachers that the test had some negative effects on curricula and students. The teachers stated that the test intensified the curriculum due to the need to study for the test and engage in test-solving exercises, reducing the time allocated for the lessons other than mathematics and literacy. Furthermore, teachers commented that because of the test, some of the students felt stressed or sick, were tearful before the exam or afraid of their parents' reactions, and had insomnia (Dulfer, Polesel, & Rice, 2012). According to a study conducted by the New York State Education Department (2004), the positive effects of high-stakes exams on students and teachers were that they provided students with clear information about their skills and encouraged them to study harder. However, this type of test had negative effects, such as students being disappointed with their results, discouraging them from studying, making them more competitive, and lowering their opinion of the value of grades and evaluation.

Thus, central common examinations have positive and negative effects on students and education. In Turkey, research was carried out concerning the impact of the TEOG system and common exams on education and teaching; however, there is no research on how the test affects the teachers ' classroom practices and their relations with the stakeholders regarding education.

The aim of this study is to determine the effects of the TEOG system on teachers 'practices in classroom and teachers' relationships with students and other stakeholders in visual arts, technology and design, music and physical education lessons that are not covered by the TEOG exams. This study is important in terms of providing an understanding of the impacts of the TEOG central exam system which is applied in the transition to high schools in Turkey. Such high-stake exams generate exam

based teaching and learning activities at schools and manipulate teachers' relationships with their students in whole lessons. The study is also considered important in terms of identifying problems in these dimensions and the planning of educational policies that might resolve them.

Purpose of the Study

The general aim of the research is to gather the opinions of teachers who teach visual arts, technology and design, music and physical education lessons that are not covered by the common exams in the TEOG system, and to determine how this affects their relations with the students and other stakeholders, and education and training practice and evaluation activities in the schools. For this purpose, answers to the following questions were sought:

1. How does the TEOG system affect the communication between teachers and students?
2. How does the TEOG system affect classroom teaching activities?
3. How does the TEOG system affect the activities of evaluating student achievement?
4. How does the TEOG system affect the teachers' relationships with school administrators and their assistants?
5. How does the TEOG system affect the teachers' relationships with parents?

METHOD

This research was conducted as a survey study, in which the purpose is to collect data to determine a specific feature of a group and create a picture of the existing situation in the field of research (Büyüköztürk et al. 2013).

Study Group

In the research, a typical case sampling method was chosen from the purposeful sampling strategy. This method requires the collection of information from the sample by determining a situation that is typical of a number of situations in the universe (Büyüköztürk et al. 2013). Fowler (2009) recommends that the sample size should be determined in accordance with the analysis plan of the studying data (as cited in Creswell, 2014, pp. 159). In this study, the group size was determined in accordance with the data analysis plan and easy accessibility. This research was carried out with 35 teachers who teach visual arts, physical education, technology and design, and music courses in seven public schools in Ankara. Table 1 shows the distribution of the participant teachers according to the branch, gender, service year, and grade that they teach.

Table 1. Information on the Study Group

		Number of Teachers
Branches	Music	5
	Visual arts	6
	Physical education,	9
	Technology and design	15
Gender	Woman	23
	Man	12
Service year	Between 1-5 years	4
	Between 6-10 years	3
	Between 11- 15 years	7
	Between 16-20 years	11
	More than 20 years	10
Grade that they teach	5-8	21
	7-8	14
	Total	35

Instrument and Data Collection

In this study, a semi-structured survey form with five open-ended questions was used. This instrument was to allow teachers to express their opinions freely on how the TEOG system affects their teaching and evaluation processes and relations with stakeholders in their courses. Teachers were asked to respond in writing to the questions in the survey form created by the researchers. With the survey form, no personally identifiable information other than their demographic and professional characteristics was received. Researchers' contact information was also given to them to learn the results of the study. The following open-ended questions were asked to the teachers who were included in the purpose of the study:

1. How does TEOG affect your general communication with students?
2. How does TEOG affect your classroom teaching activities?
3. How does TEOG affect your assessment activities?
4. How does TEOG affect your relationships with administrators and their assistants?
5. How does TEOG affect your relationships with the parents?

Data Analysis

A descriptive analysis approach was applied to the written opinions of the teachers. In this approach, the data are summarized and interpreted according to the previously determined themes. The purpose of this analysis is to present the findings to the readers in an organized and interpreted way. In descriptive analysis, direct quotations are often given in order to reflect the views of individuals in a striking way (Yıldırım & Şimşek, 2005). In the study, firstly, according to the research questions, each of the 35 forms containing the teachers' responses was examined one by one. Essential opinions/themes and keywords in the opinions were extracted from the forms. Secondly, the selected words for each question and corresponding to the sub-themes were identified. This study was examined under five themes and 22 sub-themes. Expressions describing the sub-themes specified in all response papers were counted. As a result, the frequencies of each question and sub-themes were calculated according to the themes. In order to ensure reliability in coding, the consistency between the coding for each question and the subcategory was examined by two experts in addition to two measurement experts. The level of reliability was calculated based on the total consistency and inconsistency ratio of two experts regarding the codes (Tavşancıl & Aslan, 2001). It was determined that the reliability levels of the evaluators calculated for each question and sub-theme ranged between .100 and .33 in this study. Sub-theme expressions with low reliability were reviewed and changed, and returned to evaluators, after which the final level of reliability was found to be .80 for the whole themes and sub-themes. The placement consistency of opinions in the sub-themes should be between 70% and 100% for each question and sub-theme. There should be at least 70% consistency between the coders (Hall & Houten 1983). In this regard, it was understood that an acceptable level of consistency was achieved; thus, the frequencies of responses in each question and sub-theme were calculated according to the themes.

RESULTS

In this section, the findings obtained in relation to the objectives of the research are summarized and interpreted.

The Impact of the TEOG System on Teacher and Student Relations

Table 2 shows the frequency distributions of teachers' responses to the first research question: "How does the TEOG system affect the communication between teachers and students?"

Table 2. Distribution of Teachers' Views According to Sub-Themes

Sub-Themes	Frequency
Establishing a note-based communication with students	4
Conflict with students in the course process	2
Conflicts due to students' stress / fear / anxiety	6
Negative impact on relationships in general	8
No effect	7

Table 2 shows that one-third of the respondent teachers (f = 8) reported that the TEOG system generally had a negative impact on their communication with their students, while the other third of the respondent teachers indicated that they were in conflict with students due to the process of the course, lack of discipline, and test anxiety. However, some of the teachers (f = 7) stated that the test had no effect on their communication with the students. Some teachers (f = 4) also pointed out that their communication with students was focused on the grade and students only participated in this course to obtain high grades. Examples of the teachers' views on the effects of the TEOG system on teachers' communication with students are given below:

K2B "Since our lesson does not include the TEOG group, the students evaluate it as not important."

K6S "It affects our communication with the students negatively. Apart from TEOG, everything seems to be futile, and I'm having trouble controlling the class."

K3T "Communication with students always takes place on the bases of grades."

K12S "Students perceive the TEOG test as a vital event. Students are not interested in our courses. "

The Impact of the TEOG System on In-Class Teaching Activities

Table 3 shows the frequency distributions of teachers' responses to the second question of the research, "How does the TEOG system affect the activities in the classroom?".

Table 3. Distribution of Teachers' Views According to Sub-Themes

Sub-Themes	Frequency
Students do not care about the lesson	20
Students do not show interest in the lesson	21
Students do not undertake the duties and responsibilities related to the lesson	12
Students want to take practice tests for TEOG during class hours	12

Table 3 shows that the majority of the teachers (f = 20) stated that the students did not care about the course and this affected the education and training activities negatively. Most of the teachers (f = 21) stated that because of the common exams, the students did not show interest in the activities of these courses and they did not participate in these activities. In addition, some of these teachers (f = 12) expressed that students did not want to undertake their duties and responsibilities related to the courses, and they only wanted to study for the exam in the courses. The teachers also commented that the students always wanted to take practice tests. Some of the teachers' views on the effectiveness of the TEOG system in in-class teaching activities are as follows:

K6S "I cannot carry out effective teaching in the classroom. It takes a short time to draw the attention of students. Students always tend to study for TEOG and answer the questions in practice tests. "

K9B "Because they want to be successful in the exam, the children want permission to study for the TEOG exam in this lesson."

K10T "Students do not care about to other courses apart from of TEOG exam. For this reason, in terms of the students the purpose of the courses cannot be realized exactly.

K21T "Children especially want to solve the test in our courses."

The Impact of the TEOG System on the Evaluation Activities

Table 4 shows the frequency distribution of teachers' answers on the third question of the research, "How does the TEOG system affect the activities of evaluating students' success?".

Table 4. Distribution of Teachers' Views According to Sub-Themes

Sub-Themes	Frequency
Students want to get high grades.	6
Administrators interfere with teachers' grades.	10
Parents expect and request high grades from teachers	6
Teachers do not make fair evaluations / They give inflated grades	12
No effect	5

Table 4 shows that the majority of participant teachers stated that students, parents, and administrators (f = 22) expected or demanded high grades from the teachers in these courses. A significant number of teachers (f = 13) also stated that they had to give an inflated grade to their students and sometimes they did not make a fair evaluation because of the TEOG system. Conversely, some teachers (f = 5) stated that the TEOG system had no effect on the evaluation activities. Some of the teacher's views about the effects of the TEOG system on assessment activities are as follows:

K1S "Students want to get high grades. They can ask for it as natural rights. "

K2B "There is a pressure on us …. We cannot give low grades. We give high grades to not lower the grade average."

K8T "School administrators say that for all of the courses, which are not covered by TEOG exam, grades should be high. That's why they interfere with the teachers' grades and say that all grades should be higher."

K9B "We have to raise the grades so that students who are good at other courses do not fail."

The Impact of the TEOG System on Teacher's Relationship with School Administrators

Table 5 shows the frequency distributions of teachers' responses to the fourth question of the research, "How does the TEOG system affect the relationship of teachers with school administrators and assistant managers?".

Table 5. Distribution of Teachers' Views According to Sub-Themes

Sub-Themes	Frequency
There is conflict due to the pressure to give high grades	9
These courses are not considered / unnecessary	13
No negative effect	12

Table 5 shows that a significant number of participant teachers (f = 12) stated that the TEOG system did not adversely affect their relations with school administrators and manager assistants, others teachers (f = 13) stated that school administrators and their assistants cared about the courses which were within the scope of the common exam, that they did not care about the courses outside the scope

of the common exam. In addition, they asked their teachers to implement practice TEOG tests in these courses. Some of the teachers (f = 9) also mentioned that administrators asked them to give high grades to their students, which also caused tension and conflict between them. Some of the teachers' views on the effects of the TEOG system on school administrators and manager assistants and teacher relations are as follows:

K6S "The test (TEOG) negatively affects our relationships. When parents complain about the grades, the administrators tell us to raise the grades. Because there is no one regards to our course at our school, there is permanent stress in such events."

K10T "Generally, like students, administrators do not give importance to the courses outside of the scope of TEOG. They do not provide any opportunity to achieve the purpose of the lesson; on the contrary, they generally put obstacles in the way."

K11B "The relationships are bad because our course is seen as unessential and worthless. For students, only mathematics, English etc. are important."

The Impact of the TEOG System on Teachers' Relationship with Parents

Table 6 shows the frequency distributions of the teachers' responses to the fifth question of the survey, "How does the TEOG system affect teachers' involvement with parents?".

Table 6. Distribution of Teachers' Views According to Sub-Themes

Sub-Themes	Frequency
Parents do not care about the courses	13
Conflict due to parents' request for high grades	6
Parents want students not to participate in activities or be assigned homework	7
Exam-oriented training expectations	4
No effect	7

Table 6 shows that a significant number of the participant teachers (f = 13) stated that the parents did not give importance to these courses. Some of the teachers (f = 7) also mentioned that the parents did not ask their children to participate in activities in these courses and asked the teachers not to give assignments related to these courses. However, many teachers commented that the TEOG system had no effect on the relationship with the parents. Some teachers (f = 6) also addressed the presence of conflict between parents and teachers due to the expectation of test-oriented education and high grades for their children. Some of the teachers' statements about the effects of the TEOG system on the relationship with the parents are as follows:

K2B "The families are very worried. They do not allow students to participate in extracurricular activities."

K6 S "Parents complain about homework, research, etc. in the visual arts course because they think that their children cannot study for the TEOG exam. That's why they say that assignments like this should not be given."

K19 T "The parents does not care about the courses that are not covered by the TEOG exam, and they ask the teachers –to give the student 100 points."

K8 T "When the time for school report arrives, we are in conflict with the parents because of our grades. We are trying to convince them that our courses are important. But they think that if a lesson is not included in the TEOG exam, it is not important or necessary."

DISCUSSION and CONCLUSION

The findings of this research which aimed to determine the effects of TEOG system on teaching and learning processes in visual arts, technology and design, music and physical education courses and teachers' relations with students and other stakeholders were evaluated in terms of how the TEOG application was related to the objectives of MoNE. For MoNE, one of the aims of TEOG is to strengthen the relationship between student, teacher and school, and to make the role of teachers and school more effective in the education process. According to the findings of this research, the TEOG system, in contrast to the purpose of MoNE, has made the situation "more tense and negative" instead of "strengthening student, teacher and school relations". The exam creates conflict between the teachers and students due to the teaching process and the anxiety and stress experienced by the students. Thus, this finding does not coincide with the aim of MoNE, which states that "the central common examination will reduce the test anxiety by spreading it to the process". Similar results regarding the TEOG system were obtained by Öztürk and Aksoy (2014) and Demirtaşlı (2016). The results were similar in other studies, which also showed that central examinations caused stress, anxiety and insomnia in students, and students felt sick, depressed, anxious and embarrassed upon receiving the test results (Cornell, Krosnick & Chang, 2006; Dulfer, Polesel, & Rice, 2012).

According to the results of the research, students did not care about these courses, pay attention, or fulfill their duties or responsibilities related to these courses, and this exam had a negative impact on educational activities. These results contradict Özkan and Özdemir (2014)'s conclusion that "TEOG will lead to an increase in students' interest in the courses outside the scope of the exams because of the effect of year-end scores on the TEOG process." These findings also contradict the result of the research of Şad and Şahiner (2016), in which the participant teachers stated that the importance of the courses and teacher evaluation would increase because of the TEOG tests which affected assessment in the classroom. In addition, the teachers thought that the student's interest and motivation related to the course would be positively affected and student absenteeism could be reduced." A similar result was reached by Sullivan (2006). In that research, the teachers perceived that high-stake testing had changed the focus of instruction toward subjects that were tested by high-stake exam. Overall, 52% of the teachers perceived that this increased focus on the test and decreased instructional time for other non-tested lessons reduced student access to non-core content, such as art, music, physical education and computer/technology.

According to the teachers' opinions in the current study, one of the results obtained related to the course process is that "students want to do practice tests for the common tests and study on the courses within the scope of the common test.". This outcome coincided with the findings of Atila and Özeken (2015) in that "a significant part of the science teachers stated that TEOG exams affected the way the course was given and engaged students in taking practice tests during the class hour". The research of Zorlu and Zorlu (2015) concurred, stating that "the students in the seventh and eighth grades asked the teacher to teach the courses for the exams.". These results from different research are supported by the conclusion reached by Buyruk (2014) that "central exams led to a more exam-centered form of education in schools."

For MoNE, one of the other aims of the TEOG system is to monitor and evaluate student achievement objectively. According to the findings obtained from the research, a significant number of the teachers were unable to measure and evaluate their students objectively due to the TEOG exams, and they were forced to give inflated grades to their students. In addition, more than half of the participants in the current study stated that students, administrators and parents expected or demanded the teacher to give higher grades. These results contradict the aim of the TEOG system, which is "objectively monitoring and evaluating student achievements". In particular, it is not possible to determine whether the evaluation is appropriate for this purpose in the courses covered by the exams. This finding similar to the research of Atila and Özeken (2015), who found that due to the TEOG tests, the teachers hesitated to give grades. Almost all science teachers give their students nearly the same performance and project grades with the scores obtained from the TEOG exams. However, this finding contradicts with the results of the study by Şad and Şahiner (2016). The authors determined that the results of the TEOG

exam will increase the importance of the courses and teacher assessment because of the impact on the student's achievement grade in classroom.

Although one-third of the teachers who participated in the study stated that the TEOG system had no negative impact on their relations with the administrators, more than one-third of the teachers suggested that the administrators did not care about the courses that were the subject of the research or considered the courses unnecessary. Similarly, about one-third of the teachers stated that they received requests from the administrators to give higher grades to students, which negatively affected their relations with the administrators. These findings are similar to the results of Atila and Özeken (2015)'s research in that some of the science teachers said that because administrators asked the teachers to increase their students' achievement in their own institutions, this caused pressure on the teachers.

According to the findings from the research, some teachers who participated in the study reported that central examinations conducted within the scope of the TEOG had no impact on their relations with parents, but more than one-third of the participant teachers believed that parents did not care about their courses. So, the teachers stated that the test system had a negative impact on their relationship with parents. In addition, some teachers reported that the parents requested them to give higher grades for children, which led to conflict between them. This finding is in agreement with the research of Atilla and Özeken (2015), who reported on the awareness of parents concerning the importance of grades given by teachers for the placement of their children in secondary education; thus, they wanted to meet the teachers to discuss students' grades, and this caused pressure on the teachers. Moreover, in this study, some teachers stated that parents did not want their children to participate in the internal and external activities in school because of the exam. They also did not want the teachers to give any homework in these courses. As a result of research conducted by MoNE (2010), it was found that the SBS preparation process led children to be significantly distanced from sports, social and cultural activities. Similar results were obtained in a study conducted by Cornell, Krosnick and Chang (2006), who showed that about half the students reported a decrease in extracurricular activities due to the central examination.

In conclusion, students, school administrators and parents do not appear to care about the courses that are not within the scope of the TEOG exams, and this adversely affects the relations between the teachers who present these courses and the students, administrators, and parents. In addition, such high-stake exams have a negative impact on the educational practices and evaluation activities in these courses. However, as stated in the Turkish National Education (MEB, 1973), it is aimed that young people who will determine the future of our country will be educated to be "constructive, creative and productive people who have a balanced and healthy personality and character from the care of body, mind, morality, spirit and emotion; the power of free and scientific knowledge. In addition, the 10th Development Plan (Ministry of Development, 2013, s. 32) and higher-level policy documents, such as the MoNE Strategic Plan (Strategy Development Presidency, 2015) stated that in order to increase the students' mental and physical development and their skills in all educational levels, social, artistic, sportive and cultural activities should be given more attention, and the participation rate of the students should be increased. To achieve these objectives, courses, such as physical education, music and visual arts have an important role, as well as the academic content applied in educational institutions. However, it will be difficult to achieve these objectives if these courses are not properly taught in schools. For this reason, to exclude these courses from the TEOG exams will make it difficult to educate individuals in the qualifications that are planned in the educational institutions. In order to achieve the objectives of the Turkish National Education and higher-level policy documents, emphasis should be given on the courses mentioned above, as well as the academic content of courses in schools. For this reason, weight of student's grade point average based on primary academic courses and other courses should be increased in the computation of the TEOG exam score. In this way, school achievement can be relatively important to entrance secondary education.

As in all research, there are some limitations to this study. The main limitation is that the results were obtained from a small group. In subsequent studies, it would be useful to conduct a comprehensive survey of more structured materials with the views of more teachers working in schools in different

regions and characteristics. Thus, comparing teachers' opinions according to variables, such as duration of service, class and school size, as well as eliciting students' views in a similar way will contribute to the field.

REFERENCES

Atila, M. E. & Özeken, Ö. F. (2015). Temel eğitimden ortaöğretime geçiş sınavı: Fen bilimleri öğretmenleri ne düşünüyor? *Ondokuz Mayis University Journal of Faculty of Education*, *34*(1), 124-140. DOI: 10.7822/omuefd.34.1.7

Buyruk, H. (2014). Öğretmen performansının göstergesi olarak merkezi sınavlar ve eğitimde performans değerlendirme. *Trakya University Journal of Education, 4*(2), 28-42.

Büyüköztürk, Ş., Çakmak, E. K., Akgün, Ö. E., Karadeniz, Ş. & Demirel, F. (2013). *Bilimsel araştırma yöntemleri*. Ankara: Pegem Akademi.

Cornell, D. G., Krosnick, J. A., & Chang, L. (2006). Student reactions to being wrongly informed of failing a high-stakes test: The case of the Minnesota basic standards test. *Educational Policy, 20*(5), 718-751.

Creswell, J. W. (2014). *Nitel, nicel araştırma deseni ve karma yöntem yaklaşımları* (S. B. Demir, Translation). Ankara: Eğiten Kitap.

Demirtaşlı, N. (2016, September). *Hesap verebilir eğitim sisteminde bir paydaş olarak öğretmenlerin TEOG (temel eğitimden orta öğretime geçiş) uygulamasına yönelik görüşleri*. Paper presented at the 5th International Congress on Measurement and Evaluation in Education and Psychology, Antalya, Turkey.

Dulfer, N., Polesel, J. & Rice, S. (2012). *The Experience of Education: The impacts of high stakes testing on school students and their families*. An Educator's Perspective, Sydney, Whitlam Institute

ERG. (2017). *PISA 2015: Genel bulgular ve eğilimler*. Retrieved from http://www.egitimreformugirisimi.org/wp-content/uploads/2017

Hall, R. V., & Houten, R. V. (1983). *Managing behavior, behavior modification: The measurement of behavior*. Austin, Texas: Pro-ed.

Kalkınma Bakanlığı. (2013). *Onuncu kalkınma planı 2014-2018*. Ankara: TC. Kalkınma Bakanlığı.

Milli Eğitim Bakanlığı. (1973). *Milli eğitim temel kanunu*. Retrieved from http://www.mevzuat.gov.tr/MevzuatMetin/1.5.1739.pdf.

Milli Eğitim Bakanlığı EARGED. (2010). *Seviye belirleme sınavının değerlendirilmesi*. Ankara: Millî Eğitim Bakanlığı.

Milli Eğitim Bakanlığı. (2010). *PISA 2006 projesi ulusal nihai rapor*. Ankara: Millî Eğitim Bakanlığı.

Milli Eğitim Bakanlığı. (2011). *TIMSS 2007 ulusal matematik ve fen raporu 8. sınıflar.* Ankara: Millî Eğitim Bakanlığı.

Milli Eğitim Bakanlığı. (2012). *İlköğretimden ortaöğretime ortaöğretimden yükseköğretime geçiş analizi*. Ankara: Millî Eğitim Bakanlığı.

Milli Eğitim Bakanlığı. (2013). *Temel eğitimden ortaöğretime geçiş sistemi*. Retrieved from http://www.meb.gov.tr/duyurular/duyurular2013/bigb/tegitimdenoogretimegecis/sunum.pdf.

Milli Eğitim Bakanlığı. (2014). *OGES sunum.* Retrieved from http://oges.meb.gov.tr/docs2104/sunum.pdf

Milli Eğitim Bakanlığı. (2015). *PISA 2012 araştırması ulusal nihai rapor.* Ankara: Millî Eğitim Bakanlığı.

Milli Eğitim Bakanlığı Strateji Geliştirme Başkanlığı. (2015). *Millî Eğitim Bakanlığı 2015–2019 stratejik planı.* Retrieved from http://sgb.meb.gov.tr/www/mill-egitim-bakanligi-2015-2019.

Milli Eğitim Bakanlığı. (2016a). *2016-2017 öğretim yılı ortak sınavlar e-kılavuzu*. Retrieved from https://oges.meb.gov.tr/meb_iys_dosyalar.

Milli Eğitim Bakanlığı. (2016b). *PISA 2015 ulusal raporu.* Ankara: Millî Eğitim Bakanlığı.

New York State Education Department. (2004). *The impact of high-stakes exams on students and teachers.* Retrieved from http://www.oms.nysed.gov/faru/TheImpactofHighStakesExams_files/ The_Impact_of_High-Stakes_Exams.pdf.

Özkan, M. & Özdemir, E. B. (2014). Ortaokul 8. sınıf öğrencilerinin ve öğretmenlerinin ortaöğretime geçişte uygulanan merkezi ortak sınavlara ilişkin görüşleri. *Journal of History School (JOHS),* DOI:10.14225/Joh641

Öztürk, F. Z. & Aksoy, H. (2014). Temel eğitimden ortaöğretime geçiş modelinin 8. sinif öğrenci görüşlerine göre değerlendirilmesi (Ordu ili örneği). *Ondokuz Mayis University Journal of Faculty of Education*, DOI: 10.7822/omuefd.33.2.8

Sullivan, G. P. (2006). *The impact of high stakes testing on curriculum, teaching, and learning.* (Doctoral Dissertation, Faculty of the Virginia Polytechnic Institute and State University). Retrieved from https://vtechworks.lib.vt.edu/handle/10919/

Şad, S. N. & Şahiner, Y. K. (2016). Temel eğitimden ortaöğretime geçiş (TEOG) sistemine ilişkin öğrenci, öğretmen ve veli görüşleri. *Elementary Education Online,* DOI:10.17051/io.2016.78720.

Tavşancıl, E. & Aslan, E. (2001). *İçerik analizi ve uygulama örnekleri*. İstanbul: Epsilon.

Yıldırım, A. & Şimşek, H. (2005). *Sosyal bilimlerde nitel araştırma yöntemleri*. Ankara: Seçkin.

Zorlu, Z. & Zorlu, F (2015). Fen ve teknoloji dersinde öğrenme ortamına yönelik öğrencilerin düzeyleri ve öğretmenlerin görüşleri. *Route Educational and Social Science Journal*, *2*(1), 103-114.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(1);80-95

An Analysis Program Used in Data Mining: WEKA

Gökhan AKSU* Nuri DOĞAN**

Abstract
In this study, it is aimed to introduce one of the data mining methods which is very popular in recent years and commonly used in this area. For this purpose, the WEKA program and the decision trees, which is one of the methods used to estimate the dependent variable through independent variables, will be introduced. In today's age of technology, the amount of information at hand is constantly increasing and the derivation of meaningful results from this information is seen as a valuable field of study. Data mining aims to reveal the information that is hidden in a large amount of data after a series of operations, which is very useful for researchers. Regarding this approach that is mostly based on estimation and classification, there is a lot of new and unvalidated software that has not yet been fully tested. In this study, we discuss WEKA software, which is one of the programs in the field of data mining, how to run the program and the content of the analyzes and output files. The study also contains some suggestions for the practitioners who want to use this program about the superior aspects of the software and what kind of analysis can be done with it.

Key Words: Data mining, WEKA, Classification, Prediction, Algorithm

INTRODUCTION

Thanks to the Internet, a major revolution has been occurred in accessing and using information over the last decade (Jain, 2015). At this stage, researchers and scientists focused on storing, recalling and using data when needed. From time to time, the data at hand is likened to a gold mine for conducting research and development in a particular area. Data mining is a process that defines the data obtained as input and output information (Weiss and Davison, 2010). Fayyad, Piatetsky-Shapiro, and Smyth (1996), one of the most cited researchers in the field, describe data mining as the execution of certain algorithms to elicit certain patterns from the available data.

Simple structures of different types are formed on the data set, in a way that they can be easily displayed. For example, in a data set, there may be examples where one property works very well, whereas the others are unrelated or unnecessary. In one data set, there may be examples where the properties contribute independently and equally to the output variable. In another data set, there may be a number of properties that have a simple logical structure and that are obtained by the decision tree. In a different data set, there may be several independent sets of rules that assign examples to different classes. In another data set, interdependencies between different subsets of properties can be examined. In a data set, linear dependence between weighted sums of numerical properties determined by appropriate weighting methods can be examined, whereas in another it is possible to place these examples on specific regions of the sample space based on the distances between the examples. In a different set of data, there may be no class value as seen in the algorithms where learning is uncontrolled (Witten and Frank, 2005). There are different examples where different structures and different data mining tools are used in the infinite equality of possible data sets. In these examples, there are also algorithms that may completely overlook the appropriateness of different types and make correct classifications for only one class (Holte, 1993).

Classification rules are alternatives to decision trees. The given rules are first executed for the first line, and then for the second and the last line respectively. The lines setting the rules are generally

*Dr., Adnan Menderes University, Aydın Vocational School, Aydın-Turkey, e-mail: gokhanaksu1983@hotmail.com ORCID ID: 0000-0003-2563-6112
** Prof. Dr., Hacettepe University, Education Faculty, Ankara-Turkey, e-mail: nuridogan2004@gmail.com, ORCID ID: 0000-0001-6274-2016

To cite this article:
Aksu, G., & Doğan, N. (2010). An analysis program used in data mining: WEKA. *Journal of Measurement and Evaluation in Education and Psychology, 10*(1), 80-95. DOI: 10.21031/epod.399832

Received: 28.02.2018
Accepted: 29.11.2018

named as "Decision List". These lists show the rules required for the correct classification of the examples in the decision table (Chadha and Singh, 2012).

The rules of association are not different from the classification rules except that they are not limited with classification, they can also estimate any property. In this approach in which the patterns, relationships and causal structures in the data set are revealed, different combinations of properties can be estimated as well. In addition, the association rules do not require the use of a set of sets of rules, as in the rules of classification (Han, Kamber and Pei, 2000). A large number of different association rules can be derived from a small data set. For this reason, applying it to a large number of examples is not needed. The rules of association are generally formed by "if…. then" cycle. For example, "if there is no wind and the game is not played, the humidity is high" association rule indicates that the humidity should be high if the game was not played when there is no wind (Witten and Frank, 2005)

Clustering rules take the form of a diagram showing how the examples in the data set fall into specified clusters. The results obtained in this method, where the clusters' belonging is determined, are sometimes given as a dendrogram and sometimes as a table. The cluster, to which each example in the data set belongs, is shown according to its characteristics regarding the similarity or dissimilarity criteria with the other examples. On the dendrogram, each cluster is depicted on a sub-level, as decomposed into its subsets (Karypis, Han, and Kumar, 1999).

In the prediction procedure, numerical properties are considered instead of categorical variables. In data mining, even though categorical variables are considered in decision trees or association rules, linear regression models are used to estimate the numerical value of the property (Padmavathi, 2012). Apart from classical regression methods, a more accurate and consistent estimation can be made by logistic regression method. In this approach, where input variables are weighted according to their contribution to the model, a large number of logistic equations are obtained instead of a single regression equation (Perlich and Provost, 2002).

Decision Trees

The problem of creating a decision tree is expressed self-repetitively. First of all, a property is determined to be set for the root node (a node without ancestor node and therefore the node at the top) and a branch is created for each possible value (Fayyad and Irani, 1992). This process breaks the data set into subsets for each property value. The process is repeated successively for each branch by using only the examples that reach this branch. If, in any case, all examples in a node belong to the same class, then the development of that branch of the tree (node) is stopped. Because from this point on there will be no decomposition into different classes (Quinlan, 1993). The only thing left to make a decision is to determine the way of dividing each property when a series of different classes is given. Below are the results of the game played for the general view and temperature characteristics of the weather. There are 2 possible (Yes-No) alternative for each branch and they are divided into classes at the top, as shown in Figure 1.

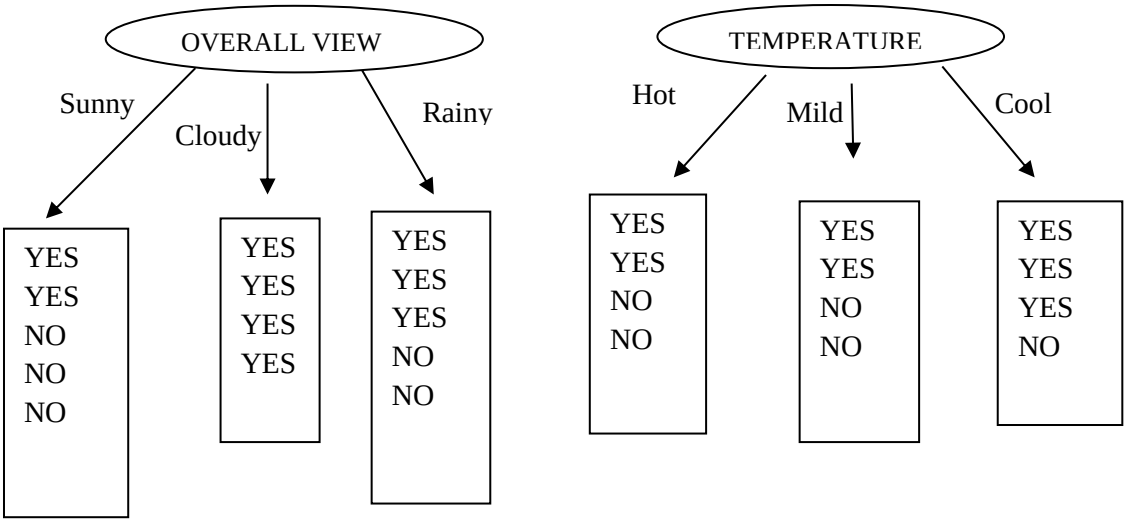


Figure 1. Tree Structure for Weather Data (Witten and Frank, 2005)

We can decide which of the branches are the best choices for the tree structure given in the figure regarding the class of the leaves shown in the rectangular shape. The numbers of yes and no classes are shown on the leaves. If only a single class is created on the sheet (Yes or No), there will be no need to divide again and the iterative branching process will end.

We want this process to be as short as possible, thus we examine small trees. If we measured the purity of each node, we would have to choose the properties that produced the purest child nodes. Now take some time and consider which feature will be the best.

The measure of purity that we will use is called information and it is measured by units designated as bits. The bit represents the amount of information required for a new Yes (played) or No (not played) classification related to a node of the tree. Of course, there is nothing special about these numbers and there is a similar relationship between them, regardless of their actual value. Thus, we can add another criterion to the list. The information obtained must be in accordance with the multi-stage property shown previously. Remarkably, it appears that there is only one function that meets all these characteristics, and this function is called the amount of information or entropy (Rokach and Maimon, 2008). The following equation shows how the amount of information is obtained mathematically.

Entropy (p_1, p_2, …, p_n) = - p_1.logp_1 – p_2.logp_2, … , - p_n.logp_n

The negative signs (-) here are due to the rules of the logarithm coming from the fractions p_1, p_2, … , p_n, while calculating the logarithm of the fractions in the form of A / B, the base remains unchanged and a minus sign is placed before the expression in the denominator ($\log_a {}^{A}/_{B} = \log_a A - \log_a B$). In spite of the minus sign in the expression, in reality, the amount of information will be positive.

Usually, logarithms are given at the base of 2 ($\log_2 x$) and in this case the amount of information called entropy is defined as bit. These bits are a generic type of bits used in the computer (1-0).

p_1,p_2,…,p_n expressions of the entropy indicates fractions that sum up to 1. For example;

info ([2,3,4]) = entropy (2/9, 3/9, 4/9)

As shown in the above formula, the sum of 2/9, 3/9 and 4/9 fractions specified as p_1, p_2, p_3 will be equal to 1. Therefore, decisions regarding multi-stage characteristics can be obtained by the general formula given below.

entropy (p,q,r) = entropy (p,q+r) + (q+r). entropy ($\frac{q}{q+r}$, $\frac{r}{q+r}$)

It should be kept in mind that p+q+r = 1 in the given formula.

info ([2,3,4]) = $-\frac{2}{9}$ x log $\left(\frac{2}{9}\right)$ - $\frac{3}{9}$ x log $\left(\frac{3}{9}\right)$ - $\frac{4}{9}$ x log $\left(\frac{4}{9}\right)$ = $-\frac{2}{9}$ x (log2 – log9) - $\frac{3}{9}$ x (log3 – log9) - $\frac{4}{9}$ x (log4 – log9)

$$= \frac{2\log2+2\log9}{9} - \frac{3\log3+3\log9}{9} - \frac{4\log4+4\log9}{9}$$

$$= \frac{-2\log2-3\log3-4\log4-9\log9}{9}$$

The above formula is a formula that shows how the amount of information is calculated in practice.

Reliability: Evaluation of What We Learn

Evaluation plays a key role in achieving real progress in data mining. There are many different ways to extract patterns from the data file and to make deductions from these patterns. However, while determining which method should be used for a particular problem, we should have systematic ways of evaluating and comparing different approaches in which different methods are used. The evaluation process is not as simple as it seems.

To see which of the two different methods works better, training and test sets are needed. But in fact, learning the performance in the training set will not be a good indicator of performance in an independent test set. We will need ways to estimate performance limits based on the experimental studies to be conducted on the data sets (Breiman and Friedman, 1984). More clearly, training the learning method through a data set and then test this method over the same data set would not be an approach that reflects reality. Therefore, the error rate on the training set may not be a good indicator of actual performance. Because, since the classifier learned from the same training data, any calculation process based on this data will be optimistic (Witten & Frank, 2005).

Our concern in data mining will be the future performance of the method based on the new data rather than past performance based on the historical data. The class of each example in the training set is already known. Even though we are not interested in the classification, and our aim is to clear the data rather than making a prediction, the classifications should also be taken care of. To estimate the performance of a classifier on the new data, we need to evaluate the error rate on a different data set than the data set from which the classifier is obtained. In the studies, we should assume that both training data and test data are representative examples of the problem that we are investigating (Sumathi & Sivanandam, 2006).

In some cases, the test data may differ from the training data in nature. For example, suppose we have examples of credit risk problems. Let us assume, for example, that the bank has received training data from branch offices in New York and Florida, and wants to learn how well the classifier to be obtained through this training data will perform at a new branch in Mexico. It would probably be appropriate to use New York data as the test data to evaluate Florida classifier and to use Florida data to evaluate New York classifier. However, if the data sets were combined in the process of training, the performance in the test data would probably not be a good indicator of the performance of data obtained from a different state in the future (Witten, Frank and Hall, 2016). In the studies conducted in the field of education, it has been investigated whether a different learning method is effective on the courses that students attend school. In a study by Lopez et al. (2012), it was investigated whether the forum usage data in the Moodle learning management system is a significant indicator of the success of the course. In the study, it was tested whether the same result can be achieved with clustering algorithms in cases where the class variable (course achievement) was not known.

In most cases, the training data should be classified by hand and, of course, the test data should also be classified likewise in order to obtain the error rate. This limits the amount of data that can be used for training, verification, and testing, and how to achieve the best performance with this limited data becomes a question. The answer is, a certain amount (20% - 30%) of the data set is kept for testing, which is called the retention procedure, and then the remaining amount is used for training. However,

if necessary, a part of the data to be used for training can also be separated as the validation data (Mitchell, 1997).

Cross-validation

What will you do if you have limited amounts of data for training and testing? In the retention method, a portion of the available data is used for testing and the rest is used for training. However, a part of the data allocated to training can be used for verification. In practice, allocating one-third of the available data for testing and the remaining two-thirds for education is a very common method (66% = Education, 34% = Test data).

Of course, the sample group that you use for training (or test) may not be a good representative of the universe. In general, you cannot directly understand whether a sample is a good representative of the universe or not. But there is a very simple control that can be significant for you. In this approach, each class of your universe, in which all the data is included, should be represented in the training and test data with an accurate ratio. Only this will increase the representation power of the universe. However, if you are unlucky and you have lost data in all examples of a class in your sample, you cannot expect the classifier, which will be obtained through this data, to be able to perform well on the test data. In this case, the results will be worsened as the data in the test set is over-represented, since none of these data is included in the training set. Instead, you must ensure that random sampling was carried out to ensure that each class in the universe is properly represented in the training and test data. This method is called the stratification or stratified retention approach. Although it is often noteworthy to apply this method, this is seen only as primitive protection to eliminate an uneven or unbalanced representation in the training and test data (Witten, Frank, & Hall, 2016).

A more generic way of reducing the bias arising from a particular sample selection is repeating the whole process several times for training and test data across different sample groups. In each repetition process, a certain proportion of the data (for example two-thirds for the training) can be used for training (if possible, in a stratified and randomly selected way), whereas the rest can be used for testing. A more general error rate can be obtained by taking the average of error rates from different repetitions. This is called the repetitive retention method for calculating the error rate.

In cases where only one retention method is applied, you may think that the training and test data exchange roles. In other words, you can train the system with the test data while testing the system with the training data. This way, you reduce the problem of unequal representation in training and test data by averaging the two results you get.

Unfortunately, this only makes sense for 50% - 50% splitting of training and test data, which is not generally considered as a suitable split. Instead, using more than half of the available data for training is a better approach. On the other hand, there is an important statistical technique called cross-validation. In cross-validation, we set a constant folding or splitting number for the data, for instance, let's set this constant number as 3. The data is then divided into three equal parts and each of them is used for testing while the remaining portion is used for training. That is to say, two-thirds of the available data are used for training while one-third of the data at hand is used for testing. When this process is repeated, each example will be used once for testing. This process is called the triple cross-validation and if stratification is carried out, it is called the stratified triple cross-validation (Kohavi, 1995).

When there is a single and constant data available, a 10-fold cross-validation approach is applied to the standard method used to estimate the error rate of the learning technique. In this approach, the universe of the study, which includes all of the data, is randomly divided into 10 classes and each class must be represented in the whole data set at approximately the same rate. 10-fold cross-validation process is illustrated in Figure 2 in order to make it easier to understand.

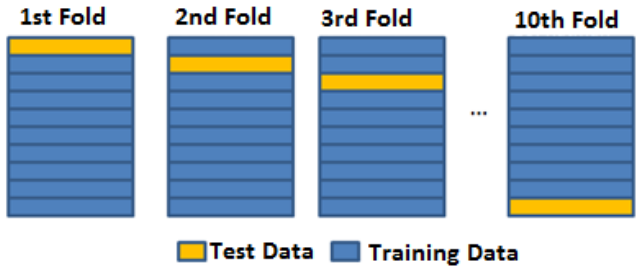


Figure 2. Ten-fold cross-validation

In this process, each fold is held separately for the test, while the learning process is performed through the remaining nine folds and the error rate is calculated from the 1/10 data held separately. Thus, the learning method is repeated for a total of 10 times over different training data (each with many common points). Finally, a general error rate is obtained by averaging 10 different error rates (Refaeilzadeh, Tang and Liu, 2007).

Bootstrap Method

This method is based on statistical procedures that can be defined as repeated sampling. A sample is taken from the current data set for training or testing beforehand and these examples are excluded. That is, once an example is selected, it cannot be selected again. This approach is similar to building teams for a football game. Just as you cannot choose a player for two different teams, in the Bootstrap method the same example cannot be selected twice. However, the examples in the data set are not like humans. Most learning methods use the same example twice, and the results will differ if it takes place twice in the training set. Mathematically speaking, in case of having more than one image of the same object, it is not meaningful to mention the groups defined as "cluster" (Ibarguren et al., 2014).

The logic of the Bootstrap method is to be able to sample the data set at hand by relocation in order to be able to create a training set. This stage will show the mysterious Bootstrap number (0.632) and how it is obtained. For this, a data set consisting of n examples is relocated so that another sample consisting of n examples is sampled. Since some of the examples from this second data set will be repeated, the original data set should contain examples used in somewhere else, to be used as test samples

But what is the probability that a particular example is not included in the training set? The probability of being included in the training set is $\frac{1}{n}$ and the probability of not taking part in this group is calculated as $(1 - \frac{1}{n})$. The multiplication of the probabilities of not being in the given cluster is calculated with the help of the equation given below.

$$(1-\frac{1}{n})n \approx e^{-1} = 0{,}368$$

In the equation given above, the symbol shown by e is the base of the natural logarithm and it is equal to 2.7168. However, you should keep in mind that this display is not equal to the error rate. This equation shows the possibility of not being selected in any way for a particular example. This way, the test set will cover approximately 36.80% of the data for a fairly large data set and therefore the training set will cover 63.20% of the data at hand. So, we learned where the number of Bootstrap 0.632, which was previously defined as a mysterious number, is coming from. Some examples are re-used in the training set to reach the total number of n in the original data set (Kushary, 2012).

This numerical value obtained by training a learning system through the training set and the error value obtained from the test set will be a pessimistic estimate of the actual error. Because even though the size of the training set is n, it still covers 63% of the samples and considering that 90% of the data is used in 10-fold cross-validation, it can be seen that this is not a larger proportion. To compensate this, the error obtained from the relocation process through the examples in the training set is combined

with the error rate obtained from the test set. The error value obtained as a result of the relocation should not be used as the error term alone because it will be a very optimistic estimate of the actual error value. However, the Bootstrap method generates the error value in the combination of the errors obtained from the training and test data and the value is shown below is calculated as the final error value (Bradley and Tibshirani, 1993).

$$e = 0{,}632 \times e_{\text{test data}} + 0{,}368 \times e_{\text{training data}}$$

The entire Bootstrap method is then repeated several times by relocating different examples for the training set and the results are averaged.

The Bootstrap method can be considered as the best method for estimating the error rate for small data sets. However, just as in the method of excluding one group, it is possible to show its disadvantage through a completely artificial and special situation. In fact, assume that the data set that we previously considered had been randomly divided into two classes. The actual error rate of any prediction rule is assumed to be 50%. But let's consider the training set; as a result of the relocation, a draft learning method will give 100% success with zero error (e = 0). In this case, the 0.632 Bootstrap method will weigh this with 0.368 and accordingly the overall error term will be calculated as 31,60% (0,632 × 50% + 0,368 × 0%), which will be misleadingly optimistic.

Power of Estimation: Considering the Magnitude of the Error

The evaluation process performed according to the accuracy of the classification implicitly assumes that different errors do not have the same meaning. An example of this can be seen in credit debts. The mistake of giving a loan to someone who did not pay his/her previous loan is much bigger than the mistake made by not giving credit to someone who has never taken a loan.

For the cases where there are two classes in the form of yes-no, borrowing-not borrowing money, giving-not giving a loan, classifying a suspicious part as scrap or not, etc., there are 4 different outcomes, as shown in Table 1. The true positive (TP) and the true negative (TN) are the correct classification results. TP are the cases when the test result is positive when the actual situation is positive, whereas TN are the cases when the test result is negative when the actual situation is negative. False positive (FP) is predicting the output incorrectly as positive (YES) while the result is actually negative (NO). False negative (FN) is predicting the output incorrectly as negative (NO) while the result is actually positive (YES). True positive ratio (TPR) is obtained by dividing positive cases to all cases [TP/(TP+FN)], whereas false positive ratio is obtained by dividing false positives to all negative cases [FP/(FP+TN)] (Powers, 2011).

Table 1. Different Results Regarding a Two-class Estimate

		REAL CASE		
		POSITIVE	NEGATIVE	TOTAL
TEST RESULT	POSITIVE	TRUE POSITIVE (TP)	FALSE POSITIVE (FP)	TP+FP
	NEGATIVE	FALSE NEGATIVE (FN)	TRUE NEGATIVE (TN)	FN+TN
	TOTAL	TP+FN	FP+TN	TP+FP+FN+TN

The overall success rate after computing the ratios of the outcomes is the division of the number of correct classifications to the total number of classifications.

$$\text{Overall Success} = \frac{TP+TN}{TP+TN+FP+FN}$$

The error rate for this classification is calculated by subtracting the overall success rate from 1. Sensitivity and selectivity concepts are also important to know. Sensitivity is the ability of the test to

distinguish positive cases from actual positive cases. At the same time, the true positive rate is defined as sensitivity.

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

Selectivity: It is the ability of the test to distinguish negative cases from actual negative cases.

$$\text{Selectivity} = \frac{TN}{TN+FP}$$

The sensitivity and selectivity values described above define how well the test differentiates non-relevant and relevant cases (Johari, 2016). False positive rate is considered as Type I Error and False negative rate is considered as Type II error.

In a multi-class prediction, the results obtained from the test are shown on a table called an error matrix, having rows and columns for each two-dimensional class. The matrix is defined in a way that each column represents the actual value of an example, whereas each line represents predicted value (test result) of the example (In some cases the rows show the actual value, and the columns the test results). Good results are obtained when the numbers on the main diagonal (the line drawn from the upper left corner to the lower right corner) are quite large and the non-diagonal elements are small (ideally zero).

The curves which are known as ROC (Receiver Operating Characteristic) curves and trying to select test samples with high positive rate attempt to describe the performance of a classifier regardless of the type of error and class distribution. In these graphs, the number of positive examples corresponding to the negative examples is shown in the horizontal axis (x-axis) as a percentage, whereas the number of negative examples corresponding to the positive examples is shown in the vertical axis as a percentage (Sinha & May, 2005). The Kappa statistic or Kappa value is a numerical value in which the expected and observed values are compared. Besides, it is less misleading because it takes the chance factor into account. Expected and observed accuracy calculation is used simultaneously in Kappa statistics and it can be easily determined through the confusion matrix. In this method, the values in the rows show the actual values, while the values in the columns show the estimated values; the expected accuracy rate is subtracted from the actual accuracy rate and this value is divided to (1-expected accuracy rate). Although there is no definitive standard interpretation of the Kappa statistic, in general, the values in the range of 0.00-0.20 are considered as low, 0.21-0.40 as notable; 0.41-0.60 as mediocre; 0.61-0.80 as important and 0.81-1.00 as excellent (Landis and Koch, 1977). There are different error criteria in the evaluation of numerical estimation in data mining. Mean square error, root mean squared error, mean absolute error, squared relative error, root relative squared error and relative error are the error criteria used to determine how accurate numerical estimation is (Witten and Frank, 2005). Relative error values attempt to compensate the fundamental predictability or unpredictability of the output variable, whereas square and square root criteria of the errors allow reducing the errors to the same size. The F-criterion used in data mining is obtained from certainty and recall values. Precision, which is of great importance particularly in the medicine and medical field, shows the success in a condition predicted as true (positive). Recall is more important in marketing and marketing research and shows how successfully positive situations are predicted (Schwenke and Schering, 2007)

DESCRIPTION OF THE PROGRAM, ANALYSIS, EXAMPLES OF ESTIMATION AND CLASSIFICATION

Explorer, which is the main interface of the WEKA program allows easy access to all operations by providing menu selection and form filling options. As shown in Figure 3, the window on the WEKA main screen displays the different mining tasks that the WEKA program supports under six different tabs. Knowledge Flow is a feature of WEKA that allows multiple uses of the features across a screen. Even though only one operation can be performed on the Explorer screen, Knowledge Flow is a feature that allows the tasks to run repeatedly over a number of different operations. Under the Explorer

screen, everything is automatic and ready. In Knowledge Flow, these processes should be created by the user (Şeker, 2016).

Figure 3. WEKA Main Screen

Imagine that you have some data and you want to obtain a decision tree from this data. First, you need to prepare your data, then run the explorer and upload the data to the program. Then choose a decision tree creation method, create a tree, and interpret the output you get. It is easy to repeat this process with a different decision tree algorithm and a different evaluation method. Under the Explorer menu, you can make back and forth transitions between the outputs you get, you can evaluate the models built on different data sets and you can graphically visualize both the models and the datasets including classification errors made by the models.

Below and in Figure 4 & 5, six different tabs at the top of the explorer window are briefly defined. Each of these tabs shows different actions that you can perform with the data at hand.

1. PREPROCESS: allows you to select the data set and edit it in different ways.
2. CLASSIFY: allows you to train the learning method that will classify or predict and evaluate them
3. CLUSTER: allows you to learn clusters of the data sets
4. ASSOCIATE: allows you to learn the rules of association for your data set and evaluate them
5. SELECT ATTRIBUTES: allows you to select relevant properties in your data set
6. VISUALIZE: allows you to see different two-dimensional graphs of your data set and to determine the interaction between them.

Each option provides access to a range of possibilities. Up to now only preprocess and classification options have been considered superficially. For the researchers who want to conduct further analysis, it is recommended to examine the options of cluster, associative, select attributes and visualization.

Preparing Data for Analysis

Data is usually presented in the form of tables or databases. However, the data storage method of the WEKA program contains the aggregated list of examples in ARFF file format, with the data you have entered in the table in ARFF format, in which attribute values specified for each example are shown as separated by commas. Most tables and database programs allow you to convert your data file into .csv (comma-separated value) type. In this file type, the data is stored by inserting a comma between the values in the data file. Once you have done this, it will be sufficient to upload your file to a text

document or to the software. Then save the relations with the "@relation" extension, attribute information with "@attribute" extension, and your data file with "@data" extension as a raw text document. For example, an Excel data file for the PISA data that will be mentioned later is saved differently with a .csv extension and this file is opened from the program and converted to a WEKA data file with .arff extension. However, you don't really need to follow these steps to create the ARFF file yourself, because the explorer can directly read the files with CSV extension.

Introducing Data to the Program

Now, let's upload the data you have at hand to the explorer and try to analyze it. Start the WEKA program to access the screen shown in Figure 4. Then select the explorer, which is one of the five different interfaces under the applications heading.

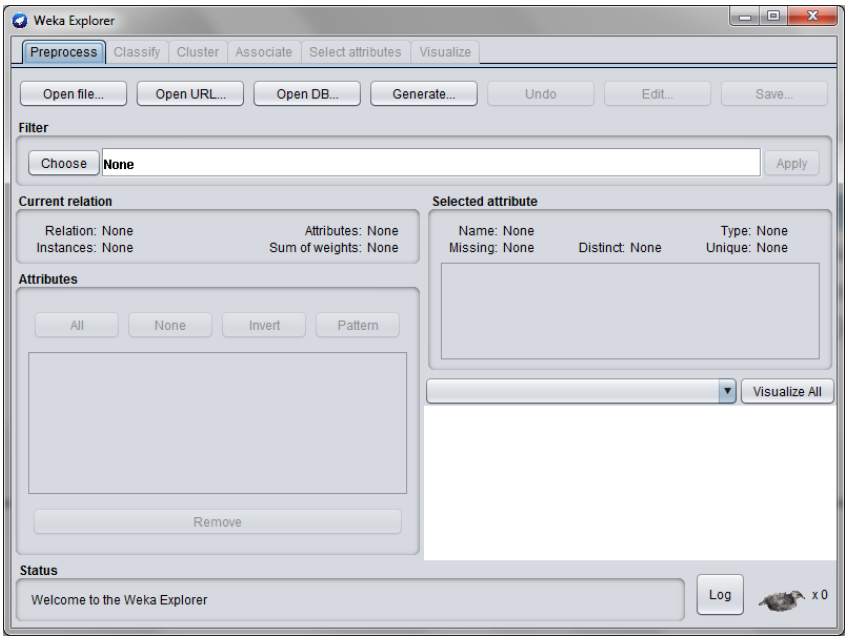


Figure 4. Data Preparation Screen

Then, you can see one of your previous files by clicking the Open File button. By defining your file saved on your computer, you will show your data to the program. If your data is not in ARFF format, which is the basic data file extension of the WEKA program, you will need to select the file type as CSV. When you specify a file with a .csv extension in the program, this file will be automatically converted to a file type with ARFF extension. The screen that you'll face after uploading the file provides you with information about the data set you have.

Performance of Analysis: Creating the Decision Tree

C4.5 decision tree learning method which is one of the most used algorithms in data mining works with J.48 algorithm and this algorithm is a version of the WEKA program that can be used by everyone before the launch of C5.0 application (Kaur and Chhabra, 2014). When you click on the CLASSIFY button shown in Figure 5 (a) and then click on the CHOOSE button from the screen, the screen shown in Figure 5 (b) will appear. Since there is no analysis on the screen, there is no result in the output window in the lower right corner. Once the algorithm and the test type to be used for the classification are set, all you have to do is click the START button.

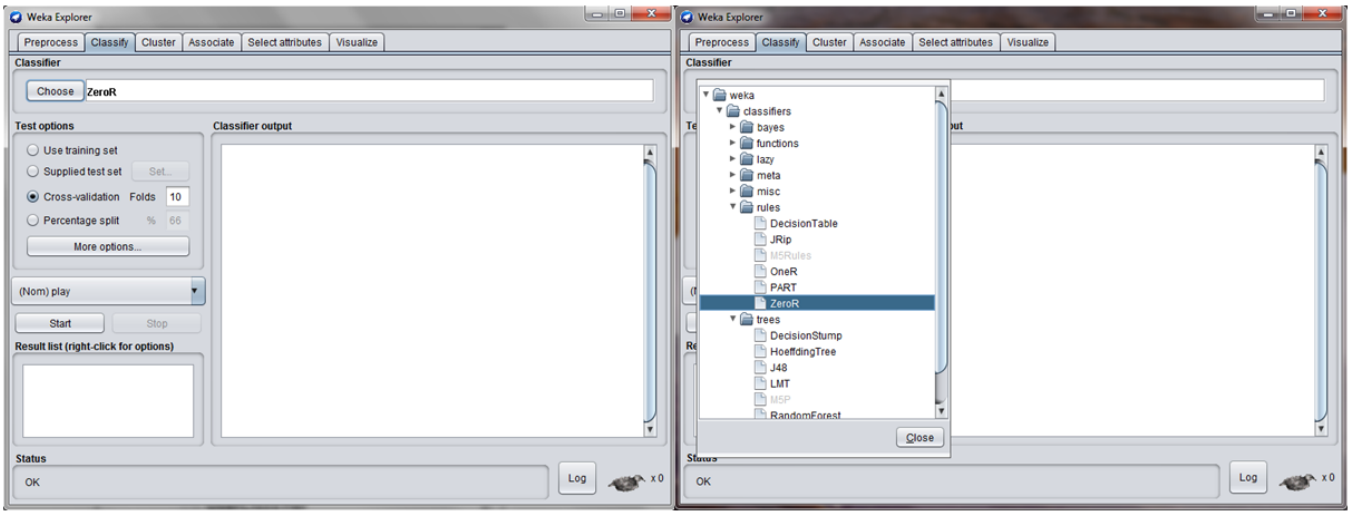


Figure 5. Startup Screen: (a) Choose Window and (b) Classify Window

In order to perform the analysis first, click the choose button located at the top left, then click on Trees section from the hierarchically listed menu and choose J.48 method. At the moment, all you have to do is to choose the method or algorithm that is always at the lowest level from the methods presented sequentially. When you select the algorithm that you will use, you will see the parameters for the method or algorithm you selected in the row next to the choose button. If you double click on this line, J4.8 classifier editor will be opened, the parameters and the numerical values corresponding to the desired values accepted by the program will be displayed. Of course, these default values are determined to make more precise measurements.

After selecting the classifier, you can run the program by clicking the Start button. The WEKA program has a very fast processing capability and it can perform analysis in a short time and while the analysis is in progress a bird will move in the lower right corner of the main screen shown in Figure 5 (a).

As an example, the screenshot obtained after uploading the variables covered in the PISA 2015 student questionnaire, namely the duration of science learning (smins), the total learning time at school (tmins) and the socioeconomic status index of the student (escs) and the input and output variables used in the process of estimating science literacy levels (pv1scie) to WEKA is shown in Figure 6.

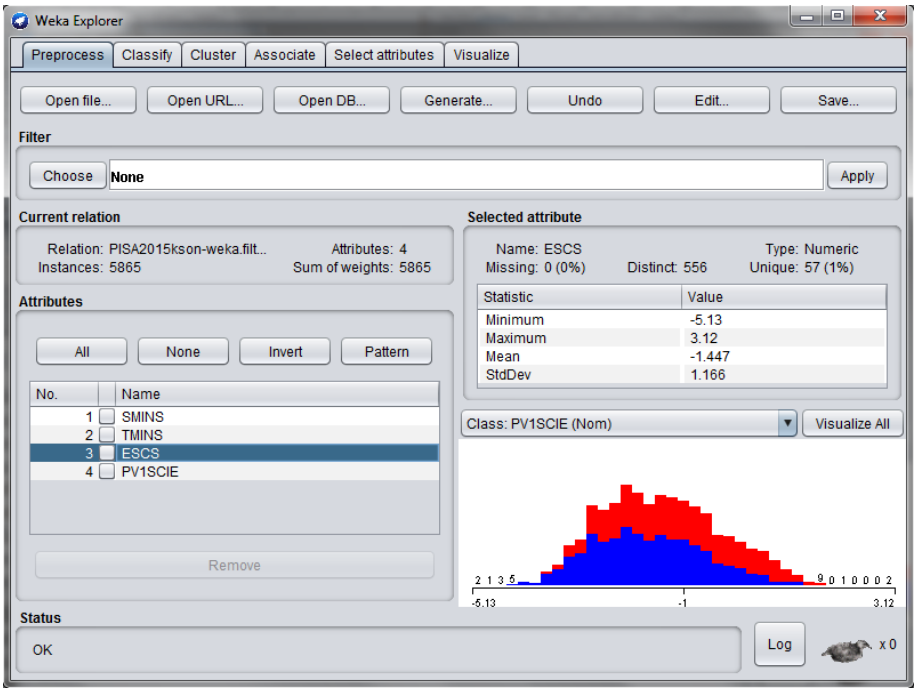


Figure 6. Uploading Input and Output Variables to the Program

After this process, Classify interface is opened, the window shown in Figure 7 is reached by clicking the Choose button in the Classifier window and one of the decision trees under the Trees tab is chosen. As an example, analyzes were performed by selecting J.48 method.

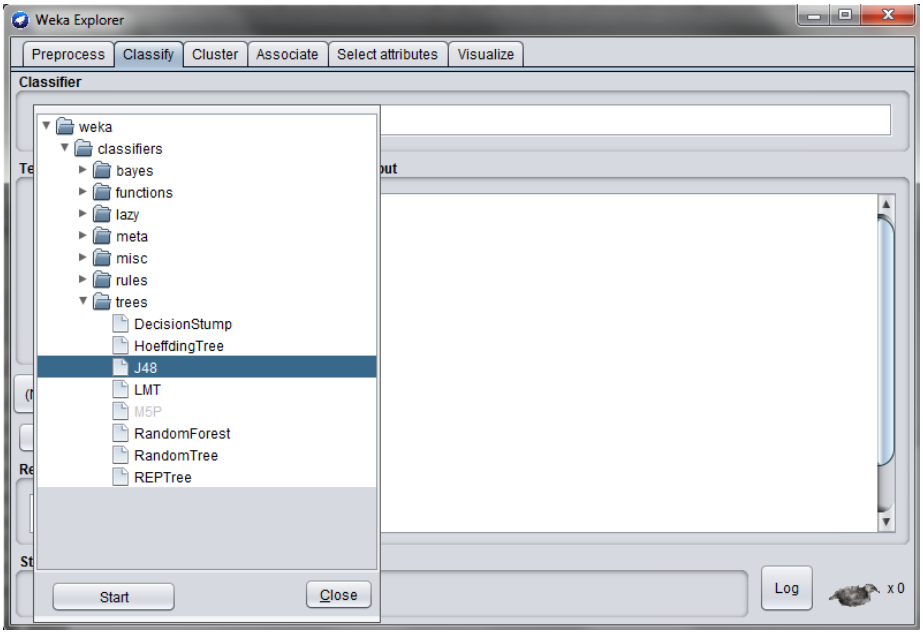


Figure 7. Determination of Decision Tree Method

The analysis process is completed by clicking the Start button on the screen shown in Figure 7. If researchers want to use a different method instead of a 10-fold cross validation method, one of four different validation types can be selected in the test options window. After this process, the window shown in Figure 8 is reached by clicking the Start button.

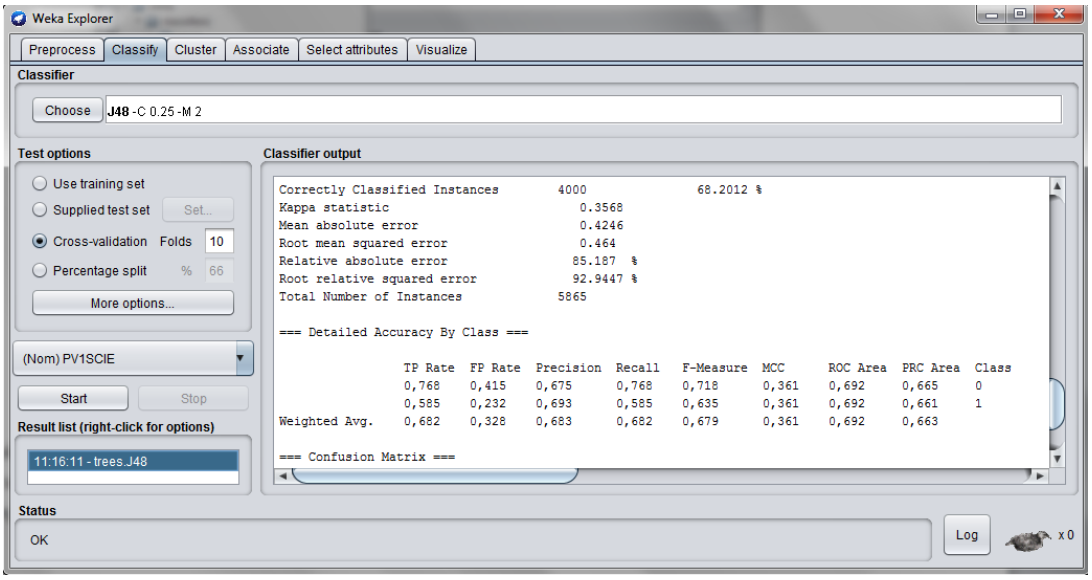


Figure 8. Decision Tree Analysis Outputs

In the classifier output window shown in Figure 8, the number of correctly classified students, statistics of the classification process and confusion matrix are given. Other outputs of the classification process can be seen by moving the cursor down. In order to see the decision tree, researchers should go over the trees.J48 line in the Results list window in the lower left corner of the screen shown in Figure 8 and activate Visualize tree option by right clicking it.

Evaluation of the Results

Within the scope of the study, duration of science learning (smins), total time of learning in school (tmins) and student's socio-economic status index (escs) variables covered in PISA 2015 student survey are defined as input, whereas science literacy level (*pv1scie*) is defined as the output variable, then the decision tree for estimating the output variable with input variables is shown in Figure 9.

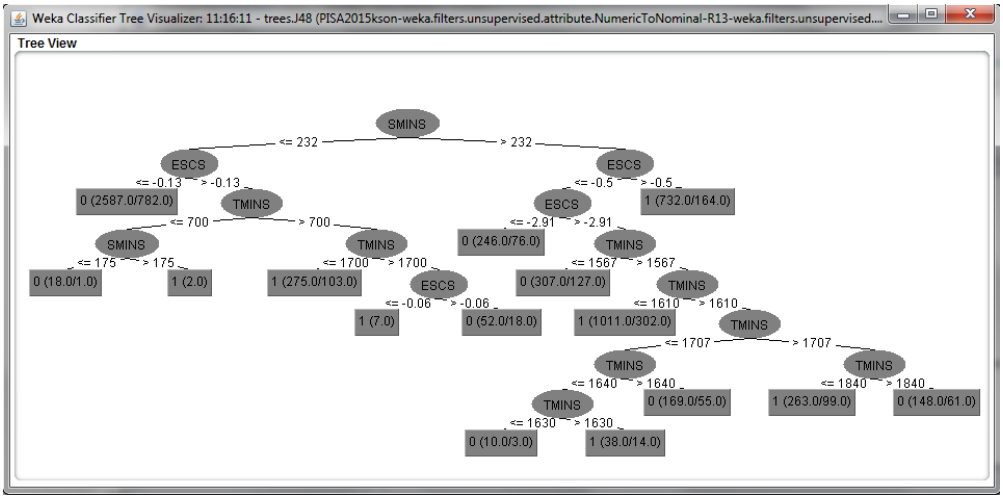


Figure 9. Decision Tree Obtained Through J.48 Method

As can be seen in Figure 9, it is found that duration of science learning (smins) variable has the most significant effect among the three input variables which are covered in the study to classify students in terms of PISA Science literacy. Students are divided into two classes according to the cutting point of the duration of science learning, which is 232. Regarding the students whose science learning time is below 232, a socio-economic status index has the most important effect in the classification of students followed by total learning time variable. At the second level of tree branching, the cut-off point according to socio-economic status index is determined as 0.13 and students below this value are classified as Failed (0). On the same branch, the total learning time of those with a socio-economic status index over 0.13 is checked and the cut-off point for this property is set at 700.00.

In the output file, the classification error of the learning method, and the evaluation results of the tree classification process, including Kappa statistics, the mean absolute error, and the root mean squared error are reported. The root mean squared error is the squared average of the second order loss function. The mean absolute error is calculated by taking the absolute value of the differences instead of taking the errors. Moreover, the output also includes relative error values calculated based on the a priori probability distributions. These statistics are obtained using the ZeroR learning method. Lastly, when the output file is examined, it is seen that detailed accuracy, precision, recall, and F-criterion values are reported for each class.

RESULTS AND DISCUSSION

WEKA program, which is one of many different programs that allows drawing meaningful results from the existing patterns in a data file, seems to be more popular than others because of having a user-friendly interface and being an open source software (Zupan and Demsar, 2008). Especially in quantitative estimation, LMT algorithms based on J4.8, ID3, M5P, and logistic model offer you many options besides Bayesian methods. In addition, it is possible to make estimations with a high level of accuracy and precision with random tree structures such as Random Trees, Raptree and Random Forest. To move one step further than the classical methods of estimating the dependent variable using independent variables, logistic regression model, which has different equations for each branch, is used instead of a single regression equation for the estimation with higher accuracy (Robu and Hora, 2012). Again, offering more than 20 different options under the rules and functions menu is one of the important advantages of the program. In addition, another advantage of the program is being able to read data files with .csv extension directly, without needing any converters.

Although the decision trees for Hoefding tree, J.48, Logistic model, Reptree and Random Tree algorithms can be easily created by the program, the decision trees are not given for the other algorithms, which can be considered as one of the limitations of the program. In addition, the numerical properties with a negative sign and a decimal value cannot be directly read by the program, which is also seen as a limitation.

Data mining is mainly based on classification and prediction algorithms. Although you have obtained a categorical variable when you divide a numeric property into two classes according to a certain threshold value, there is a significant difference between these two. In the classification properties, all information about the property is used in the branching process, whereas in numeric properties you continue to use information about the property in consecutive nondisjunction. In other words, sequential branching in digital properties will continue to produce new information. A classified property can only be tested once, from the root of the tree to a certain leaf of the tree, whereas a numerical feature can be tested several times. Therefore, trees may be more complex and more difficult to understand. Because the tests for a property are not performed together and there may be dispersion along the way. The way to create a tree that is more readable but more difficult to achieve is allowing a multidimensional test for the specified property and testing a few constants on a single node of the tree. For this reason, it is more useful to continue analyzing numeric properties instead of classification. However, at this point, the WEKA program is, unfortunately, unable to predict what will be the numerical value of the dependent variable through independent variables. Although it is possible to obtain the result variable of the relevant example by defining the data related to a single property in

the model with the help of the logistic model established by the learning method using the necessary codes, failing to estimate the dependent variable numerically for all examples is considered to be an important limitation in data mining where very large data is used.

Researchers who are going to study about data mining are recommended to review one of the latest versions of the free program offered by Waikato University by downloading from https://www.cs.waikato.ac.nz/ml/weka/downloading.html, within the framework of the superiorities and limitations mentioned above. In particular, they are advised to repeat the analyzes on the sample data files that the program provides and to interpret obtained results. Besides, Select Attributes feature based on the cross-validation method is thought to be useful for the researchers in reducing the number of variables, especially if there are too many variables at hand. Although the program is open source, researchers who study on the literature are advised to compare the results that they obtained with the command files that they have written with the results of other software, preferably with the assistance of a programmer.

REFERENCES

Bradley, E., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Boca Raton, USA: Chapman and Hall.

Breiman L., Friedman J. H., Olsen E. A., & Stone C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth International Group.

Chadha, P., & Singh, G. N. (2012). Classification rules and genetic algorithm in data mining. *Global Journal of Computer Science and Technology Software & Data Engineering*, *12*(15), 50-54.

Fayyad, U., & Irani, K. (1992) On the handling of continuous-valued attributes in decision tree generation. *Machine Learning*, *8*, 87–102. doi: 10.1007/BF00994007

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process of extracting useful knowledge from volumes of data. *Communications of the. ACM*, *39*(11), 27–34. doi: 10.1145/240455.240464

Han J., Kamber, M., & Pei, J. (2000). *Data mining: Concepts and techniques*. Massachusetts: Morgan Kaufmann.

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning 11*, 63–91. doi: 10.1023/A:102263111

Ibarguren, I., Perez, J. M., Muguerza, J., Gurrutxaga, I., & Arbelaitz, O. (2014). *An update of the J48Consolidated WEKA's class: CTC algorithm enhanced with the notion of coverage*, (Technical Report EHU-KAT-IK-02-14), Spain: University of the Basque Country

Jain, A. K. (2015). *Indian ethnobotany: Emerging Trends*. Jodhpur: Scientific Publisher.

Johari, R. (2016). MS&E 226: "Small" data lecture 8: Classification problems, lucture notes, Retrived from: http://web.stanford.edu/~rjohari/teaching/notes.html

Karypis, G., Han, E. H., & Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Institute of Electrical and Electronics Engineers Computer Society*, *32*, 68-74. doi: 10.1109/2.781637

Kohavi, R. (1995, August) *A study of cross-validation and bootstrap for accuracy estimation and model selection*. Paper presented at the In Proceedings of International Joint Conference on AI. Montreal, Canada.

Kushary, D. (2012). Bootstrap Methods and Their Application, *Technometrics*, *42*(2), 216-217. doi: 10.1080/00401706.2000.10486018

Landis, J. R.,& Koch, G. G. (1977) The measurement of observer aggreement of categorical data. *Biometrics*, *31*(3), 159-174.

Lopez, M. I., Luna, J. M., Romero, C., & Ventura, S. (2012). *Classification via clustering for predicting final marks based on student participation in forums*. Paper presented at the 5th International Conference on Educational Data Mining, Chania, Greece.

Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-Hill.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning.* San Mateo, CA: Morgan Kaufmann.

Padmavathi, J. (2012). Logistic regression in feature selection in data mining. *International Journal of Scientific & Engineering Research*, *3*(8), 1-4.

Perlich, C., Provost, F., & Simonoff, J. S. (2002). Tree induction vs. logistic Regression: A learning-curve analysis. The Journal of Machine Learning Research, *4*(1), 211-255. doi: 10.1162/153244304322972694

Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63. doi: 10.9735/2229-3981

Refaeilzadeh P., Tang L., & Liu, H. (2007). *On comparison of feature selection algorithms*. In Proc. AAAI-07 Workshop on Evaluation Methods in Machine Learing II. Vancouver, British Columbia, Canada, July 2007.

Robu, R., & Hora, C. C. (2012, June). *Medical data mining with extended WEKA*, Paper presented at the IEEE 16th International Conference on Intelligent Engineering Systems (INES), Lisbon, Portugal.

Rokach, L., & Maimon, O. Z. (2008). *Data mining with decision trees: Theory and applications*. Singapore: World Scientific Publishing Co., Inc.

Schwenke, C., & Schering, A. (2007). *True positives, true negatives, false positives, false negatives.* New Jersey, USA: Wiley Encyclopedia of Clinical Trials.

Sinha, A. P., & May, J. H. (2005). Evaluating and tuning predictive data mining models using receiver operating characteristic curves. *Journal of Management Information Systems*, *21*(3), 249-280.

Sumathi, S., & Sivanandam, S. N. (2006). *Introduction to data mining principles*. Berlin: Springer-Verlag.

Şeker, S. E. (2016). *Weka ve veri madenciliği*, Retrieved from https://www.dr.com.tr/ekitap/weka-ile-veri-madenciligi

Weiss, G. M., & Davison, B. (2010). Data mining, In H. Bidgoli (Ed.), *The handbook of technology management* (pp.2-17), New Jersey: John Wiley and Sons.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann.

Witten, I. H., Frank, E., & Hall, M. (2016). *Data minig: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann.

Zupan, B., & Demsar, J. (2008). Open-source tools for data mining, *Clinics in Laboratory Medicine, 28(1)*, 37-54. doi: 10.1016/j.cll.2007.10.002

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(1);96-116

Effects of Students and School Variables on SBS Achievements and Growth in Mathematic*

Emine YAVUZ ** Şeref TAN *** Hakan Yavuz ATAR ****

Abstract
The purpose of this study is to investigate the effects of several student and school level variables on students' mathematical achievement and mathematical achievement growth in middle school. The research population consisted of students in Ankara who started middle school in 2008 and graduated in 2011. Using a non-random typical case sampling method, 3715 students were sampled from 40 middle schools. The data used in this study were obtained from the Ministry of National Education with express written permission. Student Placement raw mathematics subtest scores from 2009 to 2011 were used as the dependent variable in the analysis. "Gender, mathematics grade point average in 6th grade, and school attendance at sixth grade" comprised the student level variables whereas "school type and school size" variables comprised the school-level variables. A three-level hierarchical linear growth model was used to investigate the effects of these variables. Students' raw mathematics subtest scores in Students Placement Test were included in the model after these scores were equated. The results of the model analysis indicated that there was no growth in students' mathematical achievement. It was also observed that school attendance, sixth-grade mathematics grade point average and school type had a statistically significant effect on students' sixth-grade mathematical achievement.

Key Words: Mathematical Achievement, Mathematical Achievement Growth, Three-Level Hierarchical Linear Model, Growth Model, Equipercentile Method

INTRODUCTION

Mathematical achievement and mathematical achievement growth in middle school have an effect on future performance and career of students. Various factors (such as variables related to school and students' own backgrounds, past achievement, gender, parental education, etc.) are sources of information for students' future achievements. This study is aimed at examining the students' mathematical achievement growth as well as the effects of student and school characteristics on the achievement of students in the Student Placement Test (SBS) mathematical. SBS is a test put into practice by the Ministry of National Education (MoNE) in 2007 for the follow-up and evaluation of students' mathematical achievement growth. MoNE carried out the process from the preparation of the SBS questions to the application of the examination and scoring. MoNE prepares a common guideline for the collection of data in similar environments and with a similar system. The guideline is handed over to the examiners, and it is tried to avoid bias arising from the differences between the examiners and their environment. The validity and reliability studies of the data obtained from the exam are done by the respective experts in the MoNE (Milli Eğitim Bakanlığı, 2011).

There are studies in the literature on the use of national like SBS and international (The Programme for International Student Assessment - PISA and The Trends in International Mathematics and

*This study was derived from master's thesis named "Effects of Students and School Variables on SBS Achievements and Growth in Mathematic: An Investigation with Three-Level Hierarchical Linear Growth Model" completed on 26.10.2015 under the guidance of Prof. Dr. Şeref TAN.

** Research assisstant, Erciyes Üniversity, Ziya Eren Faculty of Education, Kayseri-Turkey, yavuzemine0@gmail.com ORCID ID: 0000-0002-1991-1416
*** Professor, Gazi University, Gazi Faculty of Education, Ankara-Turkey, sereftan4@yahoo.com ORCID ID: 0000-0002-9892-3369
**** Associate Professor, Gazi University, Gazi Faculty of Education, Ankara-Turkey, hakanyavuzatar@gmail.com ORCID ID: 0000-0001-5372-1926

To cite this article:
Yavuz, E., Tan, Ş., & Atar, H., Y. (2019). Effects of students and school variables on SBS achievements and growth in mathematic. *Journal of Measurement and Evaluation in Education and Psychology, 10*(1), 96-116. DOI: 10.21031/epod.493297

Received: 02.12.2018
Accepted: 20.03.2019

Science Study - TIMSS) test data applied to large samples in determining student and school variables that affect mathematical achievement (Altun, 2007; Karabay, 2012; Ötken, 2012; Özdemir, 2010; Reçber, 2011; Savaşçı, 2011 & Yılmaz, 2006). Multiple regression analysis was used to answer research questions without taking the sample structure into consideration in most of these studies. PISA, TIMSS, ÖBBS and SBS data are obtained from large samples, and these data often have multilevel structures. Students, for instances, are nested in classrooms, classrooms in schools and so forth. Students in the same class or school in this data structure show more similar characteristics to each other than students who are selected at random from the class. For this reason, it cannot be said that the observations obtained from the individuals in the same social unit are completely independent of each other (Raudenbush & Bryk, 2002). Furthermore, some schools are more homogeneous based on certain features (e.g., socioeconomic status, region, etc.) while others are heterogeneous. This means that the assumption of equality of variances is not achieved in large samples (Hox, 2010, pp. 4-7; Raudenbush & Bryk, 2002). Also, for this reason, the assumption of independence of observations and homogeneity of variances is not achieved. In addition, multiple regression analysis can yield biased results since multilevel data structures are not considered in large sample trials (Raudenbush & Bryk, 2002, pp. 5-99).

Another method of analysis used in the literature to study the effect of student and school variables on mathematical achievement is the ordinary least squares (OLS, linear least squares) method. However, according to the relevant literature, the OLS regression methods underestimate standard errors compared to the multilevel model (Hox, 2010). Underestimation of standard errors increases the probability of Type I error in the estimation of regression parameters, which is not desirable (Hox, 2010). Furthermore, the OLS regression poses a problem in interpretation, such as bias of aggregating the individual-level variables to the higher level and the determination of heterogeneity between schools (Raudenbush & Bryk, 2002, p. 253). As indicated above, multiple regression and OLS methods yield biased results due to aggregation bias, underestimation of standard errors, and regression heterogeneity (Bryk & Raudenbush, 1988). To avoid these biased results, multilevel analysis models should be used to analyze the data from multilevel samples (Atar, 2010; Atar & Atar, 2012; Demir & Kılıç, 2010; Güzel, 2006; Sevgi, 2009). In essence, multilevel models are generalized regression methods and can be used for causal interpretation, data reduction, and various estimation purposes (Hox, 2010).

This study is aimed at examining the students' mathematical achievement growth and the effects of student and school characteristics on students' SBS mathematical achievement. Growth in the study refers to the change in the SBS mathematics scores of the students that they received over the years. Based on the fact that the data structure is multilevel, three-level hierarchical linear growth model was used in this study to avoid the bias of single-level analysis methods for this data. In light of this, level-1 consists of a repeated measure of students' SBS mathematics scores, level-2 consists of students' characteristics variables and finally level-3 consists of school characteristics variables. After a thorough review of Turkish literature, no study was found examining the factors that affect mathematical achievement and mathematical achievement growth by a three-level hierarchical model. However, there are many studies using three-level hierarchical linear growth models abroad (Ding, Song & Richardson, 2010; Huang et al., 2009; Raymond, 2009; Shapley et al., 2011; Shim, 1995; Wu, 2004; Yang, 2000; Zhu, 1998). In addition, when the literature was examined, there were many studies examining student, school and country variables as they affect the mathematical achievement abroad with three-level hierarchical linear model (HLM) (Agodini & Harris, 2016; Kao et al., 2017; Tan & Hewer, 2018; Yi & Shin, 2018). It was found that only two studies in Turkey (Aztekin & Yılmaz, 2014; Çelik, 2016) were found to be relevant. Çelik (2016), only the studies the effects of the variables of country while Aztekin and Yilmaz (2014) examined the effects of variables of students, schools, and countries. In addition, the effect of some variables in PISA and TIMSS mathematical achievement was examined in these two studies. Unlike the international large scale exams that reflect international committees' goal and objectives, SBS reflects the goals and objectives of the Turkish education system. This study is important because it examines students' mathematical achievement growth in SBS and the effects of some student and school variables on

SBS mathematical achievement in the middle school in the Turkish education system. Studies conducted in Turkey mostly examine the effect of variables such as gender, school attendance and teacher qualifications on mathematical achievement using one or two level models. Although there seems to be a growing interest in the use of three-level models in the analysis of multilevel data in the Turkish literature, they are still too few to draw valid conclusions (Çelik, 2016). It is thought that this study will be an example of a three-level hierarchical growth model implications in Turkish literature.

Purpose of the Study

The purpose of this study is to investigate the effects of several student-level variables (gender, grade point average and school attendance in sixth grade) and school-level variables (school type and school size) on students' mathematical achievement and mathematical achievement growth in middle school using three-level hierarchical linear growth model. For this purpose, the following research questions were investigated:

1) Do the sixth-grade students' SBS mathematical achievements vary at the student level and school level? If yes, what percentage of variance is accounted for at each level?

2) Is there any growth in students' SBS mathematics raw scores in middle school from 2009 to 2011?

3) Do the sixth-grade students' SBS mathematical achievements vary according to student-level variables (Gender, grade point average in sixth grade-GPA6 and attendance at school in sixth grade-ABSENTEEISM)? If yes, what percentage of variance is accounted for the student-level variables?

4) Do the sixth-grade students' SBS mathematical achievements vary according to school-level variables (school size and type)? If yes, what percentage of variance is accounted for by the student-level variables?

METHOD

This research is a causal comparison model of quantitative research methods (Büyüköztürk, et al., 2008) as it examines the effects of variables related with students and schools on students' mathematical achievement and mathematical achievement growth in middle schools.

Sample

The research population consisted of students in Ankara who started middle school in 2008 and graduated in 2011. One of the aims of this study is to get detailed information about the universe by choosing a typical situation which is thought to represent the universe. Therefore, a typical sampling method (Büyüköztürk et al. 2008, p. 91) was used to determine a sample of middle schools in the province of Ankara and the research was conducted on the data of these schools. The data used in the study were obtained as a result of correspondence with the Ministry of National Education (MONE). So, 40 schools from the middle school universe in Ankara and 3733 students who were educated in these schools were randomly assigned in the years related to codes created by the Ministry of National Education. While the school sample consists of 39 public schools and 1 private school, a worthy observation made is that the average number of these schools in the relevant years varies between 12 and 255. The existence of one private school in the sample shows the limitation of the research in terms of school diversity. The research sample was first composed of 3733 students. However, it was decided that the outliers of the data – 18 students – should not be included in the analysis because the sample was large enough. The research sample takes its final form as 3715 students in 40 middle schools in Ankara who started in middle school in 2008 and graduated in 2011. It has been taken into consideration that the students who are in the sample group should not have changed school during their education. The research sample is suitable for HLM analysis because at

least two repetitive measurements are used for the study of change, and students (3715 persons) are placed at the second level, and the number of units in the third level is left to the researcher's ability to reach the database.

Data Collection Instrument

The aim of the study was to examine the variables affecting the achievement of the students and the changes in their achievement. Repeated measures are needed to examine the students' achievement growth. In the Turkish education system, these repeated measures were first obtained with SBS, and it is thought that there will be similar exam applications for follow-up examinations. In this study, raw mathematics scores obtained from the students' SBS were used to examine mathematical achievement. These raw scores were included in the analysis as dependent variables.

Level-1 variables (Dependent variable)

The sixth, seventh and eighth grade SBS mathematic subtest data belonging to the same students are raw scores. The raw scores are calculated by applying correction formula to the number of items answered correctly. They were included as a continuous variable in the HLM analysis and considered as dependent variables. The aim is to determine the change or students' achievement growth by examining these points together.

Level-2 variables (Student level variables)

Gender (GENDER), absenteeism (ABSENTEEISM6) and year-end mathematics grade point averages of students for the 2008-2009 academic year (GPA6).

Year-end mathematics grade point averages are based on the average grade that the students took in two semesters in mathematics lessons in related years and included as a continuous variable in this study. Gender is a dummy variable; male students are coded as "0" and female students are coded as "1". The absenteeism is the number of days during which the students did not come to school in the related education years. It was regarded as a continuous variable. While gender is a constant variable that does not change over time, the absenteeism and year-end mathematics grade point averages of students are the variables that change over time (year to year). Since the aim is to examine the effects of the initial state (grade point average in 6^{th} grade) of students on mathematical achievement and mathematical achievement growth in this study, time-varying variables (year-end absences and year-to-year mathematics grade point averages) are included as fixed variables.

Level-3 variables (School level variables)

School size (SIZE) and type (TYPE) variables are included as third level variables in the analysis. School size refers to the average number of students per school, and it has been considered as a continuous variable. School type is a dummy variable in which public schools are coded as "0" and private schools are coded as "1".

Data Analysis

Providing equivalency of SBS scores: Equipercentile method

In order to determine the students' mathematical achievements growth, it is necessary to monitor their mathematical achievements over the years. For this purpose, it is necessary to compare the SBS mathematics subtest scores of the students. The procedures for comparing student achievements are insufficient because of the differences in the number of items used in SBS exams, their reliabilities,

and difficulties. For this reason, it is necessary to convert test scores to the same score scale before including SBS raw scores in the HLM analysis. Middle mathematics curriculum has a helical structure (Ersoy, 2006) and the contents of SBS mathematics subtests are compatible with the curriculum of the related years (MEB, 2011); therefore, the structure of SBS mathematics subtests was thought to be similar and the test scores equated before the HLM analysis. Equivalence of SBS scores in this study is provided by equipercentile equation method.

The reliability of the test scores must be high in order to use equipercentile method (Schneider & Dorans, 1999). The reliability coefficients of KR-21 for the sixth, seventh and eighth grade SBS raw scores used in the study were 0.86; 0.90 and 0.92, respectively. Thus, the high reliability of the raw scores is enough to do the equation study.

The extreme values in the data set consisting of 3733 students were determined before the equipercentile method was started. So, the Z scores of the SBS raw scores were calculated, and the scores (18 students) belonging to the students who were out of the range of ± 4 were discarded from the data set (Çokluk, Şekercioğlu & Büyüköztürk, 2014, p. 14). In addition, missing values were taken as "0". Raw scores are calculated by applying correction formula to the number of items answered correctly. Scaled scores were analyzed in the Rage3.15 program. Since the sample is not very large and the four moments of the forms are not very close to each other as a result of the equipercentile method, smoothing was done. Then, beta4 value and chi-square difference values were examined and it was decided to perform polynomial pre-smoothing at C = 4. Smoothed equivalent scores were used in the HLM analysis.

HLM analysis: Three-level hierarchical linear growth models

Hierarchical linear models are the generalization of the regression methods used for various purposes such as causal interpretations, various estimates, and data reduction (Raudenbush & Bryk 2002). HLM7 software was used for conducting HLM analysis in this study. This analysis starts with a completely unconditional model, then continues with a base model where level-1 is simple linear growth. Finally, intercept and slope of level-1 becomes dependent variables at level-2 and level-3. There was not any growth in this analysis; the only intercept of level-1 becomes dependent variables at level-2 and level-3. In addition, student variables are added to the base model to answer research question three and variables related to school are added to the base model to answer research question four.

Completely unconditional model: As the name implies, there are no explanatory variables in the model except for the intercept coefficients. This model provides a basis for the definition of the level-1 model and useful evidence for the development of basic statistics for the evaluation of other models (Heck and Thomas, 2009, p. 173). In addition, this model used for checking the proportion of total variance in the outcome can be explained by group membership (e.g., with ICC). The model equation is:

Level-1: $Y_{tij} = \pi_{0ij} + e_{tij}$

Level-2: $\pi_{0ij} = \beta_{00j} + r_{0ij}$

Level-3: $\beta_{00j} = \gamma_{000} + u_{00j}$

Combining model: $Y_{tij} = \gamma_{000} + u_{00j} + r_{0ij} + e_{tij}$

In this study, the subscripts which are t, i, j represent time (test taking year: 2009, 2010, 2011), student and school, respectively.

Y_{tij}*:* SBS mathematics score at *t* time of student-i in school-j.

π_{0ij} : The intercept coefficient indicating the estimated initial state (mathematical achievement in the sixth grade) of the student-i in school-j

β_{00j} : mean score of SBS mathematics scores of students in the sixth grade in school-j

γ_{000}: The general mean of mathematics score (general average) for all students in the 2009 SBS (sixth grade SBS)

e_{tij}: Effect of repeated measurements at *t* time, first order random effect. Errors are independent and normally distributed. The mean is "0" and the constant variances are σ^2 (Raudenbush & Bryk, 2002, p. 162).

r_{0ij}: Second-level intercept (β_{00j}) effect. So the difference between mathematics score of student-i and the school-j 's average mathematics score, the deviation amount. The mean is "0" and variance is τ_{π}.

u_{00j}: Third-level intercept effect. The difference between the average mathematics score of the school-j and the general mathematics score is the amount of deviation. In other words, it is the unexplained difference in the average mathematics scores of schools. The mean is "0" and the variance is τ_{β}.

Level-1 model (Unconditional model for growth): The level-1 model was used to answer the second research question. In the demonstration of individual growth in the HLM, more polynomial curves are used because they are flexible and predict standard linear modeling procedures. Since there were three repeated measurements belonging to the students in this study, linear development model was used in the analysis of the data.

In the study, "class" (encoded as 0, 1, 2) was chosen as the time unit. The time points in the measurement (the time periods during which the exams were applied) are the 6th grade (time = 0), 7th grade (time = 1) and 8th grade (time = 2). For this reason, the intercept parameter (π_{0ij}) interpreted as the actual starting status at time point a_{it}=0 of the student i. The time point a_{it}=0 is the sixth grade SBS. Interpretation of the intercept as the real starting point forms the basis for interpreting the development of the students in the sample over time. The slope is interpreted as the annual rate of change of mathematical achievements of students, which is represented by the growth curves and is formed by time effect. Each student will have their own growth curves shaped by the raw mathematics scores they have achieved over three years. The model equations are as follows:

Level-1: $Y_{tij} = \pi_{0ij} + \pi_{1ij}*(GRADE_{tij}) + e_{tij}$

Level-2: $\pi_{0ij} = \beta_{00j} + r_{0ij}$
$\pi_{1ij} = \beta_{10j} + r_{1ij}$

Level-3: $\beta_{00j} = \gamma_{000} + u_{00j}$
$\beta_{10j} = \gamma_{100} + u_{10j}$

π_{1ij}: Coefficient indicating the rate of growth of the student-i during the specified academic year (2008-2009, 2009-2010 and 2010-2011 academic year) in school-j.

$GRADE_{tij}$: Mathematical achievement of the student-*i* at *t* time.

β_{10j}: Mean growth rates of students in school-j

r_{1ij}: Second-level slope (β_{10j}) effect. The difference between the rate of mathematics score of student i and the school j's average rate of mathematics score; the deviation amount. The mean is "0" and variance is τ_{π}.

γ_{100}: Mean of all students' growth rates in mathematical achievement

u_{10j}: Third level slope effect. That is, the difference between the growth curve of the average mathematical achievement of the school j and the growth curve of the general mathematical achievement is the amount of deviation. In other words, it is the unexplained difference in average growth rates of schools. The mean is "0" and the variance is τ_{β}.

After unconditional model analysis, conditional models (second and third-level models) were applied in three-level hierarchical linear growth model analysis.

*Level-2 model***:** The level-2 model was used to answer the third research question. Second-level variables allow for differences between individuals and groups. This model contains the explanatory variables (second level variables) of the unexplained variance at the first level intercept and slope. The second level model equation is given below:

Level-1: $SBS_{tij} = \pi_{0ij} + \pi_{1ij}*(GRADE_{tij}) + e_{tij}$

Level-2: $\pi_{0ij} = \beta_{00j} + \beta_{01j}*(ABSENTEEISM6_{ij}) + \beta_{02j}*(GPA6_{ij}-\overline{GPA6.ij})+r_{0ij}$

$\pi_{1ij} = \beta_{10j}$

Level-3*:* $\beta_{00j} = \gamma_{000} + u_{00j}$

$\beta_{01j} = \gamma_{010}$

$\beta_{02j} = \gamma_{020} + u_{02j}$

$\beta_{10j} = \gamma_{100}$

($ABSENTEISM6_{ij}$): Student-i's sixth-grade attendance in school-j (The number of days when the learner does not come to the school)

β_{01j}*: ABSENTEEISM6* effect; effect of students' sixth-grade attendance on mathematical achievement at sixth grade in school-j

($GPA6_{ij}-\overline{GPA6.ij}$): year-end mathematics sixth-grade point averages of student-i in school-j (display of centered around group average)

β_{02j}*: GPA6* effect; effect of students' year-end mathematics sixth-grade point averages on mathematical achievement at sixth grade in school-j

γ_{010}: Average effect of all of the students' sixth-grade attendance on mathematical achievement at sixth grade

The result of the level-1 analysis shows that the growth rates of the students are not statistically significant. For this reason, the π_{1ij} coefficient was included as a fixed variable in the level-2 analysis. However, since the π_{1ij} coefficient was included as a constant variable in the analysis, the residuals r_{1ij} and u_{10j} were not included in the model equation. At the second level, students' attendance in the sixth grade (ABSENTEEISM6) and year-end mathematics grade point averages (GPA6) were included. The significance of residuals of the variables has been examined before level-2 analysis. As a result of the study, it was decided that ABSENTEEISM6 variable should be included as a fixed constant and (GPA6) variable should be included as a random variable. Since ABSENTEEISM6 variable is fixed, the value u_{01j} is not included in model equality.

Level-3 model: This model was used to answer the fourth research question. Third level parameters describe the distributions of the average mathematical achievement and mathematical achievement curves as a function of school-level variables (school size and type). The third level model equation is as follows:

Level-1: $SBS_{tij} = \pi_{0ij} + \pi_{1ij}*(GRADE_{tij}) + e_{tij}$

Level-2: $\pi_{0ij} = \beta_{00j} + r_{0ij}$

$\pi_{1ij} = \beta_{10j}$

Level-3: $\beta_{00j} = \gamma_{000} + \gamma_{001}(TYPE_j) + u_{00j}$

$\beta_{10j} = \gamma_{100}$

As in the level-2 model, since the π_{1ij} coefficient was included as a fixed variable in the model, residual values of r_{1ij} and u_{10j} were not included in the model equality. Here, other than the coefficients in the first and second level model equations, the coefficient γ_{001} is explained.

γ_{001}: Effect of school mean on the general mean.

Reliability of HLM models: Assumptions were checked before starting HLM analysis. Assumptions of the three-level hierarchical linear development model are as follows: a) Metric; the dependent variable is measured over time on a general scale. b) The shape of the change is linear; change/growth increases with constant intervals. c) Distribution of errors by mean "0"; independent, normal distributions of errors with constant variance. d) Covariance structure; each variable is unrelated to its own level and another level of error (Raudenbush & Bryk, 2002, p. 255). Meeting necessary assumptions prevents bias results in HLM analysis. In this study, all the variables meet the assumptions.

As a result of the completely unconditional model analysis, the reliability of level one (π_0) and two (β_{00}) intercept coefficients were estimated as 0.82 and 0.91 respectively. This can be interpreted based on the fact that the data included in the analysis were obtained with sufficient reliability for the estimation of the mathematics averages. In addition, the deviation value for this completely unconditional model was found as 83347.167906. Estimated number of parameters was 4.

As a result of the first-level model analysis, the reliability coefficients for the intercept point (π_0, β_{00}) were estimated as 0.40 and 0.82, respectively. The reliability coefficients of slopes (π_1, β_{10}) were estimated as 0.003 and 0.35, respectively. Here, the low-reliability coefficients of the slopes are due to the slopes not being as constant as the intercept. Very low reliability in HLM analysis does not indicate that the HLM analysis is invalid. Low reliabilities indicate that the variable needs to be fixed in a top model because the variable has a really small variance, or the corresponding variances are hardly sampled (Raudenbush & Bryk, 2002). Raudenbush and Bryk (2002) also stated that they could be regarded as reliable if their reliability values are above 0.05.

In this study, the deviation value estimated for the completely unconditional model was found as 83347.167906 while the estimated deviation value for the level-1 growth model was found as 83341.505305. The estimated number of parameters is 9. It can be said that level-1 growth model did not fit well because the difference between the deviation values (5.66) is not more than twice the chi-square value (11.07), which is the difference between the estimated parameter numbers in the freedom degree models (5). As a result of the HLM growth model analysis, the inability to predict development supports this situation.

As a result of the analysis of the level-2 model, the reliability of the level-1 intercept coefficient was increased by 0.14 and estimated as 0.54. The reliability of the second-order intercept coefficient was increased by 0.14 and estimated as 0.96. In this study, the deviation value estimated for the completely unconditioned model was 83347.167906 while the estimated deviation value for the level-2 model was 79883.176924. The estimated number of parameters was 9. It can be said that the difference between the deviation values (3463.99) was better than that of the level-2 model. This is the case because the difference between the estimated number of parameters in the degree-of-freedom models (5) was at least twice as large as the chi-square value (11.07).

As a result of the third level analysis, the reliability of the intercept coefficient was estimated as 0.82 for the level-1 and 0.85 for the level-2. The deviation value estimated for the completely unconditional model was 83347.167906 while the estimated deviation value for the level-2 model was 83325.874254. The estimated number of parameters was 6. The difference between the deviation values (21.29) was at least twice as high as the chi-square value (5.99), which was the difference between the estimated parameter numbers in the freedom rank models (2).

Finally, the real ranges of 95% of the true values of the constant effect coefficients were estimated. A formula $\pm 1{,}96\sqrt{\tau_\beta}$ was used to estimate the 95% confidence interval of a true value of the coefficient (Raudenbush & Bryk, 2002, p. 55). Estimated coefficients stand between the confidence intervals generated for the variables, which is another evidence for the reliability of estimates.

RESULTS

Findings related to the first question: "Do sixth-grade students' SBS mathematical achievement vary students and schools? If yes, what percentage of variance accounted for at each level?"

For this question, a completely unconditional model was analyzed and the result of the analysis are reported in Table 1. Here, the intercept coefficient is interpreted as the mathematical achievement in the sixth grade of the student. When Table 1 is examined, it is seen that the general mean (γ_{000}) is estimated as 32.65 with a standard error of about 0.70. When a 95% confidence interval is created around the general mean, it is expected that the true value of the general mean will be within the range of 31.28 to 34.02 (%95CI(γ_{000}) = 32.65± (1.96)(0.70)).

Table 1. Completely Unconditional Model Analysis Results

Fixed Effect	Coefficients	Standard error	*t*	Approximate *d. f*
Intercept, π_0				
Intercept2, β_{00}				
Intercept3, γ_{000}	32.65*	0.7	46.66	39
Random Effect	**Standard Deviation**	**Variance components**	***d. f***	$\chi 2$
Intercept1,r_0	9.48*	89.92	3675	20920.02
Level-1, *e*	7.58	57.49		
Intercept1/ Intercept2,u_{00}	4.21*	17.75	39	559.55

*p<0.05

Furthermore, if 95% confidence interval is formed around the general average, it is expected that 95% of the average mathematics raw scores of the students will be in the range of 14.07 to 51.23($\widehat{\gamma 00} = 32.65 \overline{+} (1.96)\sqrt{89.92}$). Likewise, it is expected that 95% of the average mathematical achievements of the schools will be between 24.40 and 40.90 ($\widehat{\gamma 00} = 32.65 \overline{+} (1.96)\sqrt{17.75}$).

As seen in Table 1, the estimated value of variability in inter-student level (*e*) is 57.49, the estimated value of variability in intra-student level, the variance of (r_0) is 89.92 and the estimated value of variability in school level (u_{00}) is 17.75. It can be said that there are statistically significant differences in means of mathematical achievements of students and general means of mathematics scores of schools since the *p*-values of coefficients are smaller than 0.001 alpha level. Furthermore, this indicates that it is necessary to establish a three-level hierarchical linear growth model (p<0.001, $d.f$= 3675) and schools (p<0.001; $d.f$ = 39). In addition, if the shared variance ratio in the upper levels is greater than 0.10, the multi-level analysis is allowed to continue (Lee, 2000). The extent to which the levels explain the variance in the SBS mathematics scores is calculated by means of interclass correlation (ICC) (Raudenbush & Bryk, 2002, p. 230):

Test (intra-student) level ICC: $\hat{p} = \frac{57.49}{57.49 + 89.92 + 17.75} = 0.35$

Inter-student level ICC: $\hat{p} = \frac{89.92}{57.49 + 89.92 + 17.75} = 0.54$

School level ICC: $\hat{p} = \frac{17.75}{57.49 + 89.92 + 17.75} = 0.11$

By calculating the interclass correlation, it was seen that the majority of the variance in mathematical achievement (0.54) could be explained by the student level, then the test (intra-students) level and at least the school level. It can be stated that the variables that most affect the students' mathematical

achievement are the characteristics related to them, the characteristics related to the tests and the characteristics related to the schools, respectively.

Findings related to the second question: "Is there any growth in students' SBS mathematics raw scores in middle school from 2009 to 2011?"

To answer this question, the first level model (unconditioned model for growth) was analyzed. Intercept and GRADE variables are included as random variables to the model. Since the SBS raw scores are taken with their equivalence to HLM analysis, the raw scores for mathematics range between 0 and 78. The level-1 (growth model) analysis results are given in Table 2. The growth rate of students' mathematical achievements is estimated as γ_{100} = 0.03 over the course of time. In other words, the average growth rate of the students in each year shows an increase of 0.03. The *p* value of γ_{100} coefficient is not statistically significant (p>0.05; $d.f$ = 39), so the change in the students' mathematical achievements can be described as a small coefficient that can be explained by sampling error. In other words, it can be said that the inclination coefficient is not statistically significant because the students' development in mathematical achievement in the three education-training processes is too small to estimate the difference.

Table 2. Level-1 (Growth Model) Analysis Results

Fixed Effect	Coefficients	Standard error	*t*	Approximate *d. f*
Intercept1, π_0				
Intercept2, β_{00}				
Intercept3, γ_{000}	32.61*	0.69	47.08	39
GRADE slope, π_1				
Intercept2, β_{10}				
Intercept3, γ_{100}	0.03	0.12	0.22	39
Random Effect	**Standard Derivation**	**Variance components**	***d. f***	χ^2
Intercept1,r_0	9.38*	87.87	3675	5534.12
GRADE slope, r_1	0.28	0.08	3675	3035.41
Level-1, *e*	7.57	57.25		
Intercept1/ Intercept2,u_{00}	4.00*	16.03	39	281.71
GRADE/ Intercept2,u_{10}	0.44*	0.19	39	62.93

*p<0.05

Findings related to the third question: "Do sixth-grade students' SBS mathematical achievement vary student-level variables (Gender-GENDER, grade point average in 6th grade-GPA6 and attendance at school in 6th grade-ABSENTEEISM6)? If yes, what percentage of variance accounted for the student-level variables?"

When an exploratory analysis was performed before starting HLM analysis, it was seen that the most important variables were ABSENTEEISM6 and GPA6. For this reason, the second level model was analyzed by adding "ABSENTEEISM6 and GPA6" variables to answer this problem. The variables were added only to the intercept coefficient because the students' progress was not statistically significant. The significance of the residuals of variables in the analysis has been examined. As a result of the study, it was decided that the variable ABSENTEEISM6 should be taken as the model constant while the variable GPA6 should be taken as the random variable. Furthermore, GPA6 was centered on the group average and tried to avoid possible multiple-connection problems. Since the residual variance in the second level of the "GRADE" variable does not make statistical sense, this variable is kept constant at the second and third levels. Two different tables were created for more favorable reporting of fixed and random effects.

The results of the level-2 fixed effects analysis are given in Table 3. When Table 3 is examined, it can be seen that the coefficient of change of the variable ABSENTEEISM6 (γ_{010}) is estimated as -0.05 with a standard error of about 0.02. When 95% confidence interval is established, the actual value of the variable ABSENTEEISM6 is expected to be in the range of -0.09 to -0.01 (%95CI(γ_{000}) = -0.05± (1.96)(0.02)). The *p* value of the coefficient γ_{010} was examined to determine whether the effect of the variable ABSENTEEISM6 on the general mean is different from zero. The H_0 hypothesis was rejected because the *p* value of the coefficient is statistically significant (p <0.05; d. f = 3634). In other words, the mathematical achievement of the student who goes to school regularly is 0.05 units more than the student who is absent from school in sixth grade.

Table 3. Level-2 Fixed Effect Analysis Results

Fixed Effect	Coefficients	Standard error	*t*	Approximate *d.f.*	Effect size
Intercept1, π_0					
Intercept2, β_{00}					
Intercept3, γ_{000}	32.92*	0.73	45.17	39	---
ABSENTEEISM, β_{01}					
Intercept3, γ_{010}	-0.05*	0.02	-2.27	3634	-0.005
GPA6, β_{02}					
Intercept3, γ_{020}	0.42*	0.02	27.18	39	0.04
GRADE slope, π_1					
Intercept2, β_{10}					
Intercept3, γ_{100}	0.006	0.09	0.07	7349	---

*p<0.05

When Table 3 is examined, it is seen that the GPA6 coefficient (γ_{020}) is estimated as -0.42 with a standard error of about 0.02. When 95% confidence interval is established, the true value of GPA6 variable is expected to be in the range of 0.38 to 0.46 (%95CI(γ_{000}) = 0.42± (1.96)(0.02)). To determine whether the effect of GPA6 variable on the general mean is different from zero, the *p* value of γ_{020} coefficient is examined. The H0 hypothesis was rejected because the *p* value of the coefficient was statistically significant (p<0.001, *d.f* =39). In other words, the mathematical achievement of a student who has high year-end mathematics grade point average is higher than that of a student who has low year-end mathematics grade point average in the sixth grade. It can be said that the common effect of GPA6 variable is reported here. In other words, this variable has different effects on students in different schools.

When random effects are examined in Table 1, it is seen that level-1 intercept variance (r_0) is estimated as 89.92. When the level variables are added to the level-2 model, this variance is estimated as 22.78 (Table 4). With the proportion of the difference between the two variances to the variance in the growth model explained, the variance of student level by the student level variables was calculated. The student-level estimates of the ABSENTEEISM6 and GPA6 variables explained 0.75 of student-level variance. Given that the student-level explains 0.54 of the variance in mathematical achievement, these variables explain 0.41 of the variance in mathematical achievement. When the effect of sizes of the variables of GPA6 and BASARIORT6 were examined, it was observed that the effects on mathematical achievement were too small to be felt in daily life (Ferguson, 2009).

Table 4. Level-2 Random Effect Analysis Results

Random Effect	Standard Deviation	Variance components	*d. f*	χ^2
Intercept1,r_0	4.77*	22.78	3634	7942.70
Level-1, *e*	7.58	57.49		
Intercept1/ Intercept2,u_{00}	4.30*	18.48	39	1435.93
Intercept1/GPA6,u_{02}	0.09*	0.008	39	218.86

*p<0.05

To determine whether the variance of mean mathematical achievements of students is different from zero, the *p* values of the coefficient (r_0) are examined. The H_0 hypothesis is rejected because the coefficient *p* is less than 0.001 alpha. In other words, some of the variance in student-level mathematical achievement remained unexplained.

Findings related to the fourth question: "Do sixth-grade students' SBS mathematical achievement vary school-level variables (school size-SIZE and type-TYPE)? If yes, what percentage of variance accounted for the student-level variables?"

It was seen that the most important variable was TYPE when an explanatory analysis was performed before starting HLM analysis. For this reason, the third level model was analyzed by adding only the "TYPE" indicator to the third level in order to answer this problem. The variables were added only to the intercept coefficient because the students' progress was not statistically significant. Two different tables were created for more favorable reporting of fixed and random effects. The results of the level-3 fixed effects analysis are given in Table 5.

Table 5. Level-3 Fixed Effect Analysis Results

Fixed Effect	Coefficients	Standard error	*t*	Approximate *d.f.*	Effect size
Intercept1, π_0					
Intercept2, β_{00}					
Intercept3, γ_{000}	32.04*	0.62	51.80	38	---
TYPE, γ_{001}	13.45*	2.36	05.70	38	3.19
GRADE slope, π_1					
Intercept2, β_{10}					
Intercept3, γ_{100}	0.006	0.12	0.05	7389	---

*p<0.05

When Table 5 is examined, the TYPE variable coefficient (γ_{001}) is estimated as 13.45 with about 2.36 standard error. When the 95% confidence interval is established, the actual value of the TYPE variable is expected to be in the range of 8.82 to 18.08 (%95CI(γ_{000}) = 13.45± (1.96)(2.36)). The *p* value of γ_{001} coefficient was examined to determine whether the effect of the TYPE variable on the general mean is different from zero. The H_0 hypothesis was rejected because the *p* value of the coefficient is statistically significant (p<0.001, $d.f$= 38). In other words, the mathematical achievement in the sixth grade of the student who is attending private school is 13.45 which is more than the attendance in the public school.

When the random effects are examined from Table 1, it is seen that the second level intercept variance (u_{00}) is estimated as 17.75. When the level variables are added to the third level model, this variance is estimated as 9.97, Table 6. With the proportion of the difference between the two variances to the variance in the growth model explained, the variance of school level by the school level variables was calculated. The TYPE variable explained 0.44 of the school-level variance. Given that the school level explains 0.11 of the variance in mathematical achievement, the TYPE variable accounts for 0.05 of the variance in the mathematical achievement. When the effect of the size of the TYPE variable was examined, it was observed that the variable had a large effect on size for the school level (Ferguson, 2009).

Table 6. Level-3 Random Effect Analysis Results

Random Effect	Standard Deviance	Variance components	*d. f*	χ^2
Intercept2,r_0	9.48*	89.89	3675	20920.04
Level-1, *e*	7.58	57.49		
Intercept1/ Intercept2,u_{00}	3.16*	9.97	38	414.31

*p<0.05

The p-values of the (u_{00}) coefficient have been examined to determine whether the variance of mean mathematical achievements of schools is different from zero. The H_0 hypothesis is rejected because the coefficient *p* is less than 0.001 alpha value. In other words, the variance in mathematical achievement at school level remained unexplained. To explain the remaining variance, the analysis should be repeated by adding different demographic variables to the model.

DISCUSSION and CONCLUSION

The aim of this study is to determine the effects of variables related with students (gender, year-end grade point, school attendance) and schools (type and size) on SBS mathematical achievement and its growth by three-level HLM growth model. It is believed that using the three-level HLM growth model will remove the bias of single-level analysis. Aside from this, this will also serve as an example in the Turkish literature of three-level hierarchical growth model applications. The variables examined in this study, composed of some data that are collected and stored as fixed by the MoNE for all schools in Turkey every year. In this study, the SBS test is applied only to middle school students and in those years the middle schools are three years, so a student has no more than three repeated measures. In addition, vertical scaling was not found to be suitable for the structure of the SBS scores obtained. For this reason, the equivalence of the scores has been established by applying an equal percentage method instead of vertical scaling. The equivalents of the scores were used in the analysis of the research. The research findings were discussed within the framework of these limitations.

For studies looking at the effects of student and school characteristics on mathematical achievement, it is expected that most of the variance in mathematical achievement will be explained by student characteristics (Odden, Borman & Fermanich, 2009; Zvoch & Stevens, 2003). In this study, it was observed that the majority of students' variance in the sixth grade of mathematical achievement was explained by student characteristics, followed by intra-student (test) characteristics and finally by school characteristics. This can be substantiated by the fact that the vast majority of the variables affecting mathematical achievements are student-level, or that student-level variables have a large influence on predicting students' mathematical achievement. Turhan, Şener and Gündüzalp (2017) examined 39 studies related to school effectiveness and found that schools had less impact on student achievement than other factors (students, parents). Similar findings were obtained in Akyüz (2014), Aydın (2015), Sevgi (2009), Tavşancıl and Yalçın (2015). In this respect, it can be said that the effects of these characteristics on the academic achievement of students need to be studied on a bigger scale. It is also seen that the variance ratio which can be explained by intra-student (test) characteristics is also high. However, this level of variance remains unexplained because the variables related to the tests were not included in the model in this study. In the later models, student and school level variables were included in the model in an attempt to explain the variance in mean mathematical achievement.

In this study, unlike many studies (Ai, 1999; Ding vd., 2010; Green, 1995; Huang, et al., 2009; Raymond, 2009; Shapley et al., 2011; Shay, 2000; Shim, 1995; Wu, 2004; Yang, 2000; Zhu, 1998; Zvoch & Stevens, 2006) using a three-level linear growth model, the mathematical achievements growth of students could not be estimated. It is thought that the reason why any growth has not been observed in this study might be due to the correlation between the tests. The correlation coefficient between the tests in the studies involving equated test scores is expected to exceed 0.87 (Schneider & Dorans, 1998 as cited in Dorans, 1999). However, the correlation coefficients between SBS scores belonging to students in this study vary between 0.62 and 0.72. The low correlation between the test scores reduces the comparability of test scores (Schneider & Dorans, 1998 as cited in Dorans, 1999). In this context, it can be said that the different designs of the tests (such as the fact that the tests given to the pupils according to years have different numbers of items and that there is no anchor item in the tests) according to the class levels negatively affect the comparability of SBS achievement scores at different grade levels. Another reason for the non-observation of growth may be related to the data structure. The data used in the study were mostly obtained in a rather cumulative manner. This prevented the modeling of the correspondence between the answers given to the items. By modeling the answer pattern, the equivalence between the tests becomes more

sensitive (Kolen & Brennan, 2014). By providing a more sensitive equivalent of the mathematics test scores at different grade levels with vertical scaling, it is thought that small differences in mathematical achievement of students can be modeled.

Similar to the study of Zvoch and Stevens (2006), in this study, it was seen that the majority of the variance in the mathematical achievements of the sixth graders could be explained by the student-level variables (0.54). For this reason, the desire is to add the demographic variables (gender-GENDER, grade point average-GPA6 and attendance at school in sixth grade-ABSENTEEISM6) of the students, who are thought to influence the mathematical achievement, to the model. However, it had been decided to include only the variables of the students' "grade point average and attendance at school in sixth grade" in the explanatory analysis carried out before the HLM analysis. It is necessary to create learning experiences by using the time carefully and efficiently for the student's mathematical achievement growth in a middle school to occur and continue. A teaching activity in which the student does not exist during the time allocated for learning causes the learning experiences to be incomplete (Altınkurt, 2008; Fidan, 2004; Özbaş, 2010; Sulu Çavumirza, 2012). In this context, the mathematical achievements of the students with high absenteeism are expected to be lower. From this study, it was observed that the variable "absenteeism" influenced the students' mathematical achievements in parallel with this expectation. Similarly, in the study carried out by Yavuz and Atar (2016), it was observed that the attendance of the students in the school affected the students' academic achievement. Another variable handled at the second level is the average grade point in sixth grade. There is no knowledge about mathematics grade point averages provided by the same teacher within a school year. This ambiguity introduced limitations into the study and the discussion was made considering this limitation. There are studies in the literature that examine the effect of the average yearly grade on the future test achievement of students. For example, Cyrenne and Chan (2012) examined students' mathematical achievement using HLM on data obtained from 5136 students from 84 schools. Likewise, Finn, Gerber and Wang (2002) and Kim (2006) examined the effect of the students' grade point average on the examinations taken the following year. As a result of the investigations, the researchers determined that the students' yearly mathematics grade point average predicted the achievement in the next mathematics examination. In other words, it was seen that the yearly mathematics grade point average affected the achievement of the next mathematics exam. Similarly, in this study, it was determined that the average mathematics grade point in the sixth grade affected SBS mathematical achievement of the students.

In the school level (level-3) analysis, it was desirable to add both the school type (TYPE) and the size (SIZE) as the school variables in the model. However, it was decided that only the "TYPE" variable was included in the analysis by the explanatory analysis carried out prior to the HLM analysis. As a result of the school level analysis, it was determined that the school type affected the students' SBS mathematical achievements. Similar to this research finding, Kim (2006), who studied the effects of student and school variables on the eighth, tenth, and twelfth-grade mathematical achievements, observed the impact of school achievement on student achievement in these three grades. It was determined that the mathematical achievements of students attending non-government schools at every grade level were higher than those of other students. It is known that some of the non-government schools in Kim's (2006) study sample selected their students through the exam. Similarly, some private schools in Turkey also determine their students by examinations. The schools in the sample of this study (data) were obtained from the sample of Ankara of the MoNE with the request for random school selection. The names of the schools were given a different code for the researcher, for ethical considerations. For this reason, it is not known whether the private school in the study sample selected its students by a special examination. If the private school in the study sample is selected by the exam, it is expected that the SBS achievements of the students going to this school will be high. Another study examining the effect of school type on student achievement is Lee and Smith's (2001) study. Lee and Smith (2001) examined the mathematical achievement of students with low and high socio-economic status according to school types. At the end of their study, students with high socioeconomic status were found to have a high level of mathematical achievement in all school types while those with low socioeconomic status were

influenced by the types of schools they attended. Finally, the studies carried out by Arslan, Satıcı and Kuru (2006) determined the effectiveness of the public and private primary schools through teachers' opinions. From the studies, they conclude that private schools, to a large extent, are more effective than public schools. Turhan et al., (2017), in their study of effective school studies, found that only four studies from 39 studies made a comparison between the effectiveness of public and private school. They concluded that this number was quite low. Researchers emphasized that there must be further research on this issue and stated that due to the low number of studies, a clear and emphatic judgment on school effectiveness could not be achieved.

Suggestions

In this study, it is concluded that the type of school that students attend and their absenteeism in their schools affect their mathematical achievement. There are many reasons for these variables to be effective on students' achievement. One reason for this may be that private schools provide students with an effective learning environment. To have an effective learning environment, the size of the class should be small, the physical characteristics of the class should be suitable for learning activities, and the teachers should be equipped with the necessary knowledge (Engin, Özen & Bayoğlu, 2009; Özden, 2017). In this context, additional buildings or new schools can be built to reduce class size. The physical properties of the new classes constructed can be designed in a teachable way. Existing classes can be used in the most appropriate way for teaching activities as much as possible. Finally, the qualifications of the teachers can be determined at certain intervals and their problems can be solved with the necessary courses. In an effective and efficient learning environment, students do not get bored and they maintain their interest and desire for learning (Engin et al., 2009; Özden, 2017). Students who do not get bored, and who are persistent about their interest and desire in learning, are expected to spend more time in such environments. Taking this into account, effective learning environments can be prepared at schools to prevent student absenteeism. Another reason for student absenteeism may be that the students may not find the education in their schools sufficient to be able to succeed in SBS (Yılmaz, 2011). For this reason, the students do not attend the schools by taking reports and they might continue their education taking private lessons or in some other ways in this process. It is necessary to change the perception that the education given in schools is not enough. To do so, effective learning environments can be created, the necessary tools for teaching can be provided, or the teachers' lack of pedagogical knowledge can be eliminated if necessary. In addition, student absenteeism can contribute a certain percentage to the placement score in the SBS.

In this study, unlike some studies carried out abroad, no change was observed in the mathematical achievements of the students. This may be due to the fact that there was no real growth in mathematical achievement or that the data were not appropriate for the growth analysis. The number of items and item difficulties in the SBS exams taken in different years can be shown as an indicator of the inappropriateness of the data. In this context, the tests that can provide equivalence should be planned, prepared and implemented. Including vertically scaled data in the HLM analysis may facilitate the observation of growth if the data is appropriate. In addition, within the scope of this study, the achievement scores of the students were obtained as the total number of true and false responses. Therefore, the response patterns of students were not modeled. It is thought that mathematical achievement growth can be observed by modeling the students' response patterns. Finally, there is a need for more studies investigating mathematical achievement growth.

The effects of the variables (attendance of students in school, the grade point average in mathematical achievement and school type) on mathematical achievement were examined. This analysis can be repeated by obtaining different variables (socio-economic status of the student, educational status of the family, frequency and duration of the mathematics course, opportunities of the school, climate, student/mathematics teacher ratio, etc.), and the effects of variables used in the study with new variables on mathematical achievement can be examined. Similarly, the academic achievement of students and their developmental achievements in other domains apart from mathematics can be examined with similar variables using a three-level hierarchical linear growth model. Finally, in causal-comparative design of this study, it was seen that the variables of "students'

attendance to school, average mathematics achievement, and school type" were found to affect students' mathematical achievements. However, the lack of an experimental study in this study constitutes the limitation of this study. For this reason, experimental studies including these variables are needed to determine the real effects discussed in the research.

REFERENCES

Ai, X. (1999). *Gender differences in growth in mathematics achievement: three-level longitudinal and multilevel analyses of individual, home, and school influences* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database (UMI No. 9940503).

Agodini, R. & Harris, B. (2016). How teacher and classroom characterıstıcs moderate the effects of four elementary math currıcula. *The Elementary School Journal, 117*(2), 216-236.

Akyüz, G. (2014). TIMSS 2011'de Öğrenci ve Okul Faktörlerinin Matematik Başarısına Etkisi. *Eğitim ve Bilim, 39*(172), 150-162.

Altınkurt, Y. (2008). Öğrenci devamsızlıklarının nedenleri ve devamsızlığın akademik başarıya olan etkisi. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi, 20*, 129-142.

Altun, A. (2007). *Effects of student and school related factors on the mathematics achievement in Turkey at eight grade level* (Yayınlanmamış yüksek lisans tezi). Orta Doğu Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

Arslan, H., Satıcı, A. ve Kuru, M. (2006). Devlet ve özel ilköğretim okullarının etkililiğinin araştırılması. *Eğitim ve Bilim Dergisi, 31*(42), 15-25.

Atar, B. (2010). Basit doğrusal regresyon analizi ile hiyerarşik doğrusal modeller analizinin karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 1*(2), 78-84.

Atar, H. Y. & Atar, B. (2012). Türk eğitim reformunun öğrencilerin TIMSS 2007 fen başarılarına etkisinin incelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri, 12*(4), 26-36.

Aydın, M. (2015). *Öğrenci ve okul kaynaklı faktörlerin TIMSS matematik başarısına etkisi* (Doktora Tezi). Necmettin Erbakan Üniversitesi, Eğitim Bilimleri Enstitüsü, Konya.

Aztekin S. & Yılmaz, H. B. (2014). The effects of human and material resources on students' math achievement in 45 countries. *Problems of Education in the 21st Century*, 62, 8-20.

Bryk, A., & Raudenbush, S. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education, 97*(1), 65-108.

Büyüköztürk, Ş., Çakmak, E. K., Akgün, Ö. E., Karadeniz, Ş. & Demirel, F. (2008). *Bilimsel araştırma yöntemleri* (14. Baskı). Ankara: Pegem A Yayıncılık.

Cyrenne, P. & Chan, A. (2012). High school grades and university performance: A case study. Economics of *Education Review, 31*, 524– 542.

Çelik, İ. (2016). *Ülke özelliklerinin TIMSS 2011 sekizinci sınıf matematik başarısına çok düzeyli etkileri* (Yüksek lisans tezi). Gazi Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Çokluk, Ö., Şekercioğlu, G. & Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik SPSS ve LISREL Uygulamaları* (3. Baskı). Ankara: Pegem A Yayıncılık.

Demir, İ. & Kılıç, S. (2010). Öğrencilerin matematik başarısına etkileyen faktörlerin PISA 2003 kullanılarak incelenmesi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 38, 44-54.

Ding, C. S., Song, K. & Richardson, L. (2010). Do mathematical gender differences continue? A longitudinal study of gender difference and excellence in mathematics performance in the U.S. Educational Studies. *A Journal of the American Educational Studies Association, 40*(3), 279-295.

Dorans, N.J. (1999). *Correspondence between ACT and SAT I Scores*. College Board Research Report 99-1. NY: The College Board.

Engin, A. O., Özen, Ş. & Bayoğlu, V. (2009). Öğrencilerin okul öğrenme başarılarını etkileyen bazı temel değişkenler. *Sosyal Bilimler Enstitüsü Dergisi,* 3, 125-156.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40*(5), 532–538.

Fidan, F. (2004). Çalışan çocuk olgusuna sosyo-psikolojik bakış. *Danişma Kurulu, 4*(1), 30-49.

Finn, J. D., Gerber, S. B., & Wang, M. C. (2002). Course offerings, course requirements, and course taking in mathematics. *Journal of Curriculum and Supervision, 17*, 336-366.

Green, J. H. (1995). *Is there inequality of educational opportunity? A new look using longitudinal data and a hierarchical model* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database (UMI No. 9621011).

Heck R. H. & Thomas, S. L. (2009). *An introduction to multilevel modeling techniques* (2. Ed.). New York: Routledge.

Hox, J.J. (2010). *Multilevel analysis: techniques and applications* (2. Ed.). Great Britain: Routledge.

Huang, D., Leon, S., La Torre, D. & Mostafavi, S. (2008). *Examining the relationship between LA's best program attendance and academic achievement of LA's best students.* National Center for Research on Evaluation, Standards, and Student Testing (CRESST) Report 749, University of California, Los Angeles.

İş Güzel, Ç. (2006). *Uluslararası Öğrenci Değerlendirme Programında (PISA 2003) insan ve fiziksel kaynakların öğrencilerin matematik okuryazarlığına olan etkisinin kültürler arası karşılaştırılması* (Doktora tezi). Orta Doğu Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

Karabay, E. (2012). *Sosyo-kültürel değişkenlerin PISA fen okuryazarlığını yordama gücünün yıllara göre incelenmesi* (Yüksek lisans tezi). Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Kao, Y., Davenport, J., Matlen, B., Thomas, L., & Schneider, A. (2017). *Bridging research and practice: Efficacy of a research-based redesign of a grade 7 mathematics curriculum.* Paper presented at the Annual Meeting of the American Educational Research Association, San Antonio, TX.

Kolen, M. J. & Brennan, R. L. (2014). *Test equating, scaling, and linking methods and practices*. USA: Springer.

Lee, V. E. & Smith, J. B. (2001). *Restructuring high schools for equity and excellence: What works.* New York: Teachers College.

Lee, V. E. (2000). Using hierarchical linear modeling to study social contexts: The case of school effects. *Educational Psychologist, 35* (2), 125-141.

Milli Eğitim Bakanlığı (2011). *Ortaöğretim Kurumlarına Geçiş Sistemi - Seviye Belirleme Sınavı e-Başvuru Kılavuzu.* Erişim adresi: http://www.meb.gov.tr

Odden, A., Borman, G. & Fermanich, M., (2009). Assessing teacher, classroom, and school effects, including fiscal effects. *Peabody Journal of Education, 79*(4), 4-32.

Ötken, Ş. (2012). *İlköğretim 7. sınıf SBS başarısını yordayan değişkenlerin belirlenmesi* (Yüksek lisans tezi). Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Özbaş, M. (2010). İlköğretim okullarında öğrenci devamsızlığının nedenleri. *Eğitim ve Bilim, 35*(156), 32-44.

Özdemir, F. (2010). *PISA 2003'de genel lise öğrencileri ve Kanuni Lisesi öğrencilerinin matematik başarısını etkileyen faktörlerin incelenmesi* (Yüksek lisans tezi). Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Özden, Y. (2017). Sınıf içinde öğrenme öğretme ortamının düzenlenmesi. Emin Karip (Ed.), *Sınıf yönetimi* içinde (s. 36-69). Ankara: Pegem Akademi Yayıncılık.

Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical linear models* (2. Ed.). Newbury Park, CA: Sage

Raymond, K. J., (2009). *Sensitivity of HLM growth and school effect estimates to the number of waves of data utilized* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database (UMI No. 3410501).

Reçber, Ş. (2011). *An investigation of the relationship among the seventh grade students' mathematics selfefficacy, mathematics anxiety, attitudes towards mathematics and mathematics achievement regarding gender and school type* (Yüksek lisans tezi). Orta Doğu Teknik Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

Savaşçı, H. S. (2011). *Sosyoekonomik değişkenlerin ve okulun eğitim kaynaklarının ilköğretim 7. Sınıf öğrencilerinin akademik başarı düzeyleri ile ilişki durumu* (Yüksek lisans tezi). Mehmet Akif Ersoy Üniversitesi, Sosyal Bilimler Enstitüsü, Burdur.

Schneider, D. & Dorans, N. (1999). *Research notes: Concordance between SAT® I and ACT™ scores for individual students.* Office of Research and Development, College Entrance Examination Board: New York.

Sevgi, S. (2009). *Türkiye'de okul ve öğrenci özelliklerinin matematik başarısı ile ilişkileri* (Yüksek lisans tezi). Orta Doğu Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

Shapley, K., Sheehan, D., Maloney, C. & Caranikas-Walker, F., (2011). Effects of technology immersion on middle school students' learning opportunities and achievement. *The Journal of Educational Research, 104*(5), 299-315.

Shay, S. A. E. (2000). *A longitudinal study of achievement outcomes in a privatized public school: a growth curve analysis* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database (UMI No. 9972554).

Shim, M. K. (1995). *A longitudinal model for the study of equity issues in mathematics education* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database (UMI No. 9543721).

Sulu Çavumirza, E. (2012). *İlköğretim 8. sınıf öğrencilerinin sahip oldukları bazı değişkenler ve algıladıkları okul iklimi bakımından seviye belirleme sınavında aldıkları puanların değerlendirilmesi* (Yüksek lisans tezi). Necmettin Erbakan Üniversitesi, Eğitim Bilimleri Enstitüsü, Konya.

Tan, C. Y. & Hew, K. F. (2018). The impact of digital divides on student mathematics achievement in confucian heritage cultures: A critical examination using PISA 2012 data. *International Journal of Science and Mathematics Education, 16*, 1-20. Available from https://doi.org/10.1007/s10763-018-9917-8

Tavşancıl E. & Yalçın, S. (2015) A Determination of Turkish Student's Achievement Using Hierarchical Linear Models in Trends in International Mathematics-Science Study (TIMSS) 2011. *The Anthropologist, 22*(2), 390-396.

Turhan, M., Şener, G., & Gündüzalp, S. (2017). Türkiye'de Okul Etkililiği Araştırmalarına Genel Bir Bakış. *Turkish Journal of Educational Studies, 4*(2). 103-151.

Wu, C. (2004). *The educational aspirations and high school students' academic growth: A hierarchical linear growth model* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database (UMI No. 3145778).

Yang, J. (2000). *The effects of school community on students' academic learning growth: A multilevel analysis of nels:88 for high schools* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database (UMI No. 9972937).

Yavuz, E. & Atar, H. Y. (2016). Examining the effects of students and school variables on PISA 2012 problemsolving achievement in Turkey. *New Trends and Issues Proceedings on Humanities and Social Sciences,* 05, 24-30.

Yılmaz, Ç. (2011). Öğrenci devamsızlıklarının nedenleri. *Eğitişim Dergisi*, 29. Erişim adresi: http://www.egitisim.gen.tr/tr/index.php/arsiv/21-30/sayi-29-cesitleme-ocak-2011/830-ogrenci-devamsizliginin-nedenleri

Yılmaz, E. T. (2006). *Uluslararası öğrenci başarı değerlendirme programı (PISA)'nın Türkiye'deki öğrencilerin matematik başarılarını etkileyen faktörler* (Yüksek lisans tezi). Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Yi, P. & Shin, I. (2018). Multılevel relatıons between external accountability, internal accountability, and math achievement: A cross-country analysıs. *Problems of Education in The 21st Century, 76*(3), 318-332.

Zhu, R. (1998). *Application of hierarchical linear model (3L) to the study of student and school effects on elementary student' math performance over time* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database (UMI No. 9919001).

Zvoch, K. & Stevens, J. J. (2006). Successive student cohorts and longitudinal growth models: An investigation of elementary school mathematics performance. *Education Policy Analysis Archives, 14*(2), 1068–2341.

Öğrenci ve Okul Değişkenlerinin Matematik Başarısı ve Gelişimine Etkileri

Giriş

Öğrencilerin matematik alanında akademik başarılarının gelişmesinde okulların önemli bir işlevinin olması beklenir. Bu bağlamda küçük yaşlardan itibaren üstelik zorunlu olarak verilen eğitim hizmetinin yürütüldüğü okullar, sıkça araştırma konusu olmuştur. Ayrıca okullar, aile dışındaki sosyal ortamlar olarak öğrenmenin en çok gerçekleştiği yerler olması bakımından dikkate değerdir. Böylece birey için okul hayatına başlamak önemli bir dönüm noktası olarak kabul edilebilmektedir.

Bu çalışmanın amacı, öğrenci (cinsiyet, yılsonu başarı ortalaması, okula devam durumu) ve okul (türü ve büyüklüğü) değişkenlerinin ortaokulda matematik başarısı ve gelişimine etkilerini belirlemektir. Çalışmadaki gelişim, öğrencilerin yıllar içinde aldıkları SBS matematik puanlarındaki değişimi ifade etmektedir. Veri yapısı çok düzeyli olduğundan bu çalışmada tek düzeyli analiz yöntemlerinin yanlılığını önlemek için üç seviyeli hiyerarşik lineer gelişim modeli kullanılmıştır. Modelde düzey-1, öğrencilerin SBS matematik puanlarının tekrarlı ölçümlerinden, düzey-2 öğrencilere ait değişkenlerden ve son olarak düzey-3 okullara ait değişkenlerden oluşmaktadır.

Türkiye'deki alan yazın incelendiğinde matematik başarısını ve matematik başarısındaki gelişimi etkileyen faktörleri lineer gelişim modeliyle inceleyen bir çalışmaya rastlanmamıştır. Bununla

birlikte yurt dışında üç düzeyli hiyerarşik lineer gelişim modellerinin kullanıldığı pek çok çalışma mevcuttur (Ding, Song ve Richardson, 2010; Huang, Leon, La Torre ve Mostafavi, 2009; Raymond, 2009; Shapley, Sheehan, Maloney ve Caranikas-Walker, 2011; Shim, 1995; Wu, 2004; Yang, 2000; Zhu, 1998). Ayrıca ilgili alan yazın incelendiğinde yurt dışında matematik başarısına etki eden öğrenci, okul ve ülke değişkenlerini üç düzeyli hiyerarşik lineer model (HLM) ile inceleyen bir çok çalışma (Agodini ve Harris, 2016; Kao, Davenport, Matlen, Thomas ve Schneider, 2017; Tan ve Hew, 2018; Yi ve Shin, 2018) bulunurken Türkiye'de yalnızca iki çalışmaya (Aztekin ve Yılmaz, 2014; Çelik, 2016) ulaşılmıştır. Çelik (2016) çalışmasında sadece ülke değişkenlerinin etkilerini ele alırken, Aztekin ve Yılmaz'ın (2014) çalışmalarında öğrenci, okul ve ülkelere ait değişken etkilerini incelemişlerdir. Ayrıca bu çalışmalarda PISA ve TIMSS matematik başarısındaki bazı değişkenlerin etkileri incelenmiştir. PISA ve TIMSS gibi sınavlar uluslararası komitelerin amaç ve hedeflerini yansıtırken, SBS uluslararası sınavlardan farklı olarak, Türk eğitim sisteminin amaç ve hedeflerini yansıtmaktadır. Bu çalışmanın, öğrencilerin SBS matematik alt testinde yıllar içerisindeki başarı değişimleri ve bazı öğrenci ve okul değişkenlerinin SBS matematik başarılarına etkilerinin incelemesi açısından önemli olduğu düşünülmektedir. Türkiye'de yapılan çalışmalar incelendiğinde cinsiyet, okula devam durumu, öğretmen nitelikleri gibi değişkenlerin matematik başarısına olan etkilerinin çoğunlukla tek ya da iki düzeyli modellerle incelendiği görülmüştür. Türk alan yazınında çok düzeyli verilerin analizinde üç düzeyli modellerin kullanımına artan bir ilgi gözükse de, geçerli sonuçlar elde etmek için hala çok azdır (Çelik, 2016). Tek düzeyli analiz yöntemlerinin yanlılığından kaçınmak ve alan yazındaki üç düzeyli hiyerarşik lineer gelişim modeli uygulamalarının eksikliğinin giderilmesi için bu çalışmada üç düzeyli hiyerarşik lineer gelişim modeli kullanılmıştır. Bu şekilde öğrenci ve okul değişkenlerinin matematik başarısına etkilerinin incelenmesinin yanı sıra matematik başarısındaki gelişimin de incelenmesi hedeflenmiştir.

Yöntem

Öğrencilerin matematik başarıları ve matematik başarılarındaki gelişimi etkileyen öğrenci ve okul değişkenleri incelendiği için bu çalışma nedensel karşılaştırma modelindedir. Araştırma evrenini, Ankara ilinde 2008 yılında ortaokula başlayıp 2011 yılında ortaokuldan mezun olan ortaokul öğrencileri oluşturmaktadır. Örneklem için seçkisiz olmayan, tipik durum örnekleme yöntemi kullanılmıştır. Araştırma örneklemini, Ankara ilindeki 40 ortaokuldan, 2008 yılında ortaokula başlayıp 2011 yılında mezun olmuş, 3715 öğrenci oluşturmaktadır. Öğrencilerin üç yıllık eğitim-öğretim sürecini aynı okulda geçirmeleri dikkate alınmıştır. Çalışmada kullanılan veriler Milli Eğitim Bakanlığı ile yapılan yazışmalar sonucu elde edilmiştir. Bu nedenle çalışmada incelenen değişkenler, her yıl MEB tarafından Türkiye'deki bütün okullar için sabit olarak toplanıp saklanan bazı verilerden oluşmaktadır. Öğrencilerin ilgili eğitim-öğretim dönemlerinde uygulanan SBS matematik alt testi ham puanları bağımlı değişken olarak analize dahil edilmiştir. "Cinsiyet, altıncı sınıftaki yılsonu matematik not ortalaması ve altıncı sınıftaki okula devam durumu" öğrenci düzeyi değişkenlerini oluştururken "okul türü ve büyüklüğü" okul düzeyi değişkenlerini oluşturmaktadır.

Değişkenlerin etkilerinin incelenebilmesi için verilerin analizinde üç düzeyli hiyerarşik lineer gelişim modeli kullanılmıştır. Öğrencilerin SBS matematik alt testi ham puanları, eşit yüzdelikli eşitleme çalışması yapıldıktan sonra üç düzeyli hiyerarşik lineer gelişim modeline dahil edilmiştir.

Sonuç ve Tartışma

Bu çalışmada, öğrenci (cinsiyet, yılsonu not ortalaması, okula devam durumu) ve okul (türü ve büyüklüğü) değişkenlerinin öğrencilerin SBS matematik başarısı ve gelişimine etkilerini üç düzeyli HLM gelişim modeliyle belirlemek amaçlanmıştır. Çalışmada, SBS sınavı sadece ortaokul öğrencilerine uygulandığı ve o yıllarda ortaokul üç yıl olduğu için bir öğrenci en fazla üç tekrarlı ölçüme sahiptir. Ayrıca çalışmada, elde edilen SBS puanlarının yapısına uygun dikey ölçekleme yöntemi (vertical scaling) bulunamamıştır. Bu nedenle dikey ölçekleme yerine eşit yüzdelikli yöntem uygulanarak puanların eşdeğerleri oluşturulmuştur. Araştırmanın analizinde puanların eşdeğerleri kullanılmıştır. Araştırma bulguları bu sınırlılıklar çerçevesinde tartışılmıştır.

Okul ve öğrenci özelliklerinin matematik başarısına etkilerinin incelendiği çalışmalarda, matematik başarısındaki varyansın büyük bir kısmının öğrenci özellikleri tarafından açıklanması beklenir (Odden, Borman ve Fermanich, 2009; Zvoch ve Stevens, 2006). Bu çalışmada, öğrencilerin altıncı sınıftaki matematik başarılarındaki varyansın büyük çoğunluğunun öğrenci özellikleri, sonra öğrenciler-içi (test) özellikleri en son olarak da okul özellikleri tarafından açıklanabileceği görülmüştür. Bu durum matematik başarılarını etkileyen değişkenlerin büyük çoğunluğunun öğrenci düzeyinde olduğu veya öğrenci düzeyi değişkenlerin öğrencilerin matematik başarılarını yordama da büyük öneme sahip olduğu şeklinde ifade edilebilir. Turhan, Şener ve Gündüzalp (2017), okul etkililiği ile ilgili 39 çalışmayı çeşitli yönleriyle incelemişler ve okulların diğer faktörlere (öğrenci, veli) göre öğrenci başarıları üzerinde daha az etkileri olduğu bulgusuna ulaşmışlardır. Benzer bulgunun Akyüz (2014), Aydın (2015), Sevgi (2009), Tavşancıl ve Yalçın'ın (2015) çalışmalarında da elde edildiği görülmüştür. Bu doğrultuda öğrenci özellikleri ile bu özelliklerin öğrencilerin akademik başarılarına etkisinin daha çok çalışılmaya ihtiyaç duyulduğu söylenebilir. Ayrıca test özellikleri tarafından açıklanabilecek varyans oranının da yüksek olduğu görülmektedir. Fakat bu çalışmada testler ile ilgili değişkenler modele dahil edilmediği için bu düzey varyansı açıklanmadan kalmıştır. Sonraki modellerde öğrenci ve okul düzeyi yordayıcıları modele dahil edilerek ortalama matematik başarısındaki varyans açıklanmaya çalışılmıştır.

Bu çalışmada, üç düzeyli lineer gelişim modeli kullanılan birçok çalışmanın aksine öğrencilerin matematik başarılarındaki gelişim gözlenememiştir. Gelişimin gözlenememesinin bir nedeni testler arasındaki korelasyonun olabileceği düşünülmektedir. Eşitleme çalışmalarında testler arasındaki korelasyon katsayısının 0.87'yi geçmesi istenir (Dorans 1998'den akt. Schneider ve Dorans, 1999). Bu çalışmada öğrencilere ait SBS puanları arasındaki korelasyon katsayıları 0.62 ile 0.72 aralığında değişmektedir. Test puanları arasındaki korelasyonun düşük olması test puanlarının karşılaştırılabilirliğini olumsuz etkilemektedir (Dorans 1998'den akt Schneider ve Dorans, 1999). Bu bağlamda testlerin sınıf düzeylerine göre farklı dizayn edilmeleri (yıllara göre öğrencilere verilen testlerin farklı sayıda maddeye sahip olmaları ve testlerde ortak maddenin bulunmaması gibi) de farklı sınıf düzeylerindeki SBS başarı puanlarının karşılaştırılabilirliklerini olumsuz etkilediği söylenebilir. Gelişimin gözlenememesinin bir diğer nedeni de veri yapısı olabilir. Çalışmada kullanılan veriler soru bazından ziyade kümülatif olarak elde edilmiştir. Bu durum maddelere verilen cevaplar arasındaki örüntünün modellenmesini engellemiştir. Cevap örüntüsünün modellenmesiyle testler arasındaki eşitleme çalışması daha duyarlı olmaktadır (Kolen ve Brennan, 2014). Dikey ölçeklemeyle farklı sınıf düzeylerindeki matematik test puanlarının daha duyarlı bir eşdeğerliğin sağlanmasıyla öğrencilerin matematik başarılarındaki küçük değişimlerin modellenebileceği düşünülmektedir.

Ortaokulda bir öğrencinin matematik başarısı gelişiminin oluşması ve devam etmesi için zamanın dikkatli ve verimli kullanılarak öğrenme yaşantılarının oluşturulması gerekmektedir. Öğrenme için ayrılan süre içerisinde öğrencinin bulunmadığı bir öğretim etkinliği, onun gerçekleştireceği öğrenme yaşantılarının eksik olmasına neden olmaktadır (Altınkurt, 2008; Fidan, 2004; Özbaş, 2010; Sulu Çavumirza, 2012). Bu bağlamda devamsızlığı fazla olan öğrencilerin matematik başarılarının daha düşük olması beklenir. Bu çalışmada, bu beklentiyle paralel olarak "devamsızlık" değişkeninin öğrencilerin matematik başarılarını etkilediği belirlenmiştir. Benzer şekilde Yavuz ve Atar'ın (2016) çalışmasında öğrencilerin okula devam durumlarının, öğrencilerin akademik başarılarını etkiledikleri görülmüştür. Düzey 2'de ele alınan bir diğer değişken altıncı sınıf yılsonu başarı ortalamalarıdır. Alan yazında yılsonu başarı ortalamalarının öğrencilerin gelecekteki sınav başarılarına etkisini inceleyen çalışmalar mevcuttur. Örneğin Cyrenne ve Chan (2012), 84 okuldan 5136 öğrenciden elde ettiği veriler üzerinde HLM kullanarak öğrencilerin matematik başarısını incelemiştir. Benzer şekilde Finn, Gerber ve Wang (2002) ve Kim (2006) de öğrencilerin yılsonu başarı ortalamalarının bir sonraki yıl girdikleri sınavlara etkilerini incelemişlerdir. Araştırmacılar incelemeleri sonucunda, öğrencilerin yılsonu matematik başarı ortalamaları ile bir sonraki matematik sınavı başarılarının yordandığını tespit etmişlerdir. Başka bir ifade ile yılsonu matematik başarı ortalamasının bir sonraki matematik sınavı başarısını etkilediği görülmüştür. Benzer şekilde, bu çalışmada da yılsonu

matematik başarı ortalamasının öğrencilerin altıncı sınıf SBS matematik başarılarını etkilediği belirlenmiştir.

Okul düzeyinin analizi sonucunda, okul türünün öğrencilerin SBS matematik başarılarını etkilediği belirlenmiştir. Bu araştırma bulgusuna benzer şekilde okul ve öğrenci değişkenlerinin sekizinci, onuncu ve 12.sınıf matematik başarıları üzerindeki etkisini araştıran Kim'in (2006) çalışmasında okul türünün her üç sınıfta öğrenci başarısı üzerindeki etkisini gözlemlemiştir. Her sınıf düzeyinde de devlet okulu olmayan okullarda öğrenimine devam eden öğrencilerin matematik başarılarının diğer öğrencilere göre yüksek olduğu belirlenmiştir. Kim'in (2006) çalışma örnekleminde bulunan, devlet okulu olmayan bazı okulların öğrencilerini sınav ile seçtikleri bilinmektedir. Benzer şekilde Türkiye'de de bazı özel okullar kendi öğrencilerini sınav ile belirlemektedirler. Bu çalışmanın örnekleminde bulunan okullara (verilere) MEB'in Ankara örnekleminden randum okul seçme talebi ile ulaşılmıştır. Ulaşılan okulların isimleri etik ilkeleri gözetilerek araştırmacılara farklı bir kodlama ile verilmiştir. Bu nedenle çalışma örneklemindeki özel okulların öğrencilerini özel bir sınavla seçip seçmedikleri bilinmemektedir. Eğer çalışma örnekleminde bulunan özel okul öğrencilerini sınav ile seçti ise, bu okula giden öğrencilerin SBS başarılarının yüksek olması beklenilen bir sonuçtur. Okul türünün öğrenci başarısına etkisini inceleyen bir diğer çalışma Lee ve Smith'e (2001) aittir. Lee ve Smith (2001), düşük ve yüksek sosyo-ekonomik statüye sahip öğrencilerin okul türlerine göre matematik başarılarını incelemişlerdir. Çalışmalarının sonunda yüksek sosyo-ekonomik statüye sahip öğrencilerin her türlü okulda öğrenmelerinin ve bu doğrultuda matematik başarılarının yüksek olduğunu belirlerken, düşük sosyo-ekonomik statüye sahip öğrencilerin devam ettikleri okulların türlerinden etkilendiklerini belirlemişlerdir. Son olarak devlet ve özel ilköğretim okullarının etkililiğini öğretmen görüşleriyle belirlemek isteyen Arslan, Satıcı ve Kuru'nun (2006) çalışmaları sonucunda, belirlenen boyutlarda özel okulların devlet okullarından daha etkili oldukları belirlenmiştir. Etkili okul çalışmalarının incelendiği çalışmada Turhan, Şener ve Gündüzalp (2017), 39 çalışmanın içerisinden sadece dört çalışmanın devlet ve özel okul etkililiği karşılaştırmasında bulunduğunu ve bu sayının oldukça az olduğunu belirtmişlerdir. Araştırmacılar bu konuda daha fazla araştırma yapılmasını vurgularken az çalışmanın yapılmasından dolayı okul etkililiği konusunda net bir yargıya ulaşılamayacağını belirtmişlerdir.

Sonuç olarak, üç düzeyli hiyerarşik lineer gelişim modeli analizi sonucunda öğrencilerin matematik başarılarında bir gelişme olmadığı ancak, "altıncı sınıfta okula devam durumu, yılsonu matematik not ortalamasının ve okul türünün" öğrencilerin altıncı sınıftaki matematik başarılarını istatistiksel olarak etkiledikleri görülmüştür. Araştırmada deneysel bir çalışmanın yapılmaması bu çalışmanın bir sınırlılığını oluşturmaktadır. Bu nedenle etkili bulunan değişkenlerin, etkilerinin tam olarak belirlenebilmesi için bu değişkenleri içeren deneysel çalışmalara ihtiyaç duyulmaktadır.