Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN:1309-6575 *Bahar* 2021 *Spring* 2021 *Cilt: 12- Sayı: 1* *Volume: 12- 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: epodderdergi@gmail.com
Web: https://dergipark.org.tr/tr/pub/epod

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

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

Hakem Kurulu / Referee Board

Ahmet Salih ŞİMŞEK (Kırşehir Ahi Evran Üni.)
Ahmet TURHAN (American Institute Research)
Akif AVCU (Marmara Üni.)
Alperen YANDI (Abant İzzet Baysal Üni.)
Asiye ŞENGÜL AVŞAR (Recep Tayyip Erdoğan Üni.)
Ayfer SAYIN (Gazi Üni.)
Ayşegül ALTUN (Ondokuz Mayıs Üni.)
Arif ÖZER (Hacettepe Üni.)
Arife KART ARSLAN (Başkent Üni.)
Aylin ALBAYRAK SARI (Hacettepe Üni.)
Bahar ŞAHİN SARKIN (İstanbul Okan Üni.)
Belgin DEMİRUS (MEB)
Bengü BÖRKAN (Boğaziçi Üni.)
Betül ALATLI (Gaziosmanpaşa Üni.)
Betül TEKEREK (Kahramanmaraş Sütçü İmam 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.)
Celal Deha DOĞAN (Ankara Üni.)
Cem Oktay GÜZELLER (Akdeniz Üni.)
Cenk AKAY (Mersin Üni.)
Ceylan GÜNDEĞER (Aksaray Üni.)
Çiğdem REYHANLIOĞLU (MEB)
Cindy M. WALKER (Duquesne University)
Çiğdem AKIN ARIKAN (Ordu Üni.)
David KAPLAN (University of Wisconsin)
Deniz GÜLLEROĞLU (Ankara Üni.)
Derya ÇAKICI ESER (Kırıkkale Üni)
Derya ÇOBANOĞLU AKTAN (Hacettepe Üni.)
Devrim ALICI (Mersin Üni.)
Devrim ERDEM (Niğde Ömer Halisdemir Üni.)
Didem KEPIR SAVOLY
Didem ÖZDOĞAN (İstanbul Kültür Üni.)
Dilara BAKAN KALAYCIOĞLU (Gazi Üni.)
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 Can AYBEK (Pamukkale Üni.)
Eren Halil ÖZBERK (Trakya Üni.)
Ergül DEMİR (Ankara Üni.)
Erkan ATALMIS (Kahramanmaras Sütçü İmam Üni.)
Ersoy KARABAY (Kirşehir Ahi Evran Üni.)
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.)
Fuat ELKONCA (Muş Alparslan Üni.)
Fulya BARIŞ PEKMEZCİ (Bozok Üni.)
Funda NALBANTOĞLU YILMAZ (Nevşehir Üni.)
Gizem UYUMAZ (Giresun Üni.)
Gonca USTA (Cumhuriyet Üni.)
Gökhan AKSU (Adnan Menderes Üni.)
Görkem CEYHAN (Muş Alparslan Üni.)
Gözde SIRGANCI (Bozok Üni.)
Gül GÜLER (İstanbul Aydın Üni.)
Gülden KAYA UYANIK (Sakarya Üni.)
Gülşen TAŞDELEN TEKER (Hacettepe Üni.)
Hakan KOĞAR (Akdeniz Üni.)
Hakan SARIÇAM (Dumlupınar Üni.)
Hakan Yavuz ATAR (Gazi Üni.)
Halil İbrahim SARI (Kilis Üni.)
Halil YURDUGÜL (Hacettepe Üni.)
Hatice KUMANDAŞ (Artvin Çoruh Üni.)
Hikmet ŞEVGİN (Van Yüzüncü Yıl Üni.)
Hülya KELECİOĞLU (Hacettepe Üni.)
Hülya YÜREKLI (Yıldız Teknik Üni.)
İbrahim Alper KÖSE (Abant İzzet Baysal Üni.)
Ibrahim YILDIRIM (Gaziantep Üni.)
İlhan KOYUNCU (Adıyaman Üni.)
İlkay AŞKIN TEKKOL (Kastamonu Üni.)
İlker KALENDER (Bilkent Üni.)
İsmail KARAKAYA (Gazi Üni.)
Kübra ATALAY KABASAKAL (Hacettepe Üni.)
Levent YAKAR (Kahramanmaraş Sütçü İmam Üni.)
Mehmet KAPLAN (MEB)
Mehmet ŞATA (Ağrı İbrahim Çeçen Üni.)
Melek Gülşah ŞAHİN (Gazi Üni.)
Meltem ACAR GÜVENDİR (Trakya Üni.)
Meltem YURTÇU (İnönü Üni.)
Metin BULUŞ (Adıyaman Üni)
Murat Doğan ŞAHİN (Anadolu Üni.)
Mustafa ASİL (University of Otago)
Mustafa İLHAN (Dicle Üni.)
Nagihan BOZTUNÇ ÖZTÜRK (Hacettepe Üni.)
Nail YILDIRIM (Kahramanmaraş Sütçü İmam 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)

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (Mart 2021, Sayı: 12-1)

Journal of Measurement and Evaluation in Education and Psychology (March 2021, Issue: 12-1)

İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(1); 1-14

Investigating the Performance of the Exploratory Graph Analysis When the Data Are Unidimensional and Polytomous

Akif AVCU*

Abstract
The question of how observable variables should be associated with latent structures has been at the center of the area of psychometrics. A recently proposed alternative model to the traditional factor retention methods is called Exploratory Graph Analysis (EGA). This method belongs to the broader family of network psychometrics which assumes that the associations between observed variables are caused by a system in which variables have direct and potentially causal interaction. This method approaches the psychological data in an exploratory manner and enables the visualization of the relationships between variables and allocation of variables to the dimensions in a deterministic manner. In this regard, the aim of this study was set as comparing the EGA with traditional factor retention methods when the data is unidimensional and items are constructed with polytomous response format. For this investigation, simulated data sets were used and three different conditions were manipulated: the sample size (250, 500, 1000 and 3000), the number of items (5, 10, 20) and internal consistency of the scale ($\alpha = 0.7$ and $\alpha = 0.9$). The results revealed that EGA is a robust method especially when used with graphical least absolute shrinkage and selection operator (GLASSO) algorithm and provides better performance in the retention of a true number of dimension than Kaiser's rule and yields comparable results with the other traditional factor retention methods (optimal coordinates, acceleration factor and Horn's parallel analysis) under some conditions. These results were discussed based on the existing literature and some suggestions were given for future studies.

Key Words: Exploratory graph analysis, factor analysis, network psychometrics.

INTRODUCTION

The question of how observable variables should be associated with latent structures have been at the center of the area of psychometrics (Borsboom & Molenaar, 2015). So far, various models were developed to specify this association. However, despite the quantitative increase in numbers and great flexibility of mathematical models used in psychometric studies, the models are surprisingly limited in terms of the paradigm that they are based on.

There are two large families of the models in social sciences to describe the relationships between latent variables and observed variables (Edwards & Bagozzi, 2000). In the first category, the latent traits are considered as the common cause of the observed scores. The model based on such kind of conceptualization is called reflective. Reflective models assume that latent traits cause observed variables (also known as indicators, test items, or symptoms. In reflective models, the indicators are modeled as a function of a common latent variable plus some amount of item-specific error variance. Confirmatory factor analysis (CFA) is one of the most commonly used methods representing reflective models.

Formative models are another broad category to define the relationship between latent structures and observed variables. By this conceptualization, it is accepted that observable variables define the latent structures, not caused by them. The classic example of these kinds of models is the socio-economic status defined by a set of observed variables (e.g. education, job, salary and the district of residency). Principal component analysis (PCA) can be given as a classic example of this kind of model. Using

* Research Assistant, Marmara University, Atatürk Faculty of Education, Istanbul-Turkey, avcuakif@gmail.com, ORCID ID: 0000-0003-1977-7592

To cite this article:
Avcu, A. (2021). Investigating the Performance of Exploratory Graph Analysis When the Data Are Unidimensional and Polytomous. *Journal of Measurement and Evaluation in Education and Psychology, 12*(1), 1-14. doi: 10.21031/epod.784128

Received: 22.08.2020
Accepted: 03.01.2021

PCA, data is reduced based on weighted combinations of observed variables to define latent traits (Pearl, 2000).

On the other hand, there is no *"rule of thumb"* when deciding on how many dimensions will retain. In the literature, there are many standard methods for this decision. Kaiser's rule of eigenvalues greater than one rule (KR1: Kaiser, 1960) is the most widely preferred criterion in deciding on how many factors will be retained. This popularity is partly related to its ease of application. However, this method is very sensitive to the number of variables (Gorsuch, 1983) and reliability (Cliff, 1988). Therefore, it may not be effective enough when used in factor retention decisions. An alternative method to KR1 is parallel analysis (PA) developed by Horn (1965). This method is the sample-based adaptation of the KR1 method and has been proposed to alleviate the component indeterminacy problem. Literature shows that this method shows the best performance for component analysis and factor analysis in determining the actual number of factors (Lance, Butts, & Michels, 2006; Velicer, Eaton, & Fava, 2000). Readers are encouraged to look at Kline (2014) for more technical information for KR1, PA and other methods.

More recently, the Acceleration Factor (AF) and Optimal Coordinates (OC) methods were proposed by Raiche, Riopel and Blais (2006) and Raiche (2010). These methods provide non-graphical solutions to Cattell's scree test (1966) to overcome its subjective weakness. AF shows where the elbow of the slope is on the graph and corresponds to the curve's acceleration, i.e. the second derivative. That is, it aims to determine the point where the slope changes abruptly. OC is the other method based on measuring gradients associated with the eigenvalues and preceding eigenvalues to determine the slope's location. It has been stated that AF and OC methods perform better than the KR1 method and approach the performance of PA under certain conditions. (Ruscio and Roche, 2012).

A recently proposed alternative model to traditional reflective and formative approaches is called network modeling. In this approach, there is an assumption that the associations between observed variables are caused by a system in which variables have direct and potentially causal interaction with each other (Eaton, 2015). The usage of network models has provided considerable benefit for understanding complex systems in many different disciplines (Barabási & Pósfai, 2016). In the social sciences, the application of network analysis was adopted firstly to investigate social network structures (eg. Cartwright and Harary, 1956). However, in the following decades, it has been used as an alternative to latent variable modeling in studies to analyze network models of psychological behaviors in an exploratory manner (Borsboom & Cramer, 2013; Schmittmann et al., 2013). After this shift in the application of network modeling, the popularity of the network approach increased and it started to be used intensively in psychology and led to the emergence of a new branch of psychology aimed at predicting network structures in psychological data. This new branch is called network psychometrics (Epskamp, Maris, Waldorp, & Borsboom, 2015).

As with other network models, a psychometric network model consists of a series of nodes (or vertices), a set of connections or links between the nodes (also known as edges) and information regarding the structure of nodes and edges (De Nooy, Mrvar & Batagelj, 2011). In this framework, the nodes represent the psychological indicator variables (e.g. symptoms, behaviors, or faces of latent variables). Traditionally, they are represented by circles in the network structure. On the other hand, the edges represent the node's associations and represented in a network models by lines connecting the nodes.

A more recent paper (Golino & Epskamp, 2017) introduced an innovative way to investigate the dimensionality of psychological constructs by network modeling. This new method is called the EGA. As its name implies, this model is not based on prior assumptions when investigating the dimensionality of a construct. Instead, it approaches the psychological data in an exploratory way. A fascinating feature of EGA is that it enables the visualization of the relationships between variables and allocating variables to the dimensions in a deterministic manner (Golino et al., 2020). For this reason, it is an ideal method to test or reevaluate the theoretical structure of psychological constructs.

In an EGA model, traditionally green (or blue) lines on the network represent positive partial correlations, and red lines correspond to negative partial correlations. In addition, the thickness of the

lines gives information about the amount of the correlation as the thicker lines indicate that the partial correlation values approach 1. If the partial correlation values are exactly 0, no line is drawn between the two nodes which implies that the two variables are independent when other variables in the network are conditionally controlled (Pearl, 2000). In figure 1, an exemplary graph of EGA was presented.

Like other psychometric network models, the EGA is also based on Gaussian Graphical Modeling (GGM), which was proposed by Lauritzen (1996). This model estimates the joint distribution of random variables by modeling the inverse of the variance-covariance matrix (Epskamp, Borsboom & Fried, 2018). In this type of modeling, each edge value represents the relationship between a node pair after conditioned to other variables in the model (Epskamp & Fried, 2016). In more concrete terms, partial correlations are used for the construction of networks in the models. If no edges were drawn between nodes, it implies that zero value for partial correlations is estimated. That is, the nodes are not connected in the model and show conditional independence.

Like other statistical methods that use sample data to estimate parameters, correlation and partial correlation values are also affected by sampling variation. Hence, the exact zero values in matrices are rarely be observed in real data. As a result, the estimated networks based on partial correlations become fully connected. Small weights on many edges could possibly reflect weak and potentially spurious partial correlations in this kind of network. These spurious relationships cause a threat to the clear interpretation of networks and replicability. Frequently, a statistical method is used to remove these spurious connections and control network complexity. For estimations based on partial correlations, a commonly used procedure is to apply the least absolute shrinkage and selection operator (LASSO) proposed by Friedman, Hastie and Tibshirani (2008). Because the LASSO can control spurious connections, this method can provide high precision estimates when combined with the community detection algorithm, such as the walktrap algorithm (Pons & Latapy, 2005).

LASSO uses a tuning parameter to remove spurious connections in the model by filtering the network with penalization approach to the inverse covariance matrix. In this way, partial correlation values smaller than a threshold are estimated as exactly zero. The tuning parameter was selected based on minimizing Extended Bayesian information criterion (EBIC) proposed by Chen and Chen (2008). It enables the researcher to control the sparsity of networks (Foygel & Drton, 2010). LASSO is an important part of network modeling because it determines the eventual network structure. It also enables obtaining parsimonious and more interpretable models. In EGA models, a graphical extension of LASSO is used and referred to as GLASSO. In addition, as an alternative to GLASSO, Triangulated Maximally Filtered Graph (TMFG) was proposed. This approach builds a triangulation that enables a score function to maximize. In this way, the data becomes organized in a meaningful structure and modeling becomes possible. The detailed explanations and formulations could be found in Massara, Di Matteo and Aste (2016).

As cited above, the EGA was firstly proposed by Golino & Epskamp (2017). In this paper, they compared the performance of the EGA with five different traditional factor retention methods. These methods are as follows: (a)very simple structure (VSS; Revelle & Rocklin, 1979); (b) minimum average partial procedure (MAP; Velicer, 1976); (c) fit of a different number of factors, from 1 to 10, via BIC and via EBIC; (d) Horn's Parallel Analysis (PA; Horn, 1965); (e) Kaiser-Guttman eigenvalue greater than one rule (Guttman, 1954); (f) EGA.

In the study, these methods were compared with each other by using simulated data sets across different conditions: the sample size (100, 500, 1000 and 5000), the number of factors (2 and 4), the number of items in each factor (5 and 10) and the correlation between the dimensions (.2, .5 and .7). The datasets were generated in two and four dimension structures and as having dichotomous items. The effectiveness of the methods was tested with their estimation rate of a true number of factors. These methods were compared in terms of their performance to extract the true number of dimensions. According to the findings, it was reported that EGA performed better than the traditional factor retention methods especially when the datasets were simulated as having four dimensions and when the number of items in each dimension was five. It was also stated that EGA was found to be the only method giving satisfactory results in all conditions. All in all, this study confirmed the superiority of

EGA to other traditional methods under some conditions. As this study revealed, EGA is suitable to be used with multidimensional datasets.

On the other hand, the reason why multidimensional datasets were preferred in this recent study is that EGA framework was available to be used only with multidimensional datasets, but a recent revision allowed the examination of unidimensional datasets. In this way, practical limitations to test the effectiveness of EGA with unidimensional datasets were eliminated. There are a number of important reasons to examine unidimensionality in tests. First of all, there is a need to calculate the α coefficient for the overall test (Dunn, Baguley, & Brunsden, 2014). In addition, unidimensionality indicates the presence of a common underlying cause or a coherent set of homogeneous causes (DeVellis, 2017). Based on these facts, Golino & Epskamp (2017) recommended testing the performance of EGA with unidimensional datasets composed of polytomously scored items.

Considering the richness of outputs (such as centrality measures, node strength measures, item stability statistics and entropy fit index) EGA provide to evaluate psychometrical properties of scales (Golino & Christensen, 2020), it is assumed that test developers will use EGA with increasing frequency in the future. In addition, some psychological traits like depression (Beard et al. 2016), anxiety (Fisher et al., 2017) or addiction are measured based on the symptoms they are relied on. DiFranza and his colleagues (2002) suggested considering these symptoms as interconnecting networks rather than indicators caused by latent traits. It is assumed that such kinds of understanding of psychopathological symptoms can contribute more to our understanding of disorders (Beard et al. 2016). For this reason, it is fair to assume that use of EGA will increase in the future.

Purpose of the Study

In this regard, the aim of this study was set as the comparison of the performance of EGA with traditional factor retention methods when the data is unidimensional and items are scored in polytomous response format.

METHOD

Data Simulation Procedure

In the current study, three different conditions were manipulated: the sample size (250, 500, 1000 and 3000), the number of items (5, 10, 20) and the internal consistency level (α = 0.7 and α = 0.9). The conditions of the study were determined by taking into account the features of the scales in the existing psychology literature. Related literature shows that the number of items in unidimensional measurement tools show variance. For example, the Satisfaction with Life Scale (Diener, Emmons, Larsen, & Griffin (1985) consists of five items while the Center for Epidemiologic Studies Depression Scale (Radloff, 1977) consists of twenty items. For this reason, a number of items in simulated data sets were allowed to vary between these observed values (5,10,20). In addition, in order to consider a test to be reliable, the lower threshold value was proposed as .7 (Nunnaly 1978). On the other hand, if the α level is above .90, it is regarded as the test has a good level of α. Accordingly, the data sets were simulated as half of them had α at lower threshold (α = 0.7) while another half of the datasets were simulated as having α level regarded as good (α = 0.9). Finally, the sample size of n=250 is generally regarded as the minimum number when applying factor retention methods (Cattell, 1978). For this reason, the simulated datasets were arranged to had a sample size of at least 250 while n=500, n=1000 and n=3000 conditions were also selected when generating data sets. Based on these facts, 24 different conditions were created with a 4x3x2 design. Finally, in line with the main aim of this study, all of the data sets were simulated as having unidimensional structure and datasets were generated as if the items were scored between 1-5 intervals.

For each condition, data simulation was repeated 100 times to obtain more stable results. This process resulted in generating 2400 datasets. The reported results in this study reflect the arithmetic average

of the iterations. The data simulation was performed with *mirt* package (Chalmers, 2012) in R program (R core team, 2019).

Analysis Procedure

EGA analyses were carried out using the *EGAnet* package available in R statistical environment (Golino & Christensen, 2020). The tuning parameter for GLASSO was determined based on EBIC to obtain a sparser network. In this study, this parameter was set at 0.5, which is a default option in *EGAnet*. On the other hand, the *nFactors* package (Raiche, 2010) was used for applying OC, AF, PA and KR1 factor retention methods.

The assessment of how accurate the correct number of dimensions is extracted was made based on extraction accuracy index and bias indices, as Garrido, Abad & Posada (2016). Factor extraction accuracy index was calculated at two stages: (1) coding correct estimation of the true number of factors as 1 and incorrect estimation of castors as 0, (2) taking the arithmetic mean of coded scores. For instance, when 100 datasets were analyzed, if the true number of factors extracted for 50 datasets, the accuracy index was computed as 0.5. On the other hand, the bias index was calculated as a subtraction of the estimated number of dimensions from a true number of dimensions. For instance, for a unidimensional dataset if the estimated number of the dataset is 1, the bias index is calculated as 0 while if the estimated number is 2, the bias value becomes 1. Therefore, a bias value of 0 indicates the correct number of dimensions are extracted perfectly while a bias values far from 0 indicates the poor performance of the corresponding method. Similar to the accuracy index, the values of bias in the results section represents the arithmetic mean of 100 iterations.

RESULTS

The average accuracy index values and corresponding standard deviations obtained from 100 iterations were given in Table 1. When the sample size was set as 250 and datasets contained five items, all of the methods estimated the correct number of factors perfectly regardless of the α level. As the number of items was increased to ten and α level was 0.7, EGA (LASSO) could extract unidimensional structure for 79% while this rate was 49% for EGA(TMFG). Both algorithms of EGA method outperformed the traditional KR1 method. When the α level has risen to 0.9, EGA (LASSO) method estimated the correct number of dimensions for 99% of datasets, whereas EGA (TMFG) method's percentage drops to 9%. On the other hand, for the other four traditional methods, the average accuracy rates were 100%. In particular, EGA (LASSO) method yielded comparable results with traditional methods when the alpha level was 0.9. Finally, for data sets containing twenty items, the accuracy rate of EGA(LASSO) was 2% and 52% for the conditions where the α was 0.7 and 0.9 respectively, whereas accuracy rates of EGA (TMFG) were 0% for both α levels. The only method EGA(LASSO) outperformed was KR1 while EGA (TMFG) yielded the worst accuracy rates.

For the datasets with n=500 sample size condition, all of the methods examined were perfectly estimated unidimensional structure when the sample size contained five items. This result didn't show a difference across α levels. On the other hand, when the number of items was increased to 10 and α level was 0.7, the average accuracy rate of EGA(LASSO) and EGA(TMFG) was found to be 0.99 and 0.45 respectively. EGA(LASSO) outperformed the traditional KR1 method while EGA(TMFG) method yielded the lowest accuracy levels. As the α level increased to 0.90, EGA(TMFG) was the only method that provided an imperfect accuracy rate (%22). Finally, as the number of items in the datasets was increased to 20, only AF performed a perfectly estimated true number of dimensions when the α was set to be 0.7 while AF and PA performed perfectly when the α level was 0.90. On the other hand, EGA methods yielded the worst accuracy rates.

For the n=1000 sample size condition, when the dataset contained five items, all of the methods extracted the correct number of dimensions perfectly while imperfect rates were obtained for EGA(TMFG) with accuracy rates of 0.59 and 0.26 depending on the α level for the datasets contained ten items. Finally, as the number of items was set to be 20, the EGA(LASSO) method's accuracy rates

were 68% and 99% for the α levels of 0.7 and 0.9, respectively. On the other hand, EGA(TMFG) yielded perfectly inaccurate results.

For datasets where the sample size was 3000, the accuracy rate for EGA (LASSO) was 99% when α was 0.7 and the number of items was 20, while it was 100% in other conditions. For EGA (TMFG) method, the accuracy rates for datasets with 10 and twenty items fell to 77% and 0% when the alpha was α = 0.7, while the accuracy rates for the data sets with ten and twenty items and with α value of .9, accuracy rates decreased to 36% and 0% respectively. For the KR1 method, the accuracy rate was 3% for datasets where α = 0.7 and the number of items was 2. For OC, AF and AP methods, a 100% accuracy rate was achieved under all conditions. Lastly, EGA(LASSO) yielded a 99% accuracy rate when α level was 0.7 and datasets contained twenty items while it perfectly estimated true number of dimensions for the rest of the conditions. On the other hand, EGA(TMFG) yielded the lowest accuracy rates when the number of items was 10 and 20. Especially, OC, AF and AP methods yielded perfect accuracy rates under all conditions examined. As could be inferred, based on the number of items, EGA's relative performance against traditional factor retention methods changed dramatically. In addition, for most of the conditions, GLASSO algorithm was superior to TMFG algorithm.

Table 1. Mean Accuracy of Factor Retention Methods

	EGA(LASSO)		EGA(TMFG)		OC		AF		PA		KR1	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
n=250												
α = 0.70												
5 items	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
10 items	0.79	0.41	0.49	0.50	0.91	0.29	1.00	0.00	0.91	0.29	0.02	0.14
20 items	0.02	0.14	0.00	0.00	0.58	0.50	1.00	0.00	0.58	0.50	0.00	0.00
α = 0.90												
5 items	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
10 items	0.99	0.10	0.09	0.29	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
20 items	0.52	0.50	0.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	0.03	0.17
n=500												
α = 0.70												
5 items	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
10 items	0.99	0.10	0.45	0.50	1.00	0.00	1.00	0.00	1.00	0.00	0.75	0.44
20 items	0.45	0.50	0.00	0.00	0.87	0.34	1.00	0.00	0.86	0.35	0.00	0.00
α = 0.90												
5 items	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
10 items	1.00	0.00	0.22	0.42	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
20 items	0.93	0.26	0.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	0.73	0.45
n=1000												
α = 0.70												
5 items	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
10 items	1.00	0.00	0.59	0.49	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
20 items	0.68	0.47	0.00	0.00	0.99	0.10	1.00	0.00	0.99	0.10	0.00	0.00
α = 0.90												
5 items	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
10 items	1.00	0.00	0.26	0.44	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
20 items	0.99	0.10	0.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
n=3000												
α = 0.70												
5 items	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
10 items	1.00	0.00	0.77	0.42	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
20 items	0.99	0.10	0.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	0.03	0.17
α = 0.90												
5 items	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
10 items	1.00	0.00	0.36	0.48	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
20 items	1.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00

The calculated bias values for the factor retention methods across conditions were given in Table 2. If the datasets contained five items, EGA(LASSO) provided unbiased estimates of the correct number of dimensions. As the number of items in the datasets was increased to 10 and the sample size of n=250, the bias value was estimated to be 0.33 0.01 for α levels of 0.7 and 0.9, respectively. As the sample size of datasets was increased to 500, EGA(LASSO) yielded 0.01 and 0 bias for α levels of 0.7 and 0.9. When the sample size was n=1000 and n=3000, EGA(LASSO) yielded no bias when the item number was 10. For the datasets containing twenty items, if the sample size was n=250, the bias value was 2.41 for α level of 0.7 and 1.39 for α level of 0.90. On the other than, the bias value of 1.39 has very large standard deviation value which indicated that, there was a variation across the datasets in terms of the bias value calculated. As the sample size was increased to 500, 1000 and 3000, the bias values calculated showed a decrease compared to n=250 condition. Similar changes were also observed for EGA(TMFG) across the conditions while EGA(TMFG) performed worse than EGA(LASSO) in general. On the other hand, other traditional estimation methods provided almost perfect results especially when the sample size was n=1000 and n=3000.

Table 2. Mean Bias Error of Factor Retention Methods

	EGA(LASSO)		EGA(TMFG)		OC		AF		PA		KR1	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
n=250												
α = 0.70												
5 items	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10 items	0.33	1.02	0.57	0.61	0.10	0.33	0.00	0.00	0.10	0.33	1.42	0.54
20 items	2.41	1.16	2.27	0.75	0.56	0.74	0.00	0.00	0.60	0.84	6.09	0.71
α = 0.90												
5 items	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10 items	0.01	0.10	1.00	0.45	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20 items	1.39	3.71	2.20	0.75	0.00	0.00	0.00	0.00	0.00	0.00	1.73	0.72
n=500												
α = 0.70												
5 items	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10 items	0.01	0.10	0.57	0.54	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.44
20 items	1.37	2.79	2.35	0.69	0.15	0.44	0.00	0.00	0.17	0.47	5.29	0.67
α = 0.90												
5 items	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10 items	0.00	0.00	0.85	0.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20 items	0.07	0.26	2.25	0.74	0.00	0.00	0.00	0.00	0.00	0.00	0.29	0.50
n=1000												
α = 0.70												
5 items	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10 items	0.00	0.00	0.42	0.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20 items	0.85	2.74	2.28	0.74	0.01	0.10	0.00	0.00	0.01	0.10	4.50	0.64
α = 0.90												
5 items	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10 items	0.00	0.00	0.85	0.59	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20 items	0.02	0.20	2.23	0.85	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
n=3000												
α = 0.70												
5 items	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10 items	0.00	0.00	0.25	0.48	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20 items	0.01	0.10	2.05	0.73	0.00	0.00	0.00	0.00	0.00	0.00	1.63	0.68
α = 0.90												
5 items	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10 items	0.00	0.00	0.70	0.58	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20 items	0.00	0.00	2.10	0.85	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

After calculating the accuracy rates and the bias values, a series of factorial ANOVA was performed to examine the effects of conditions altered for each factor retention method. For this analysis, the raw estimated dimension number value was used as the dependent variable. Only eta square (η2) effect size values and the significance levels of ANOVA analysis were reported. The significance levels, ** sign denotes significance at $p<0.01$ level and * implies significance at $p<0.05$. The η2 values show the magnitudes of the differences between the conditions for each method under investigation. According to Cohen (1988), η2 values of 0.14 and above can be regarded as a "*large*" effect size. On the other hand, the effect size for AF method cannot be compared because this method perfectly estimated the true number of dimensions for all 2400 datasets.

For the rest of the methods, it was found that the unique effects of the conditions examined for EGA (GLASSO) method or their two-way and three-way interactions did not have a large effect size. Similar results were observed for OC and PA methods. On the other hand, the item number condition had a large effect size for EGA(TMFG) method. Finally, for the KR1 method, large amounts of the effect size values were observed for each of the conditions examined and their two-way and three-way interactions were found as significant.

Table 3. Effect Sizes of Factorial ANOVA

	EGA(GLASSO)	EGA(TMFG)	OC	AF	PA	KR1
Sample Size (SS)	0.05**	0.01**	0.05**	-	0.05**	0.62**
Number of Items (NI)	0.09**	0.75**	0.04**	-	0.04**	0.91**
Reliability (r)	0.02**	0.01**	0.03**	-	0.03**	0.80**
SS X NI	0.07**	0.01**	0.06**	-	0.06**	0.65**
SS X r	0.01**	0.01	0.05**	-	0.05**	0.36**
NI X r	0.02**	0.03**	0.04**	-	0.04**	0.86**
SS X NI X r	0.01**	0.01	0.06**	-	0.06*	0.45**

**$p<0.01$

DISCUSSION and CONCLUSION

The current study aimed to compare the effectiveness of EGA in extracting the true number of dimensions with traditional methods when the data was unidimensional and composed of polytomous items. This aim was determined based on Golino and Epskamp's (2017) recommendations and literature review showed that no study was conducted so far considering this recommendation. Unlike this study, in the current study, OC and AF methods were included for comparison because these methods are also relatively new compared to more traditional methods like PA and KR1 and their inclusion on relatively new methods is believed to increase existing knowledge on the effectiveness of EGA.

As a result of this study, it has been observed that EGA (LASSO) successfully extracted unidimensional structure perfectly like other methods for datasets where the number of items was five. This success of EGA was valid even for data sets with a sample size as small as 250. A similar finding was obtained for EGA (TMFG). On the other hand, as the number of items increases, the performance of both EGA (LASSO) and EGA (TMFG) decreased. Even when the sample size was 3000 and the reliability level was 0.9, EGA (TMFG) could not extract the correct number of dimension with high accuracy if there were ten or more items in the data set. On the other hand, for n = 500 and n = 1000 sample size conditions, EGA (LASSO) yielded comparable accuracy rates only if the reliability level was 0.9 while it's performance decreased when the reliability dropped to 0.7 and when data sets contained twenty items.

If the methods are compared in general, AF had perfectly extracted the actual dimensional structure regardless of the conditions altered and use of it by the researchers is strictly recommended in their future studies. Overall, EGA (LASSO) algorithm outperformed EGA (TMFG) algorithm. For this

reason, it is recommended that GLASSO algorithm should be preferred over TMFG algorithm for unidimensional and polytomous data sets. The same superior performance of EGA (GLASSO) was also observed when compared with the traditional KR1 method.

Therefore, it can be said that EGA (LASSO) is an important effective alternative for researchers who prefer the traditional KR1 method, which has been used extensively because of availability on most of the commercial software programs. Considering the richness of output EGA provides (see Golino & Christensen, 2020), EGA can be a better alternative to KR1. In addition, if the sample size was increased to 1000 or 3000, EGA (LASSO) method gives results comparable to the OC and PA methods. On the other hand, EGA should be considered as a serious alternative only when the scale contains fewer items with high internal consistency for smaller sample size conditions (250 or 500). Otherwise, OC and PA provide better results.

According to factorial ANOVA results, it was found that there were no unique or interaction effects observed for EGA (LASSO) method. Similar findings were also observed for OC and PA methods. It can be said that these three methods were the most robust ones across the conditions tested. Although these statistics can not be calculated for AF, it provides perfect results under all conditions. it is also definitely correct to consider this method as robust. On the other hand, "large" effect size was observed for the EGA (TMFG) method for the sample size condition. That is, the sample size affects the performance of EGA (TMFG) method negatively regardless of other conditions. The poor performance of TMFG algorithm is understandable because it performs better when booting algorithms are used simultaneously.

Finally, for the KR1 method. "large" effect sizes were observed for all conditions and their two-way and three-way interactions. Accordingly, it can be said that the KR1 method was the least robust method within the context of the conditions examined in this study. This finding is in line with past literature (Velicer, Eaton & Fava. 2000; Ruscio & Roche, 2012).

This study is one of the few studies comparing EGA's factor retention effectiveness with other traditional methods. Contrary to the findings obtained by Golino and Epskamp (2017), EGA(LASSO) was not to be detected as clearly superior to other traditional methods. This result implies that EGA (LASSO) may not be a suitable alternative when the data is unidimensional and potential researchers should use EGA (LASSO) for scales with fewer items, higher internal consistency and a large sample size for unidimensional tests. On the other hand, EGA (TMFG) should not be an option for researchers in a wide of conditions considered in the current study.

All in all, more research is needed to examine the effectiveness of EGA in different conditions. For example, EGA's effectiveness in datasets with different ability distributions will contribute to the richness of the existing literature. In addition, in this study the effectiveness of the methods was only evaluated in terms of the number of factors. In future studies, it is suggested to evaluate the performance of EGA in terms of estimating real factor loadings.

REFERENCES

Barabási, A.-L., & Pósfai, M. (2016). *Network science*. Cambridge: Cambridge University.

Beard, C., Millner, A. J., Forgeard, M. J., Fried, E. I., Hsu, K. J., Treadway, M. T., Leonard, C. V., Kertz, S. J., & Björgvinsson, T. (2016). Network analysis of depression and anxiety symptom relationships in a psychiatric sample. *Psychological medicine*, *46*(16), 3359–3369. doi: 10.1017/S0033291716002300

Borsboom, D., & Cramer, A. O. J. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, *10*(9), 91-121. doi: 10.1146/annurev-clinpsy-050212-185608

Borsboom, D., & Molenaar, D. (2015). Psychometrics. In J. Wright (Ed.). *International encyclopedia of the social & behavioral sciences* (Second ed.. Vol. 19. pp. 418-422). Amsterdam: Elsevier.

Cartwright, D., & Harary, F. (1956). Structural balance: A generalization of Heider's theory. *Psychological Review*, *63*(5), 277–293. doi: 10.1037/h0046049

Cattell R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, *1*(2), 245–276. doi: 10.1207/s15327906mbr0102_10

Cattell, R. B. (1978). *The scientific use of factor analysis*. New York: Plenum.

Chalmers, R.P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6), 1-29. doi: 10.18637/jss.v048.i06

Chen, J., & Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, *95*(3), 759-771. doi: 10.1093/biomet/asn034

Cliff, N. (1988). The eigenvalue-greater-than-one rule and the reliability of components. *Psychological Bulletin*, *103*(2), 276-279.

Cohen J, (1988). *Statistical power analysis for the behavior science*. Lawrance Eribaum Association.

de Nooy, W., Mrvar, A., & Batagelj, V. (2011). *Exploratory social network analysis with Pajek*. Cambridge: Cambridge University.

DeVellis, R. F. (2017). *Scale development: Theory and applications*. Thousand Oaks. CA: SAGE Publications.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction with Life Scale. *Journal of Personality Assessment*, *49*(1), 71-75. doi: 10.1207/s15327752jpa4901_13

DiFranza, J. R., Savageau, J. A., Rigotti, N. A., Fletcher, K., Ockene, J. K., McNeill, A. D., Coleman, M., & Wood, C. (2002). Development of symptoms of tobacco dependence in youths: 30 month follow up data from the DANDY study. *Tobacco control*, *11*(3), 228–235. doi: 10.1136/tc.11.3.228

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of α estimation. *British Journal of Psychology*, *105*(3), 399-412. doi: 10.1111/bjop.12046

Eaton, N. R. (2015). Latent variable and network models of comorbidity: toward an empirically derived nosology. *Social Psychiatry and Psychiatric Epidemiology*, *50*(6), 845-849. doi: 10.1007/s00127-015-1012-7

Edwards, J.R., & Bagozzi, R.P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155-174. doi: 10.1037/1082-989X.5.2.155.

Epskamp, S., & Fried, E. I. (2016). *A primer on estimating regularized psychological networks* arXiv preprint Stat-Ap/1607.01367. Available at: http://arxiv.org/abs/1607.013677

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, *50*(1), 195-212. doi: 10.3758/s13428-017-0862-1

Epskamp, S., Maris, G., Waldorp, L.J., & Borsboom, D. (2015). Network Psychometrics. In Irwing. P., Hughes, D. and Booth, T. (Eds.). *Handbook of Psychometrics*. New York: Wiley.

Fisher, A. J., Reeves, J. W., Lawyer, G., Medaglia, J. D., & Rubel, J. A. (2017). Exploring the idiographic dynamics of mood and anxiety via network analysis. *Journal of abnormal psychology*, *126*(8), 1044–1056. doi: 10.1037/abn0000311

Foygel, R. and Drton, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. *Advances in Neural Information Processing Systems*, 23, 2020-2028.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics (Oxford, England)*, *9*(3), 432-441. doi: 10.1093/biostatistics/kxm045

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2016). Are fit indices really fit to estimate the number of factors with categorical variables? Some cautionary findings via Monte Carlo simulation. *Psychological Methods*, *21*(1). 93–111. doi: 10.1037/met0000064

Golino, H. F., & Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PloS One*, *12*(6), e0174035. doi: 10.1371/journal.pone.0174035

Golino, H., Shi. D., Christensen, A. P., Garrido, L. E.. Nieto, M. D., Sadana, R., Thiyagarajan. J. A., & Martinez-Molina, A. (2020). Investigating the performance of exploratory graph analysis and traditional techniques to identify the number of latent factors: A simulation and tutorial. *Psychological Methods*. Advance online publication. doi: 10.1037/met0000255

Golino, H.. & Christensen, A. P. (2020). *EGAnet: Exploratory Graph Analysis -- A framework for estimating the number of dimensions in multivariate data using network psychometrics*. R package version 0.9.4.

Gorsuch R.L. (1988) Exploratory Factor Analysis. In: Nesselroade J.R., Cattell R.B. (eds) *Handbook of Multivariate Experimental Psychology. Perspectives on Individual Differences*. Springer, Boston, MA. https://doi.org/10.1007/978-1-4613-0893-5_6

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika*, *19*(2), 149-161. doi: 10.1007/BF02289162

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179-185. doi: 10.1007/BF02289447

Kline, P. (2014). *An easy guide to factor analysis*. Routledge.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, *9*(2), 202–220. doi: 10.1177/1094428105284919

Lauritzen, S. L. (1996). *Graphical Models. Oxford Statistical Science Series*. volume 17. New York. NY: Oxford University Press.

Massara, G. P., Di Matteo, T., & Aste, T. (2016). Network filtering for big data: Triangulated Maximally Filtered Graph. *Journal of Complex Networks*, 5, 161–178. doi: 10.1093/comnet/cnw015

Nunnally. J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. New York: Cambridge University.

Pons, P.. & Latapy, M. (2006). Computing communities in large networks using random walks. *Journal of Graph Algorithms Applications, 10*(2), 191-218. doi: 10.1007/11569596_31

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna. Austria. URL https://www.R-project.org/.

Radloff, L.S. (1977). The CES–D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*(3), 385-401. doi: 10.1177/014662167700100306

Raiche, G. (2010). *nFactors: An R package for parallel analysis and non graphical solutions to the Cattell's scree test*. R package version 2.3.3.

Raiche, G., Riopel. M. and Blais, J.-G. (2006). *Non graphical solutions for the Cattell's scree test*. Paper presented at the International Annual Meeting of the Psychometric Society, Montreal.

Revelle, W., & Rocklin, T. (1979). Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research, 14*(4), 403-414. doi: 10.1207/s15327906mbr1404_2

Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological Assessment, 24*(2), 282-292. doi: 10.1037/a0025697

Schmittmann, V. D., Cramer, A. O. J., Waldorp, L. J., Epskamp, S., Kievit, R. A., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. New Ideas in Psychology, *31*(1), 43-53. doi: 10.1016/j.newideapsych.2011.02.007.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*(3), 321-327. doi: doi.org/10.1007/BF02293557

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin & E. Helmes (Eds.). *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (p. 41-71). Kluwer Academic/Plenum Publishers. doi: 10.1007/978-1-4615-4397-8_3

Tek Boyutlu ve Çok Yanıt Kategorisine Sahip Veriler İçin Açıklayıcı Grafik Analizinin Performansının İncelenmesi

Giriş

Gözlenen değişkenlerin örtük yapılarla nasıl ilişkilendirilmesi gerektiği sorusu psikometrinin merkezinde yer almaktadır (Borsboom ve Molenaar, 2015). Şimdiye kadar, bu ilişkiyi belirtmek için çeşitli modeller geliştirilmiştir. Bununla birlikte, psikometrik çalışmalarda kullanılan matematiksel modellerin niceliksel artışına ve büyük esnekliğine rağmen, örtük özellikler ve davranışlar arasındaki ilişkileri tanımlamak için sunulan modeller, dayandıkları paradigma açısından şaşırtıcı bir şekilde sınırlıdır.

Bu geleneksel yaklaşımlara ise yakın zamanda ağ modellemesi olarak adlandırılan alternatif model önerilmiştir. Bu yaklaşımda, gözlenen değişkenler arasındaki ilişkilerin, değişkenlerin birbirleriyle doğrudan ve potansiyel olarak nedensel etkileşime sebep olan bir sistem aracılığıyla kaynaklandığı varsayılmaktadır (Eaton, 2015). Ağ modellerinin kullanımı, birçok farklı disiplindeki karmaşık sistemlerin anlaşılması için büyük ölçüde fayda sağlamıştır (Barabási ve Pósfai, 2016). Sosyal bilimlerde, ağ analizi uygulaması öncelikle sosyal ağ yapılarını araştırmak için benimsenmiştir (örn. Cartwright ve Harary, 1956). Bununla birlikte, sonraki yıllarda, psikolojik davranışların ağ modellerini keşifsel bir şekilde analiz edilmesi geleneksel gizli değişken modellemelerine alternatif olarak kullanılmaya başlamıştır (Borsboom ve Cramer, 2013; Schmittmann vd., 2013). Ağ modelleme uygulamasındaki bu değişimden sonra, ağ yaklaşımının popülaritesi artmış ve psikoloji alanında

yoğun bir şekilde kullanılmaya başlanmış ve psikolojik verilerde ağ yapılarını tahmin etmeyi amaçlayan yeni bir psikoloji alanının ortaya çıkmasına neden olmuştur. Bu yeni alan, ağ psikometrisi olarak adlandırılır (Epskamp, Maris, Waldorp ve Borsboom, 2015).

Diğer ağ modellerinde olduğu gibi, psikometrik ağ modeli de bir dizi düğümden (veya köşelerden), düğümler arasında bir dizi bağlantı veya ağdan (kenarlar olarak da bilinir) ve düğümlerin ve kenarların yapısıyla ilgili bilgilerden oluşur (De Nooy, Mrvar ve Batagelj, 2011). Düğümler, psikolojik gösterge değişkenlerini (örn. gizil değişkenlerin semptomları, davranışları veya yüzleri) temsil eder. Geleneksel olarak, düğümler ağ yapısında dairelerle temsil edilirler. Öte yandan, kenarlar, düğümler arasındaki ilişkileri temsil eder ve bir ağ modelinde daireleri birbirine bağlayan çizgilerle temsil edilir.

Yakın geçmişte yayımlanan bir çalışma (Golino ve Epskamp, 2017), ağ modelleme yoluyla psikolojik yapıların boyutluluğunu araştırmanın yenilikçi bir yolunu sunmuştur. Bu yeni tekniğe Açıklayıcı Grafik Analizi (AGA) adı verilir. Adından da anlaşılacağı gibi, bu model bir yapıyı incelerken önsel varsayımlara dayanmamaktadır. Bunun yerine, psikolojik verileri keşifsel bir anlayışla ele alır. AGA'nın dikkate değer bir özelliği, değişkenler arasındaki ilişkilerin görselleştirilmesi ve değişkenlerin boyutlara atanmasını belirleyici bir şekilde sağlamasıdır. Bu nedenle psikolojik özelliklerin kuramsal yapısını test etmek veya yeniden değerlendirmek için ideal bir yöntemdir.

Kısmi korelasyonlara dayalı tahminlerle gerçekleştirilen bu yöntemde, yaygın olarak en az mutlak daralma ve seçim operatörünün (least absolute shrinkage and selection operator-LASSO) işlemi yaygın olarak uygulanmaktadır (Friedman, Hastie ve Tibshirani, 2008). LASSO, sahte (spurious) bağlantıları kontrol etmek için kullanılmaktadır. LASSO, walktrap gibi topluluk algılama algoritmalarıyla birleştirildiğinde yüksek hassasiyetli tahminler sağlayabilir (Pons and Latapy, 2005).

Optimum bir model elde etmek için Chen ve Chen (2008) tarafından önerilen genişletilmiş Bayesian bilgi kriteri (extended Bayesian information criterion-EBIC) dikkate alınarak belirlenen ayarlama parametresi kullanılır. Bu parametre, araştırmacının ağların seyrekliğini kontrol etmesini sağlar (Foygel ve Drton, 2010). LASSO, nihai ağ yapısını belirlediği için ağ modellemenin önemli bir parçasıdır. Aynı zamanda daha tutucu ve yorumlanabilir modellerin elde edilmesini sağlar. AGA modellerinde LASSO'nun grafik bir uzantısı kullanılmış ve GLASSO olarak adlandırılmıştır. Ek olarak, Üçgenleştirilmiş Maksimum Filtrelenmiş Grafik (TMFG: Triangulated Maximally Filtered Graph), GLASSO'ya alternatif olarak önerilen bir diğer tekniktir. Bu yaklaşım, bir puan fonksiyonunun maksimize etmesini sağlayan bir üçgenleme oluşturur. Bu şekilde veriler anlamlı bir yapı içerisinde organize olur ve modelleme mümkün olur. Ayrıntılı açıklamalar ve formülasyonlar için Massara, Di Matteo ve Aste (2016) 'ye bakılması önerilmektedir.

Bir AGA modelinde, geleneksel olarak, ağ üzerindeki yeşil (ya da mavi) çizgiler pozitif kısmi korelasyonları temsil ederken kırmızı çizgiler, negatif kısmi korelasyonlara karşılık gelir. Ek olarak, çizgilerin kalınlığı korelasyon miktarı hakkında bilgi verir: daha kalın çizgiler kısmi korelasyon değerlerinin 1'e yaklaştığını gösterir, Kısmi korelasyon değerleri tam olarak 0 ise, iki düğüm arasında hiçbir çizgi çizilmez. Yani, ağdaki diğer değişkenlerin etkisi kontrol edildiğinde iki değişken koşullu olarak bağımsızdır (Pearl, 2000).

AGA'nın önerildiği makalede Golino ve Epskamp (2017), AGA'nın performansını beş farklı geleneksel faktör çıkarma tekniğiyle karşılaştırmıştır. Bu çalışmada iki yanıt kategorili maddelerden oluşan iki ve dört boyutlu türetilmiş veri setleri kullanılmıştır. Çalışmanın bulguları, kontrol edilen koşullar ne olursa olsun, özellikle veri kümeleri dört boyutlu yapı olarak simüle edildiğinde AGA'nın en iyi performans gösteren yöntem olduğunu göstermiştir. Özellikle boyut sayısı dört olduğunda, AGA'nın diğer geleneksel yöntemlere üstünlüğünü doğrulanmıştır. Özellikle AGA'nın her boyuttaki madde sayısı beş olduğunda tatmin edici sonuçlar veren tek yöntem olduğu belirtilmiştir. Bu çalışmada çok boyutlu veriler kullanılmış olmasına ragmen daha sonrasında AGA algoritması tek boyutlu veri setlerinin incelenmesine izin verecek şekilde revize edilmiştir. Nitekim, aynı çalışmada AGA'nın tek boyutlu veri setleri için faktör sayısına karar vermedeki performansının incelenmesi önerilmiştir. Bu öneri dikkate alınarak gerçekleştirilen bu çalışmanın amacı veri seti tek boyutlu olduğunda ve maddeler çok yanıt kategorisine sahip olduğunda AGA'nın faktör sayısına karar vermedeki performansının geleneksel faktör çıkarma yöntemleriyle karşılaştırılması olarak belirlenmiştir.

Yöntem

Bu çalışmada üç farklı koşul kontrol edilmiştir: örneklem büyüklüğü (250, 500, 1000 ve 3000), madde sayısı (5, 10, 20) ve iç tutarlılık seviyesi ($\alpha = 0.7$ ve $\alpha = 0.9$). Buna koşullara bağlı olarak 4x3x2 tasarımı ile 24 farklı koşul oluşturulmuştur. Ayrıca, bu çalışmanın temel amacı doğrultusunda, tüm veri setleri tek boyutlu yapıya sahip olacak şekilde türetilmiştir ve maddeler 1-5 aralığında puanlanmış şekilde veri setleri oluşturulmuştur. Daha kararlı sonuçlar elde etmek için her koşul için veri üretme işlemi 100 kez tekrarlanmıştır. Bu sayede, 2400 veri kümesi türetilmiştir. Bu çalışmada bulgular kısmında sunulan sonuçlar, tekrarlar sonucunda elde edilen değerlerin aritmetik ortalamasını yansıtmaktadır. Veri üretme işlemi, R ortamında (R çekirdek ekibi, 2019) "mirt" paketi (Chalmers, 2012) ile gerçekleştirilmiştir.

AGA yönteminin performansını karşılaştırmak amacıyla beş farklı faktör sayısına karar verme yöntemi kullanılmıştır: Hızlanma Faktörü (AF: Acceleration Factor), Optimal Koordinatlar (OC: Optimal Coordinates), Paralel Analiz (PA) ve Kaiser'in özdeğer 1'den büyük kuralı (KR1). Bu dört yönteme ilişkin ayrıntılı teknik bilgiler Raiche, Riopel & Blais (2006) ve Raiche (2010) 'de yer almaktadır. AGA analizleri, R istatistik programında bulunan "EGAnet" paketi (Golino & Christensen, 2020) kullanılarak gerçekleştirilirken, OC, AF, PA ve KR1 faktör çıkarma yöntemleri için "nFactors" paketi (Raiche, 2010) kullanılmıştır. AGA tekniği GLASSO ve TMFG algoritmaları için ayrı ayrı gerçekleştirilmiştir ve çalışmanın geri kalanında sırasıyla AGA(GLASSO) ve AGA(TMFG) anılmıştır. GLASSO algoritması kullanılırken, ayarlama parametresi 0.5 olarak belirlenmiştir. Koşullara göre genel betimleyici istatistiklerin yanısıra faktöriyel varyans analizi (ANOVA) gerçekleştirilerek etkisi incelenen koşulların faktör sayısına karar verme yöntemleri üzerindeki tekil etkileri ile etkileşimlerinden kaynaklı etkilerin incelenmesi amaçlanmıştır.

Garrido, Abad ve Posada (2016) tarafından önerildiği gibi, doğru boyut sayısının ne kadar kesinlikte çıkarıldığına dair değerlendirme, çıkarma doğruluk indeksi ve yanlılık indekslerine dayanılarak yapılmıştır. Faktör çıkarma doğruluk indeksi, doğru sayıda faktörün 1, hatalı sayıda faktörün 0 olarak çıkarıldığı analiz sonuçlarının kodlanmasıyla elde edilmiştir. Örneğin 100 veri seti incelendiğinde, 50 veri seti için gerçek faktör sayısı çıkarılmışsa bu veri setlerinin her biri 1, geri kalanı ise 0 olarak kodlanmıştır. Sonuç olarak 100 veri seti için yöntemin nihai kesinlik 0.5 olarak hesaplanmıştır. Öte yandan, yanlılık indeksi, kestirilen boyut sayısının gerçek boyut sayısından çıkarılmasıyla hesaplanır. Örneğin, tek boyutlu bir veri seti için, kestirilen boyut sayısı 1 ise, sapma endeksi 0 olarak hesaplanırken, kestirilen boyut sayısı 2 ise yanlılık değeri 1 olur. Başka bir anlatımla, sıfır yanlılık değeri, boyut sayısının doğru kestirildiğini gösterirken 0'dan uzak yanlılık değerleri, ilgili yöntemin zayıf performansını göstermektedir.

Sonuç ve Tartışma

Elde edilen bulgulara göre AGA (LASSO) madde sayısının 5 olduğu veri setleri için diğer yöntemler gibi tek boyutlu yapıyı mükemmel bir şekilde kestirdiği görülmüştür. AGA'nın bu başarısı, örneklem büyüklüğü 250 olan veri setleri için bile geçerlidir. Benzer bulgular AGA (TMFG) için de elde edilmiştir. Diğer taraftan, madde sayısı arttıkça hem AGA (LASSO) hem de AGA (TMFG)'nın performansının düştüğü görülmüştür. Örneklem büyüklüğü 3000 ve güvenilirlik düzeyi 0.90 olsa bile AGA (TMFG) veri setinde 10 veya daha fazla madde olduğunda doğru boyut sayısını yüksek doğrulukla çıkartamadığı belirlenmiştir. Ayrıca, $n = 500$ ve $n = 1000$ örneklem büyüklüğü koşulları için AGA (LASSO) yalnızca güvenilirlik seviyesi 0.9 olduğunda diğer yöntemlerle karşılaştırılabilir kesinlik oranları sağlamıştır. Ancak, güvenilirlik 0.7'ye düştüğünde ve veri kümeleri 20 madde içerdiğinde performansı düşmüştür.

Yöntemler genel olarak karşılaştırıldığında ise. AF'nin control edilen koşullardan bağımsız olarak gerçek boyutsal yapıyı mükemmel bir şekilde çıkarttığı ve gelecekteki çalışmalarında araştırmacılar tarafından tercih edilebileceği görülmüştür. Genel olarak. AGA (LASSO) algoritması, AGA (TMFG) algoritmasından daha iyi performans göstermiştir. Bu nedenle, tek boyutlu ve çok yanıt kategorisine

sahip veri setleri için GLASSO algoritmasının TMFG algoritmasına tercih edilmesi gerektiği görülmüştür. AGA (GLASSO)'nın üstün performansı, geleneksel KR1 yöntemiyle karşılaştırıldığında da gözlenmiştir. Bu nedenle, AGA (LASSO)'nın geleneksel KR1 yöntemini tercih eden araştırmacılar için önemli ve etkili bir alternatif olduğu söylenebilir. AGA'nın sağladığı bilgilerin zenginliği göz önüne alındığında, araştırmacılar tarafından tercih edilmesi özellikle önerilmektedir. Ek olarak, örneklem büyüklüğü 1000 veya 3000'e yükseltildiğinde AGA (LASSO) yöntemi, OC ve PA yöntemleriyle de karşılaştırılabilir sonuçlar vermiştir. Örneklem büyüklüğü daha küçük ise (250 veya 500), AGA yalnızca yüksek iç tutarlılığa sahip ve daha az madde içeren ölçüm araçları için ciddi bir alternatif olarak düşünülmelidir.

Faktöriyel ANOVA sonuçlarına göre AGA (LASSO) yöntemi için tekil veya etkileşim etkisinin gözlemlenmediği bulunmuştur. OC ve PA yöntemleri için de benzer bulgular gözlemlenmiştir. Bu üç yöntemin test edilen koşullar arasında en dayanıklı yöntemler olduğu söylenebilir. Ayrıca bu istatistikler AF için hesaplanamamıştır çünkü bu yöntem her koşulda mükemmel sonuçlar ortaya koymaktadır. Başka bir anlatımla, bu yöntemi sağlam yöntem olarak değerlendirmek mümkündür. Öte yandan, örneklem büyüklüğü koşulunun AGA (TMFG) yönteminde "büyük" etkiye sahip olduğu gözlemlenmiştir. Yani örneklem büyüklüğü AGA (TMFG) yönteminin faktör sayısına karar verme performansı üzerinde etkiye sahiptir. Son olarak, geleneksel olarak en yaygın kullanılan KR1 yöntemi için tüm koşullar ve bunların ikili ve üçlü etkileşimleri için "büyük" etkiler gözlemlenmiştir. Buna göre bu çalışmada incelenen koşullar bağlamında KR1 yönteminin en az sağlam yöntem olduğu söylenebilir. Bu bulgu, ilgili alan yazın ile uyumludur (Velicer. Eaton. Fava. 2000; Ruscio & Roche. 2012).

Bu çalışma, AGA' nın faktör çıkarma etkinliğini diğer geleneksel yöntemlerle karşılaştıran birkaç çalışmadan biridir (Golino & Epskamp. 2017; Golino ve ark. 2020). Bu nedenle, AGA' nın farklı koşullarda etkinliğini incelemek için daha fazla araştırmaya ihtiyaç vardır. Örneğin, farklı yetenek dağılımlarına sahip veri setlerinde AGA' nın etkinliği, mevcut alan yazının zenginliğine katkıda bulunacaktır. Ayrıca, bu çalışmada yöntemlerin etkinliği yalnızca faktör sayısını kesin ve yansız çıkartabilme açısından değerlendirilmiştir. Gelecek çalışmalarda, AGA'nın performansının gerçek faktör yüklerini tahmin etme açısından değerlendirilmesi önerilmektedir.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(1); 15-27

Investigation of Classification Accuracy, Test Length and Measurement Precision at Computerized Adaptive Classification Tests *

Seda DEMİR ** Burcu ATAR ***

Abstract
This study aims to compare Sequential Probability Ratio Test (SPRT) and Confidence Interval (CI) classification criteria, Maximum Fisher Information method on the basis of estimated-ability (MFI-EB) and Cut-Point (MFI-CB) item selection methods while ability estimation method is Weighted Likelihood Estimation (WLE) in Computerized Adaptive Classification Testing (CACT), according to the Average Classification Accuracy (ACA), Average Test Length (ATL), and measurement precision under content balancing (Constrained Computerized Adaptive Testing: CCAT and Modified Multinomial Model: MMM) and item exposure control (Sympson-Hetter Method: SH and Item Eligibility Method: IE) when the classification is done based on two, three, or four categories for a unidimensional pool of dichotomous items. Forty-eight conditions are created in Monte Carlo (MC) simulation for the data, generated in R software, including 500 items and 5000 examinees, and the results are calculated over 30 replications. As a result of the study, it was observed that CI performs better in terms of ATL, and SPRT performs better in ACA and correlation, bias, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) values, sequentially; MFI-EB is more useful than MFI-CB. It was also seen that MMM is more successful in content balancing, whereas CCAT is better in terms of test efficiency (ATL and ACA), and IE is superior in terms of item exposure control though SH is more beneficial in test efficiency. Besides, increasing the number of classification categories increases ATL but decreases ACA, and it gives better results in terms of the correlation, bias, RMSE, and MAE values.

Key Words: Computerized adaptive classification testing, content balancing, item exposure control, classification criteria, item selection methods.

INTRODUCTION

Testing in education might have various objectives. These objectives include increasing the effectiveness of education, assessing students individually, making selection or placement decisions, certification, monitoring learning progress, and testing for diagnostic purposes. To achieve these objectives, it seems to be critical to have access to timely and accurate information about learners' level of ability. In this regard, Computerized Adaptive Testing (CAT) is one of the greatest reflections of developments in information and communication technologies in the field of education and contributes to making more qualified and effective evaluations.

Unlike traditional paper-pencil tests, a CAT system uses different test forms in real time based on their individualized performance to test individuals with different levels of ability (Bao, Shen, Wang, & Bradshaw, 2021). The goal of CAT is to estimate each individual's latent ability and select the most appropriate test items (i.e., the most informative item) from the item pool for an individual based on his or her current performance (Eggen & Straetmans, 2000). At the end of the process, CAT provides more reliable estimates of ability using fewer items compared to traditional tests (Bao et al., 2021;

* This study is based on doctoral dissertation entitled "Investigation of Classification Accuracy at Computerized Adaptive Classification Tests"

** Assist. Prof., Tokat Gaziosmanpasa University, Faculty of Education, Tokat-Turkey, seddadmr@gmail.com, ORCID ID: 0000-0003-4230-5593
*** Assoc. Prof., Hacettepe University, Faculty of Education, Ankara-Turkey, burcua@hacettepe.edu.tr, ORCID ID: 0000-0003-3527-686X

To cite this article:
Demir, S., & Atar, B. (2021). Investigation of classification accuracy, test length and measurement precision at computerized adaptive classification tests. *Journal of Measurement and Evaluation in Education and Psychology, 12*(1), 15-27. doi: 10.21031/epod.787865

Received: 30.08.2020
Accepted: 21.02.2021

Fan, Wang, Chang, & Douglas, 2012; Thompson, 2009). These advantages of CAT can be seen as the main reason for preferring large scale CAT applications such as the Graduate Management Admission Test (GMAT), the Graduate Record Examination (GRE), and the National Assessment of Educational Progress (NAEP). The main purpose of testing individuals may sometimes be the accuracy of classifications, such as passed or failed, apart from the effective estimate of ability. In that case, a Computerized Adaptive Classification Test (CACT) is preferred. Since important decisions are made based on the classification (e.g., retention, high school graduation, career selection), efficient and accurate classification is of critical importance (Thompson & Ro, 2007).

Additionally, test effectiveness is important for both CATs and CACTs. High test effectiveness in CAT applications with a unidimensional item pool means fewer items and lower standard errors for ability estimation (van der Linden & Hambleton, 1996 as cited in Thompson, 2009). Unlike CATs, CACTs use as few items as possible and aim at low classification errors to achieve test effectiveness (Thompson, 2009).

Purpose of the Study

An extensive review of literature on CACT applications revealed that most of the studies considered classification in only two categories (e.g., Gündeğer & Doğan, 2018a; Lau, 1996; Reckase, 1983; Spray & Reckase, 1996), and content balancing and item exposure control were not taken into account. Furthermore, classification criteria (e.g., Kingsbury & Weiss, 1980; Spray & Reckase, 1996; Thompson, 2009) and item selection methods were mostly compared (e.g., Gündeğer & Doğan, 2018b; Eggen, 1999; Lin & Spray, 2000), and the performance of different item selection methods was examined by crossing the item selection methods with classification criteria (e.g., Eggen & Straetmans, 2000; Thompson & Ro, 2007). Besides, there are a few studies that compared the performance of classification criteria in terms of Average Classification Accuracy (ACA) and Average Test Length (ATL) according to different item exposure control methods (Huebner, 2012; Lau & Wang, 1999). A study used the Sympson-Hetter (SH) item exposure control method together with the spiral method for content balancing (Huebner & Li, 2012). Considering the contribution of accurate classifications to selecting, monitoring, or placing individuals based on the test results, there seems to be a need for new research in CACT using different research designs. It is thus thought that this study will contribute to a deeper understanding of CACT applications.

The main purpose of this study was to examine the performance of different classification criteria and item selection methods used in CACT applications when weighted likelihood estimation (WLE) is used for ability estimation under various conditions of classification category numbers, content balancing, and item exposure control methods in terms of average classification accuracy, average test length, the correlation between true and estimated ability levels, bias, root mean squared error (RMSE), and mean absolute error (MAE). The research problems are as follows:

Given that WLE is the ability estimation method, and the sequential probability ratio test (SPRT) with indifference region (IR) constant value δ: .20, and the confidence interval with CI: 90% confidence level are the classification criteria, how do the values of average classification accuracy, average test length, the correlation between true and estimated ability levels, bias, RMSE, and MAE change in two, three or four-category classifications where the followings are considered together?

1. The estimate-based maximum Fisher information (MFI-EB) and cut score-based maximum Fisher information (MFI-CB) item selection methods,
2. The MFI-EB and MFI-CB item selection methods along with the constrained CAT (CCAT) and modified multinomial model (MMM) content balancing methods, and the Sympson-Hetter (SH) and item eligibility (IE) item exposure control methods.

For the purpose of the research, below are described the design of the simulation study, data generation, CACT simulation conditions, and analysis plan. Then, the results are summarized, and the main findings are highlighted. Finally, a discussion is given on the implications of this simulation

study according to ACA, ATL, measurement precision, and its results, and suggestions for future research.

METHOD

In this study, Monte Carlo (MC) simulations were performed, and CACT application results were compared using simulated datasets. If other research methods answer the questions What happened, and how, and why? simulation studies help answer the question What if ...? In simulation studies, it is possible to examine more complex systems as possible different conditions into the future can be created (Dooley, 2002). The datasets used were generated in the R program (R Core Team, 2013) based on the conditions examined in the study. The dependent variables of the study were ACA, ATL, correlation between real ability values and estimated ability values (*r*), bias, RMSE, and MAE. The independent variables were classification criteria (SPRT and CI), item selection methods (MFI-EB and MFI-CB), content balancing methods (CCAT and MMM), item exposure control methods (SH and IE), and the number of classification categories (two, three, and four). Therefore, the study had 48 simulation conditions = 2 classification criteria x 2 item selection methods x 2 content balancing methods x 2 item exposure control methods x 3 classification category numbers.

Data Generation

The data used in this study were generated by simulation in accordance with certain properties.

Generation of item and ability parameters for Monte Carlo (MC) simulation

This study was conducted as an MC simulation study by taking Thompson's (2011) study into consideration. The item pool was composed of 500 items under Item Response Theory (IRT) three-parameter logistic model (3PLM) for each of 30 replications. Since both estimate-based and cut score-based item selection methods (MFI-EB and MFI-CB) were used and two-, three- or four-category classifications were made, the item pool was composed of items that provide a high amount of information at and around the cut-point $\theta = 0$ and cover the ability level range (-3, 3). For the items in the pool, the a parameter was generated from a uniform distribution $U[0.5, 2.0]$ to represent medium and high levels of discrimination considering the study of Kingsbury and Weiss (1980), the *b* parameter was generated from a normal distribution $N(-0.5, 1.5)$ to be close to the actual values in applications as pointed out in Thompson (2009) and Warm (1989), and the *c* parameter was generated from a normal distribution $N(0.20, 0.05)$ again to be close to an actual application in keeping with Thompson (2009). In addition, ability parameters of 5000 examinees were generated from a normal distribution $N(0, 1)$ within a range of (-3, +3) for each of 30 replications.

CACT Simulation Conditions

CACT simulation conditions, used in this study, were explained in detail under subheadings.

Starting point

Available prior information about examinees can be used as the starting point in CACT (Weiss & Kingsbury, 1984; Yang, Poggio, & Glasnapp, 2006). Although not used very often, the population mean can also be defined as the starting point (Thompson, 2007b). In this research, the starting point for all conditions was determined as $\theta = 0$.

Item selection

Intelligent item selection methods where the computer program evaluates the unused items in the pool and decides which would be the best item to use next are generally classified into two groups: estimate-based and cut score-based (Thompson, 2007b). When IRT is used as the psychometric model, the cut score-based methods such as MFI, maximum Kullback-Leibler information (KLI), and log-odds ratio methods can be preferred (Lin & Spray, 2000). Traditionally, an item selection method that maximizes Fisher information at the cut-point is used with SPRT. SPRT is expected to yield better results, especially as the indifference region increases (Eggen, 1999). MFI-EB and MFI-CB methods were used for item selection in this study.

Ability estimation

Based on the literature, there are several ability estimation methods for binary scoring (1-0) and unidimensional item response theory modeling. The most common and widely used ability estimation methods include Maximum Likelihood Estimation (MLE), Marginal Maximum Likelihood Estimation (MMLE), Weighted Likelihood Estimation (WLE), and the Bayesian estimation methods such as Owen's Bayesian sequential method, Maximum A Posteriori (MAP), and expected a posteriori (EAP). Warm (1989) noted that all these methods can produce some biased estimates. Bias affects the accuracy of classification decisions systematically (Wang & Wang, 2001). Additionally, Warm (1989) concluded that, especially in fixed-length tests, estimations made by WLE had less bias compared to estimations made by MLE and MAP. He discussed that when WLE is used for various lengths of adaptive tests, the test is similar to MAP but ends with fewer items than MLE, and he proposed the WLE method, which is a modified version of MLE, for ability estimation. This estimation method may reduce item exposure and test time, thereby enhancing the usefulness of the test. Thus, it can be considered as an advantage to use WLE for CACT and CAT applications. WLE is a method that reduces bias and works on the basis of item parameters and a weighting function specific to ability levels (Warm, 1989). WLE is most often preferred in CACT applications (Eggen & Straetmans, 2000; Nydick, Nozawa, & Zhu, 2012; Wouda & Eggen, 2009; Yang et al., 2006). Considering its advantages and its position in the literature about classification, WLE was used as an ability estimation method in this study. The WLE ability estimation method is a condition that was kept constant in simulations.

Classification criteria

There are three basic classification criteria based on IRT in CACT applications: SPRT, CI, and Bayesian decision theory. All three classification criteria require fewer items than traditional fixed-form tests and provide a similar level of classification accuracy (Kingsbury & Weiss, 1983). Previous research has shown that CI is more effective in estimate-based item selections, while SPRT is more effective in cutscore-based item selections (Eggen & Straetmans, 2000; Spray & Reckase, 1996; Thompson, 2009). It has also been shown that SPRT is more effective than CI, especially in terms of classification accuracy (Eggen, & Straetmans, 2000). Furthermore, as Thompson (2009) pointed out, the most used classification criterion in CACT studies is SPRT. Against this background, the classification criteria were determined as SPRT (δ: .20) and CI (90%) in this study.

Content balancing

In the content-balanced ICT applications, examinees are measured by a test that represents each of the content areas as appropriately as possible and has higher validity. The most commonly used content balancing methods in CACT studies are the spiralling method (Kingsbury & Zara, 1989) (e.g., Finkelman, 2008; Huebner, 2012) and the constrained CAT (CCAT) method (e.g., Eggen & Straetmans, 2000; Huebner & Li, 2012). Lin (2011) used a modified multinomial model (MMM) for content balancing. However, no research has been found that compares CCAT and MMM in the literature. Therefore, in this study, unlike the previous studies, two different content balancing

methods, namely CCAT and MMM, were used. The minimum number of items to be used before terminating the test was set at 10, and the maximum number of items was set at 70 to ensure content balancing conditions. In cases where CCAT and MMM were included in the study conditions, the item pool generated with 500 items in the R program was divided into four content areas using random item assignment. Then, items were selected using the functions and loops written by the researcher in line with these content areas. The target proportions of four content areas were set at 40%, 30%, 20%, and 10%, respectively.

Item exposure

In CAT applications in which the item exposure control is not used, the selection of the items only based on maximum information could result in overexposure of items. On the other hand, both test security and more balanced use of item pool are considered while maintaining measurement precision when item exposure control techniques are implemented (Leroux et al., 2019). A search of the literature showed that the most used item exposure control methods in CACT applications are the random item selection method based on randomness strategies and the SH method (Sympson & Hetter, 1985) based on conditional selection strategies. Because randomness strategies are believed to be not effective under realistic test conditions, this research focused on the SH method and the IE method (van der Linden & Veldkamp, 2004), which is based on the same approach as the SH method. The maximum desired item exposure rate for the SH and IE methods used in the item exposure control was taken as r_{max} = .20 (Leung, Chang, & Hau, 2002), which is a frequently used value in line with the studies of Huebner (2012) and Huebner and Li (2012).

Number of classification categories

Much of the research in CACT so far has used only two categories, such as failed-passed and a single cut-point. A two-category classification such as failed-passed was used in Huebner (2012), Lin and Spray (2000), Reckase (1983), Sie, Finkelman, Riley, and Smits (2015), Thompson (2009), van Groen, Eggen, and Veldkamp (2016). Both two- and three-category classifications were used in Eggen (1999) and Thompson (2007a). A three-category classification was used in Nydick et al. (2012). Both three- and five-category classifications were used in Yang et al. (2006). This research used two-, three- and four-category classifications to compare the changes. The ability parameters generated in R for the examinees were utilized to determine the cutting points for the classifications. The generated ability parameters were ranked from the low ability level to the high ability level. Through the method used in Eggen and Straetmans (2000), a cut-point was determined for the two-category classification, two cut-points were determined for the three-category classification, and three cut-points were determined for the four-category classification. In the two-category classification, the first half of the skill levels ranked from low to high were coded as Level 1 and the second half as Level 2. Then, the cut-point (CP = 0.00) was determined by taking 70% of the highest ability level in Level 1. Similarly, in the three-category classification, the ranked ability levels were encoded as Level 1, Level 2, and Level 3, and the cut-points were defined as CP1 = -0.29 and CP2 = 0.31. In the four-category classification, the ability levels were encoded as Level 1, Level 2, Level 3, and Level 4 and the cut-points were defined as CP1 = -0.47, CP2 = -0.01, and CP3 = 0.48.

Data Analysis

Thirty replications were conducted for each of the 48 simulation conditions generated within the scope of the research, and the values of the dependent variables were obtained by calculating the average of the replications. The value of the correlation between true and estimated ability levels was calculated using the Pearson correlation coefficient (PCC), while the bias, RMSE, and MAE values were calculated following formulas written in the R program.

Bias is calculated using the formula below where the sum of the difference between the last estimated ability level ($\widehat{\theta_i}$) and the true ability level (θ_i) is divided by the number of examinees (n) (Miller, & Miller, 2004):

$$Bias = \frac{\sum_{i=1}^{n}(\widehat{\theta_i} - \theta_i)}{n}$$

RMSE is equal to the square root of the sum of squared of differences between the $\widehat{\theta_i}$ and θ_i divided by n:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{\theta_i} - \theta_i)^2}{n}}$$

MAE is calculated by dividing the sum of the absolute value of the difference between $\widehat{\theta_i}$ and θ_i by n:

$$OMH = \frac{\sum_{i=1}^{n}|\widehat{\theta_i} - \theta_i|}{n}$$

Additionally, functions and loops were written in the R program in addition to the item selection method for content balancing and item exposure control.

RESULTS

The results obtained for each subproblem of the study are presented under subheadings.

Results on the First Subproblem

Table 1 shows the values calculated by averaging 30 replications performed for each simulation condition related to the first research subproblem.

Table 1. Comparison of the Classification Criteria (CC) and Item Selection Methods (ISM) According to the Average Test Length (ATL), Average Classification Accuracy (ACA), and Measurement Precision With Correlation (r), Bias, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) Values When the number of Classification Categories (NCC) Based on Two, Three, or Four

CC	ISM	NCC	ATL	ACA	*r*	Bias	RMSE	MAE
SPRT (δ = .20)	MFB-EB	Two	24.72	.94	.94	-0.011	0.35	0.27
		Three	34.08	.88	.96	-0.012	0.32	0.24
		Four	41.34	.82	.96	-0.014	0.29	0.22
	MFB-CB	Two	22.95	.94	.90	0.019	0.44	0.32
		Three	33.93	.89	.92	0.015	0.38	0.28
		Four	42.88	.82	.93	0.012	0.35	0.26
CI (90%)	MFB-EB	Two	11.33	.89	.90	0.016	0.46	0.35
		Three	12.52	.79	.91	0.015	0.45	0.35
		Four	13.81	.71	.91	0.016	0.44	0.34
	MFB-CB	Two	11.55	.90	.87	0.019	0.49	0.38
		Three	12.62	.80	.87	0.017	0.48	0.37
		Four	13.82	.71	.88	0.020	0.47	0.36

Note. SPRT= sequential probability ratio test, CI= confidence interval, MFI-EB= maximum fisher information method on the basis of estimated-ability, MFI-CB= maximum fisher information method on the basis of cut-point.

As seen in Table 1, in the two-, three- and four-category classifications, the ACA values were quite high and ranged from .82 to .94, and the ATL values ranged from 22.95 to 42.88 when SPRT was used for classification. On the other hand, when CI was used for classification, the ACA values were relatively lower and ranged from .71 to .90, and the ATL values ranged from 11.33 to 13.82.

Accordingly, SPRT yielded better results in terms of ACA, and CI yielded better results in terms of ATL.

When the item selection methods MFI-EB and MFI-CB were used with the same classification criteria, similar results were obtained in terms of test effectiveness. In addition, an increase in the number of classification categories caused the test effectiveness to decrease for both classification criteria. In other words, it increased the ATL but reduced the ACA.

The values of the correlation (*r*) between the examinees' estimated and true ability levels ranged from .90 to .96 for SPRT and .87 to .91 for CI. With respect to the conditions in which the classification criteria were crossed by the item selection methods, higher correlations were calculated for both classification criteria in the conditions in which MFI-EB was used compared to the conditions in which MFI-CB was used. Additionally, similar correlation values were obtained in response to the increase in the number of classification categories. The bias calculated for the condition where SPRT and MFI-EB were used together (ranging from -0.014 to -0.011) was lower compared to that calculated for the condition where SPRT and MFI-CB were used together (ranging from 0.012 to 0.019). Similarly, the bias calculated for the condition where CI and MFI-EB were used together (ranging from 0.015 to 0.016) was lower compared to that calculated for the condition where CI and MFI-CB were used together (ranging from 0.017 to 0.020). The case is similar for the RMSE value, which takes into account the standard error of the estimation along with the bias, and for the MAE value. Accordingly, it can be said that lower bias, RMSE, and MAE values were found when the SPRT classification criterion or the MFI-EB item selection method was used. Furthermore, the increase in the number of categories did not exert a great effect on the bias but relatively decreased the RMSE and MAE values.

Results on the Second Subproblem

Table 2 demonstrates the values calculated by averaging 30 replications performed for each condition related to the second research subproblem, which incorporated CCAT and MMM for content balancing and SH and IE for item exposure control.

As seen in Table 2, in all conditions where the MMM content balancing method was used, the used content rates achieved the desired content rates (40%, 30%, 20%, and 10%, respectively). In the conditions where the CCAT content balancing method was used, the used content rates were above or below the desired content rates. For example, as seen in Table 2, in the condition where SPRT was used with MFI-CB, item exposure was controlled using IE, and a four-category classification was made, the CCAT content rates were found to be approximately 32%, 28%, 23%, and 16%, respectively. In addition, in the conditions where the IE item exposure control method was used, the proportion of items overexposed (OEX) was lower and the mean exposure rate of overexposed items (MOEX) achieved the desired r_{max} = .20. On the other hand, in the conditions where SH was used, OEX was higher, and MOEX was considerably higher than the desired r_{max}= .20. For example, as seen in Table 2, when SPRT and MFI-EB were used together, content balancing was done using CCAT, and a four-category classification was made, the OEX value calculated for item exposure controlled using SH was approximately .25, and the MOEX value was .29. In other words, approximately 25% of the items were above the maximum item exposure rate (r_{max} = .20), and the mean item exposure was calculated to be approximately .29.

As seen in Table 2, another comparison using the same classification criteria and item selection method showed that although the CCAT content balancing method performed better with a slight difference in terms of test effectiveness, it generally produced similar results to MMM. In addition, the SH item exposure control method performed better compared to IE in terms of test effectiveness. The best result in terms of ATL (ATL = 11.13 and ACA = .88) was recorded in the condition where CI, MFI-EB, CCAT, and SH were used together, and a two-category classification was made, while the worst result (ATL = 51.93 and ACA = .75) was recorded in the condition where SPRT, MFI-CB, MMM, and IE were used together, and a four-category classification was made. To put it differently, it can be said

that among the best and worst results, ATL was nearly five times higher, while ACA declined considerably.

Table 2. Comparison of The Classification Criteria (CC) and Item Selection Methods (ISM) According to the Average Test Length (ATL), Average Classification Accuracy (ACA), and Measurement Precision With Correlation (R), Bias, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) Values Under Content Balancing Methods (CBM) With Applied Content Rates and Item Exposure Control Methods (IECM) With Proportion of Items Overexposed (OEX), and Mean Exposure Rate of Overexposed Items (MOEX) When the Number of Classification Categories (NCC) Based on Two, Three, or Four

CC	ISM	CBM	IECM	NCC	Applied Content Rates				OEX	MOEX	ATL	ACA	*r*	Bias	RMSE	MAE
SPRT	MFB-EB	CCAT	SH	Two	36.27	29.61	21.78	12.34	.14	.28	26.91	.94	.94	-0.014	0.36	0.28
(δ= .20)			r_{max} = .20	Three	34.34	28.95	22.56	14.15	.21	.29	37.50	.87	.95	-0.015	0.32	0.24
				Four	33.32	28.65	22.94	15.09	.25	.29	44.69	.81	.96	-0.017	0.30	0.22
			IE	Two	35.56	29.38	22.09	12.97	.09	.20	29.87	.93	.94	-0.017	0.37	0.28
			r_{max} = .20	Three	33.71	28.77	22.79	14.74	.15	.20	41.71	.86	.95	-0.018	0.33	0.25
				Four	32.81	28.47	23.15	15.57	.18	.20	48.50	.78	.96	-0.018	0.31	0.23
		MMM	SH	Two	39.82	29.98	20.09	10.10	.14	.28	27.42	.94	.94	-0.015	0.37	0.28
			r_{max} = .20	Three	39.91	29.98	20.04	10.07	.21	.29	37.86	.87	.95	-0.015	0.33	0.25
				Four	39.92	29.99	20.05	10.04	.26	.29	45.35	.80	.96	-0.017	0.30	0.23
			IE	Two	39.80	30.03	20.06	10.11	.10	.20	30.82	.93	.94	-0.015	0.37	0.28
			r_{max} = .20	Three	39.84	30.01	20.08	10.08	.15	.20	42.27	.85	.95	-0.016	0.33	0.25
				Four	39.86	30.01	20.07	10.07	.17	.20	49.01	.77	.96	-0.018	0.32	0.24
	MFB-CB	CCAT	SH	Two	37.00	29.71	21.46	11.83	.13	.33	25.70	.94	.90	0.009	0.44	0.33
			r_{max} = .20	Three	34.41	28.98	22.49	14.12	.21	.34	38.35	.87	.92	0.009	0.39	0.28
				Four	32.96	28.56	23.07	15.41	.25	.36	47.02	.79	.93	0.006	0.36	0.26
			IE	Two	35.83	29.40	21.96	12.81	.11	.20	30.55	.93	.90	0.002	0.44	0.33
			r_{max} = .20	Three	33.46	28.74	22.90	14.91	.19	.20	43.94	.84	.92	0.003	0.40	0.30
				Four	32.40	28.39	23.30	15.92	.23	.20	51.14	.75	.93	0.000	0.38	0.28
		MMM	SH	Two	39.81	29.96	20.12	10.11	.13	.33	26.18	.94	.90	0.009	0.43	0.33
			r_{max} = .20	Three	39.88	30.00	20.08	10.04	.21	.34	38.61	.87	.92	0.008	0.39	0.29
				Four	39.92	30.01	20.03	10.05	.25	.36	47.36	.79	.93	0.006	0.36	0.26
			IE	Two	39.75	29.95	20.16	10.13	.12	.20	31.01	.93	.90	0.005	0.44	0.33
			r_{max} = .20	Three	39.83	30.00	20.07	10.10	.19	.20	44.85	.84	.92	0.003	0.40	0.30
				Four	39.84	30.01	20.05	10.09	.23	.20	51.93	.75	.93	0.006	0.39	0.29

(continued)

Table 2 (continue)

CC	ISM	CBM	IECM	NCC		Applied Content Rates		OEX	MOEX	ATL	ACA	*r*	Bias	RMSE	MAE
CI	MFB-EB	CCAT	SH	Two	42.13	30.95	19.09	7.83	.04	.27	11.13	.88	.89	0.011	0.48	0.37
90%			r_{max} = .20	Three	41.60	30.85	19.33	8.22	.05	.27	12.24	.78	.90	0.014	0.48	0.37
				Four	41.01	30.69	19.64	8.66	.05	.27	13.37	.70	.90	0.013	0.47	0.36
			IE	Two	42.04	30.90	19.17	7.90	.03	.20	11.19	.88	.89	0.016	0.49	0.38
			r_{max} = .20	Three	41.56	30.80	19.40	8.24	.03	.20	12.25	.78	.89	0.013	0.49	0.37
				Four	41.03	30.70	19.64	8.63	.03	.20	13.39	.69	.90	0.014	0.48	0.37
		MMM	SH	Two	39.97	30.04	19.96	10.04	.04	.27	11.18	.88	.89	0.015	0.48	0.37
			r_{max} = .20	Three	39.95	30.01	20.01	10.02	.05	.27	12.28	.78	.90	0.011	0.47	0.36
				Four	40.07	29.93	20.00	9.99	.05	.27	13.45	.70	.90	0.012	0.47	0.36
			IE	Two	40.06	29.94	20.01	9.99	.03	.20	11.21	.88	.89	0.012	0.49	0.38
			r_{max} = .20	Three	40.05	30.01	19.95	9.98	.03	.20	12.31	.78	.90	0.012	0.48	0.37
				Four	40.01	30.00	19.99	9.99	.03	.20	13.51	.69	.90	0.013	0.48	0.37
	MFB-CB	CCAT	SH	Two	41.96	30.93	19.19	7.92	.05	.36	11.36	.89	.86	0.012	0.50	0.39
			r_{max} = .20	Three	41.49	30.77	19.40	8.34	.06	.36	12.46	.79	.87	0.01	0.49	0.39
				Four	40.88	30.64	19.68	8.80	.06	.36	13.69	.70	.87	0.01	0.48	0.38
			IE	Two	42.02	30.90	19.13	7.96	.05	.20	11.42	.88	.85	0.009	0.52	0.41
			r_{max} = .20	Three	41.43	30.75	19.46	8.37	.05	.20	12.56	.77	.86	0.004	0.51	0.40
				Four	40.55	30.55	19.86	9.03	.05	.20	14.6	.66	.86	0.006	0.51	0.39
		MMM	SH	Two	39.97	30.01	20.00	10.01	.06	.35	11.37	.89	.86	0.012	0.50	0.39
			r_{max} = .20	Three	39.91	30.02	20.07	9.99	.06	.35	12.56	.79	.87	0.009	0.49	0.38
				Four	39.97	30.02	20.00	10.01	.07	.35	13.65	.69	.87	0.013	0.49	0.38
			IE	Two	40.00	30.02	19.99	9.98	.05	.20	11.49	.88	.86	0.006	0.52	0.41
			r_{max} = .20	Three	40.04	29.98	20.01	9.97	.05	.20	12.74	.78	.86	0.008	0.51	0.40
				Four	39.99	29.98	20.02	10.00	.05	.20	14.42	.67	.87	0.009	0.50	0.39

Note: SPRT= sequential probability ratio test, CI= confidence interval, MFI-EB= maximum fisher information method on the basis of estimated-ability, MFI-CB= maximum fisher information method on the basis of cut-point, CCAT= constrained computerized adaptive testing, MMM= modified multinomial model, SH= Sympson-Hetter method, IE= item eligibility method and r_{max}= maximum desired item exposure rate.

The correlation (*r*) values ranged from .90 to .96 in the conditions where SPRT was used, while they ranged from .85 to .90 in the conditions where CI was used. The bias values ranged from -0.018 to

0.009 in the conditions where SPRT was used, while they ranged from 0.004 to 0.016 in the conditions where CI was used. The highest RMSE value (0.52) and the highest MAE value (0.41) were observed when CI, MFI-CB, CCAT (or MMM), and IE were used together, and a two-category classification was made. On the other hand, the lowest RMSE value (0.30) was observed when SPRT, MFI-EB, CCAT (or MMM), and SH were used together with four-category classification, and the lowest MAE value (0.22) was observed when SPRT, MFI-EB, CCAT, and SH were used together with four-category classification.

In summary, parallel to the findings in Table 1, CI performed better in terms of ATL, while SPRT performed better in terms of ACA. As the number of classification categories increased, ATL increased but ACA decreased. With respect to the correlation (*r*), bias, RMSE, and MAE values, SPRT performed better than CI, and MFI-EB performed better than MFI-CB. Furthermore, in response to the increased number of categories, the correlation and bias resulted in similar values, while the RMSE and MAE values were relatively lower.

DISCUSSION and CONCLUSION

Because the primary focus of this study is on classification accuracy, the ACA values calculated under different conditions are of great importance in interpreting the findings. In line with the research findings, high ACA values were calculated under all research conditions. The SPRT classification criterion performed better than CI and achieved a higher rate of classifying examinees into the accurate categories. On the other hand, the CI classification criterion performed better in terms of ATL under all research conditions and required fewer items to classify examinees compared to SPRT. This finding is in agreement with those obtained by Gündeğer and Doğan (2018a), Nydick et al. (2012), Thompson (2009), and Thompson and Ro (2007). These studies, in general, reported that the classifications made using CI ended with lower ATL and ACA compared to those made using SPRT. Therefore, comparing the SPRT and CI classification criteria used in the research in terms of classification accuracy, it may be suggested to prefer SPRT which yielded higher ACA values. On the other hand, comparing SPRT and CI in terms of ATL, CI seems to be preferable as it requires fewer items to classify examinees and terminate the test. Nevertheless, it should be noted that with respect to high-risk tests (e.g., tests applied in the field of medicine and directly related to human life), it is of key importance to choose the method which achieves a higher classification accuracy despite the increasing number of items. In CACTs, ATL, and ACA are often evaluated together for test effectiveness. If a decision is to be made to choose the best performing classification criterion in terms of test effectiveness, it may be suggested to use CI for conditions where both classification criteria achieve a good level of classification accuracy.

This research found that the SPRT classification criterion performed better than CI, and the MFI-EB item selection method performed better than MFI-CB in terms of measurement precision. Accordingly, under the conditions where the SPRT classification criterion or the MFI-EB item selection method was used, the values of correlation between examinees' true and estimated ability levels were higher while the bias, RMSE, and MAE values were lower. It can thus be said that examinees' last ability levels were more precise and closer to their true ability levels when the classification criterion was SPRT or when the item selection method was MFI-EB. A possible explanation of this result might be that the item pool was composed of items that provide great information at and around the cutting point $\theta = 0$. Additionally, the MFBI-EB item selection method achieved relatively better results compared to MFI-CB in terms of test effectiveness. In other words, when MFBI-EB was used, lower ATL values and similar ACA values were obtained.

The analysis results showed that the values of correlation between examinees' true and estimated ability levels were quite high, especially when the WLE ability estimation method was used together with the SPRT classification criterion and the MFI-EB item selection method. It can thus be said that the WLE method performs successfully.

Comparing the findings presented in Table 1 and Table 2, it can be seen that relatively higher ATL and lower ACA values were obtained in line with expectations when content balancing and item exposure control were added to the research conditions. According to Thompson (2007b), content

balancing and item exposure constraints generally lead to an increase in only ATL. When content balancing and item exposure control are performed in CACT applications, it can be interpreted that the increase in ATL and the decrease in ACA may be due to the absence of an item that provides sufficient information about an examinee in the applied content area and does not exceed the item exposure rate. To solve this problem, the item pool might be expanded by increasing the number of items in each content area within the ability range which has plenty of items that exceed the maximum item exposure rate. The content balancing and item exposure control methods included in the research conditions did not change the correlation between examinees' true and estimated ability levels but caused a decrease in the bias values and an increase in RMSE and MAE values. The results obtained by the CI classification criterion were also little affected. This can be interpreted as an advantage provided by CI.

The research found that the MMM content balancing method performed better in achieving the desired content rates compared to CCAT. On the other hand, with respect to test effectiveness, CCAT performed better, especially in terms of ATL when SPRT was used although there were slight changes when CI was used. This finding is consistent with that reported by Lin (2011). Lin (2011) emphasized that although CCAT is one of the most chosen content balancing methods in CACTs, the MMM method, which is used mostly in CATs, is more successful in achieving the desired content balance. Therefore, in CACTs it is suggested to use MMM if content balancing is more critical as in high-risk tests, and CCAT if test effectiveness is more critical. The research also found that the IE method performed better in controlling item exposure compared to the SH method. This finding is in line with the work of Huebner (2012). Huebner (2012) concluded that IE works more successfully than SH in terms of item exposure control. In terms of test effectiveness, SH performed better, especially under the conditions where the SPRT classification criterion was used. When the SH method was used, lower ATL and higher ACA values were obtained. Thus, IE might be used if item exposure control, namely the safety of the test/item pool, is of critical importance in CACTs. Whereas SH might be used if test effectiveness is of more critical importance.

Under all research conditions, the increasing number of categories increased ATL while reducing ACA. To put it differently, the increasing number of categories reduced test effectiveness. This finding supports earlier observations in Eggen (1999) and Nydick et al. (2012). Eggen (1999) compared two-category and three-category classifications, and Nydick et al. (2012) compared three-category and five-category classifications. They found that the higher the number of categories was the higher the ATL values and the lower the ACA values were; thus, test effectiveness decreased. Therefore, in terms of test effectiveness, it may be suggested to keep the number of classification categories as few as possible. In addition, despite the increase in the number of classification categories, the correlation and bias values were similar, while RMSE and MAE values were relatively lower. Accordingly, examinees' last ability levels were more precisely estimated because the number of items required to terminate the test increased with the increasing number of classification categories. Therefore, it seems that the number of classification categories might be determined more optimally by considering correlation, bias, RMSE, and MAE values.

Based on the research findings, the following suggestions might be offered for future practice. If the focus of CACT is on ACA and content balancing and item exposure control are of critical importance, the SPRT classification criterion, which also performs better in terms of correlation, bias, RMSE, and MAE values, might be used together with the MFI-EB item selection method, the MMM content balancing method, and the IE item exposure control method. If the focus of CACT is on ATL and content balancing and item exposure control are performed, the CI classification criterion might be used together with MFI-EB, MMM, and IE. As for the researchers, in similar BBST studies, it can be recommended to use item pools with different properties such as multi-dimensional item pool or different pool sizes, skewness, kurtosis, etc. In addition, in similar studies to be conducted, the performances of the main BBST components can be compared over real data.

REFERENCES

Bao, Y., Shen, Y., Wang, S., & Bradshaw, L. (2021). Flexible computerized adaptive tests to detect misconceptions and estimate ability simultaneously. *Applied Psychological Measurement, 45*(1), 3-21. doi: 10.1177/0146621620965730

Dooley, K. (2002). Simulation research methods. In J. Baum (Ed.), *Companion to organizations* (pp. 829-848). London: Blackwell.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*(3), 249-261. doi: 10.1177/01466219922031365

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60*(5), 713-734. doi: 10.1177/00131640021970862

Fan, Z., Wang, C., Chang, H., & Douglas, J. (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics, 37*(5), 655-670. doi: 10.3102/1076998611422912

Finkelman, M. (2008). On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *Journal of Educational and Behavioral Statistics, 33*(4), 442-463. doi: 10.3102/1076998607308573

Gündeğer, C., & Doğan, N. (2018a). A comparison of computerized adaptive classification test criteria in terms of test efficiency and measurement precision. *Journal of Measurement and Evaluation in Education and Psychology, 9*(2), 161-177. doi: 10.21031/epod.401077

Gündeğer, C., & Doğan, N. (2018b). The effects of item pool characteristics on test length and classification accuracy in computerized adaptive classification testings. *Hacettepe University Journal of Education, 33*(4), 888-896. doi: 10.16986/HUJE.2016024284

Huebner, A. (2012). Item overexposure in computerized classification tests using sequential item selection. *Practical Assessment, Research & Evaluation, 17*(12), 1-9. Retrieved from https://pareonline.net/getvn.asp?v=17&n=12

Huebner, A., & Li, Z. (2012). A stochastic method for balancing item exposure rates in computerized classification tests. *Applied Psychological Measurement, 36*(3), 181-188. doi: 10.1177/0146621612439932

Kingsbury, G. G., & Weiss, D. J. (1980). *A Comparison of adaptive, sequential and conventional testing strategies for mastery decisions* (Research Report 80-4). University of Minnesota, Minneapolis: MN. Retrieved from http://iacat.org/sites/default/files/biblio/ki80-04.pdf

Kingsbury, G. G., & Weiss, D.J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing,* (pp. 237-254). New York: Academic Press.

Kingsbury, G. G., & Zara, A.R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*(4), 359-375. doi: 10.1207/s15324818ame0204_6

Lau, C. A. (1996). *Robustness of a unidimensional computerized testing mastery procedure with multidimensional testing data* (Unpublished doctoral dissertation). University of Iowa, Iowa City IA.

Lau, C. A., & Wang, T. (1999, April). *Computerized classification testing under practical constraints with a polytomous model*. Paper presented at the annual meeting of the American Educational Research Association (AERA), Montreal, Canada. Retrieved from http://iacat.org/sites/default/files/biblio/la99-01.pdf

Leroux, A. J., Waid-Ebbs, J. K., Wen, P-S., Helmer, D. A., Graham, D. P., O'Connor, M. K, & Ray, K. (2019). An investigation of exposure control methods with variable-length cat using the partial credit model. *Applied Psychological Measurement, 43*(8),624-638. doi: 10.1177/0146621618824856

Leung, C.-K., Chang, H. H., & Hau, K. T. (2002). Item selection in computerized adaptive testing: Improving the a-stratified design with the Sympson–Hetter algorithm. *Applied Psychological Measurement, 26*(4), 376-392. doi: 10.1177/014662102237795

Lin, C. (2011). Item selection criteria with practical constraints for computerized classification testing. *Applied Psychological Measurement 71*(1), 20-36. doi: 10.1177/0013164410387336

Lin, C. J., & Spray, J. (2000). *Effects of item-selection criteria on classification testing with the sequential probability ratio test*. ACT (Research Report 2000-8). Iowa city, IA: ACT Research Report Series. Retrieved from https://eric.ed.gov/?id=ED445066

Miller, I., & Miller, M. (2004). *John E. Freund's mathematical statistics with applications*. (7th Ed.). New Jersey: Prentice Hall.

Nydick, S. W., Nozawa, Y., & Zhu, R. (2012, April). *Accuracy and efficiency in classifying examinees using computerized adaptive tests: An application to a large-scale test.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Vancouver, British Columbia, Canada. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.476.3381&rep=rep1&type=pdf

R Core Team (2013). *R: A language and environment for statistical computing*, (Version 3.0.1) [Computer software], Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: latent trait theory and computerized adaptive testing*, (pp. 237-254). New York: Academic Press.

Sie, H., Finkelman, M. D., Riley, B., & Smits, N. (2015). Utilizing response times in computerized classification testing. *Applied Psychological Measurement, 39*(5), 389-405. doi: 10.1177/0146621615569504

Spray, J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics, 21*(4), 405-414. doi: 10.3102/10769986021004405

Sympson, J. B., & Hetter, R. D. (1985, October). Controlling item exposure rates in computerized adaptive testing. In *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 937-977). San Diego, CA: Navy Personnel Research and Development Center. Retrieved from http://www.iacat.org/content/controlling-item-exposure-rates-computerized-adaptive-testing

Thompson, N. A. (2007a). *A comparison of two methods of polytomous computerized classification testing for multiple cutscores* (Unpublished doctoral dissertation). University of Minnesota, Minneapolis.

Thompson, N. A. (2007b). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment Research & Evaluation, 12*(1), 1-13. Retrieved from http://www.iacat.org/sites/default/files/biblio/th07-01.pdf

Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement, 69*(5), 778-793. doi: 10.1177/0013164408324460

Thompson, N. A. (2011). Termination criteria for computerized classification testing. *Practical Assessment, Research & Evaluation, 16*(4), 1-7. Retrieved from https://pareonline.net/getvn.asp?v=16&n=4

Thompson, N. A., & Ro, S. (2007). Computerized classification testing with composite hypotheses. In D. J. Weiss (Ed.). *Proceedings of the 2007 GMAC conference on computerized adaptive testing*. Retrieved from http://www.iacat.org/sites/default/files/biblio/cat07nthompson.pdf

Van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics*, *29*(3), 273-291. doi: 10.3102/10769986029003273

Van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2016). Multidimensional computerized adaptive testing for classifying examinees with within-dimensionality. *Applied Psychological Measurement, 40*(6), 387-404. doi: 10.1177/0146621616648931

Wang, S., & Wang, T. (2001). Precision of warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement, 25*(4), 317–331. doi: 10.1177/01466210122032163

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427-450. doi: 10.1007/BF02294627

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*(4), 361-375. Retrieved from https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Wouda, J. T., & Eggen, T. J. H. M. (2009). Computerized classification testing in more than two categories by using stochastic curtailment. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC conference on computerized adaptive testing*. Retrieved from http://iacat.org/sites/default/files/biblio/cat09wouda.pdf

Yang, X., Poggio, J. C., & Glasnapp, D. R. (2006). Effects of estimation bias on multiple category classification with an IRT-based adaptive classification procedure. *Educational and Psychological Measurement, 66*(4), 545-564. doi: 10.1177/0013164405284031

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(1); 28-53

How Reliable Is It to Automatically Score Open-Ended Items? An Application in the Turkish Language *

İbrahim UYSAL ** Nuri DOĞAN ***

Abstract
The use of open-ended items, especially in large-scale tests, created difficulties in scoring open-ended items. However, this problem can be overcome with an approach based on automated scoring of open-ended items. The aim of this study was to examine the reliability of the data obtained by scoring open-ended items automatically. One of the objectives was to compare different algorithms based on machine learning in automated scoring (support vector machines, logistic regression, multinominal Naive Bayes, long-short term memory, and bidirectional long-short term memory). The other objective was to investigate the change in the reliability of automated scoring by differentiating the data rate used in testing the automated scoring system (33%, 20%, and 10%). While examining the reliability of automated scoring, a comparison was made with the reliability of the data obtained from human raters. In this study, which demonstrated the first automated scoring attempt of open-ended items in the Turkish language, Turkish test data of the Academic Skills Monitoring and Evaluation (ABIDE) program administered by the Ministry of National Education were used. Cross-validation was used to test the system. Regarding the coefficients of agreement to show reliability, the percentage of agreement, the quadratic-weighted Kappa, which is frequently used in automated scoring studies, and the Gwet's AC1 coefficient, which is not affected by the prevalence problem in the distribution of data into categories, were used. The results of the study showed that automated scoring algorithms could be utilized. It was found that the best algorithm to be used in automated scoring is bidirectional long-short term memory. Long-short term memory and multinominal Naive Bayes algorithms showed lower performance than support vector machines, logistic regression, and bidirectional long-short term memory algorithms. In automated scoring, it was determined that the coefficients of agreement at 33% test data rate were slightly lower comparing 10% and 20% test data rates, but were within the desired range.

Keywords: Open-ended item, machine learning algorithms, automated scoring, inter-rater reliability, coefficients of agreement.

INTRODUCTION

Individuals experience numerous tests throughout their lives. Tests show differences in individuals' knowledge, skills and abilities. Thus, decisions can be made about them (Geisinger & Usher-Tate, 2016). In recent years, the use of more than one item format in tests has become more popular. In this approach, which is referred to as a mixed-format test, open-ended items with or without restricted responses are used in addition to the multiple-choice items. In multiple-choice items, individuals encounter one right and more than one wrong answer about a problem. In open-ended items with restricted responses, individuals answer questions with a few words, sentences, or paragraphs, while in items with unrestricted responses, they respond in any length they want (Downing, 2009). The combined use of the item types allows to eliminate the limitations of each format (Messick, 1993). For example, using only the multiple-choice items in tests affects the teaching and learning process and lead individuals to study for multiple-choice tests. This situation can restrict original, critical, and higher level thinking skills. However, the use of open-ended items can overcome this limitation.

* The present study is a part of PhD Thesis entitled "The Reliability of Automated Essay Scoring and Its Effect on Test Equating Errors" conducted under the supervision of Nuri DOĞAN and completed by İbrahim UYSAL in 2019.

** PhD., Bolu Abant İzzet Baysal University, Faculty of Education, Bolu-Türkiye, e-posta: ibrahimuysal06@gmail.com, ORCID ID: 0000-0002-6767-0362
*** Prof. PhD., Hacettepe University, Faculty of Education, Ankara-Türkiye, e-posta: nurid@hacettepe.edu.tr, ORCID ID: 0000-0001-6274-2016

To cite this article:
Uysal, İ., & Doğan, N. (2021). How reliable is it to automatically score open-ended items? An application in Turkish language. *Journal of Measurement and Evaluation in Education and Psychology, 12*(1), 28-53. doi: 10.21031/epod.817396

Received: 28.10.2020
Accepted: 14.02.2021

Open-ended items are difficult to apply and take a long time and effort to score (Gierl, Latifi, Lai, Boulais & Champlain, 2014). As the number of individuals and open-ended items to be scored increases, more raters are needed. In addition, many raters need to be trained about scoring. Another problem is that scorers' emotions and cognitive abilities cause bias in scoring (Adesiji, Agbonifo, Adesuyi & Olabode, 2016). As the number of raters increases, the subjectivity in scoring decreases the reliability (Ebel & Frisbie, 1991; Hagge, 2010). Considering the large-scale test applications, one should take into account that scoring open-ended items will significantly increase the cost of the exam (Cohen, Ben-Simon & Hovav, 2003).

Automated scoring is an approach that has gained popularity in the literature among test practitioners in recent years. In automated scoring, a written text is automatically evaluated with computer-aided analysis (Shermis, 2010). The idea of automated item scoring was introduced about 50 years ago by Page (1966), a secondary school teacher, to reduce scoring difficulty (Ramineni & Williamson, 2013). Page (1966) is the developer of the Project Essay Grade (PEG) program. In this first program developed, word length, essay length, comma and preposition numbers, and number of uncommon words were utilized to predict essay scores (Wang & Brown, 2007).

Automated scoring systems can work on different lengths of answers, from short-answer items to essays (Gierl et al., 2014). In other words, automated scoring is able to score open-ended items that have restricted or unrestricted response. It is stated that 90% of the writing skill tasks currently in schools can be evaluated by automated essay scoring systems (Shermis & Burnstein, 2003). In addition to in-class applications, scoring can be done in large-scale tests with automated scoring systems. This approach is used in large-scale tests such as the International GMAT (Graduate Management Admission Test), TOEFL (Test of English as a Foreign Language), and GRE (Graduate Record Examination). The most important advantage of automated scoring systems is that immediate feedback can be given to individuals (Gierl et al., 2014). In the automated scoring process, scoring features can be defined manually on the computer (e.g., the first studies of Page), or scoring behaviors can be automatically mapped to the computer from the scoring made by human raters. Supervised machine learning algorithms, which are used in automated scoring and learn the scoring features, usually use a four-step process (Powers, 2015). These steps are; 1) defining a scoring known to be qualified to train the computer with a text-based library, 2) removing various features from the texts in the educational data, 3) developing a model about all the qualities of the text, 4) assigning points to texts which were not evaluated by using the established model or categorizing them. There are different algorithms that can be used in the supervised machine learning process. In this research, three algorithms based on classical machine learning (logistic regression [LR], multinominal Naive Bayes [MNB], support vector machines [SVM]) and two deep learning algorithms based on artificial neural networks (long-short term memory [LSTM], bidirectional long-short term memory [BLSTM]) were used. Detailed information about these algorithms can be found in Berg and Gopinathan (2017), Gierl et al. (2014), Jang, Kang, Noh, Kim, Sung, and Seong (2014), and Lilja (2018).

Using automated scoring systems in open-ended items ensures efficient use of resources, reduce scoring time, and prevent workforce loss (Attali & Burstein, 2006; Chen, Xu & He, 2014). The use of this system will eliminate the need to have a large number of raters, and this will provide a great convenience for large-scale tests with open-ended questions. Therefore, current research is important. Also, scoring bias encountered in some situations can be prevented by automated scoring. Reliability problems caused by raters with different training can be eliminated, and the generalizability issue can be overcome (Adesiji et al., 2016). However, the usage of automated scoring systems depends on the obtained scores' being as similar as possible to human raters and their not having low reliability. Human raters are an important criterion for automated scoring systems (Cohen, Levi & Ben-Simon, 2018). Automated scoring results that have poor reliability and are incompatible with human raters may cause wrong decisions about individuals. From this point of view, current research is essential as it evaluates the use of the system by comparing between human raters and automated scoring. Changes in agreement between automated scoring and human raters are likely when automated scoring conditions change (e.g. the number of data used in training and testing the system). Accordingly, it is necessary to determine the amount of data that the scores for automated scoring will be reliable

enough. This situation increases the importance of the research. The aim of the study was to examine the reliability of the data obtained by scoring open-ended items automatically. One of the objectives was to compare different algorithms based on machine learning (support vector machines, logistic regression, multinomial Naive Bayes, long-short term memory, and bidirectional long-short term memory) in automated scoring. The other objective was to examine the change in the reliability of automated scoring by differentiating the data rate (33%, 20%, and 10%) used in testing the automated scoring system. Determining the conditions for which the results are acceptable will pave the way for automated scoring studies.

When the studies in the literature are reviewed, it is seen that automated scoring procedures are carried out in languages other than Turkish. The studies of Gierl et al. (2014), Adesiji et al. (2016), Taghipour and Tou Ng (2016) can be given as examples of studies using different algorithms in machine learning. Gierl et al. (2014) used the SVM algorithm based on supervised machine learning in automated scoring, Adesiji et al. (2016) utilized a structure consisting of three modules based on unsupervised machine learning in automated scoring, and Taghipour and Tou Ng (2016) utilized three recurrent neural network algorithm based on supervised machine learning (basic recurrent units, gated recurrent units, and LSTM units). The difference in language structures is a factor that may affect automated scoring. Therefore, automated scoring in the Turkish language should be investigated. Altaic language family, which Turkish is included in, has features such as vowel harmony, agglutination, suffix, sentence order, the modifier preceding the modified, having no difference in terms of the case, gender, and number in the adjective clauses. Names that come after numbers indicating plurality do not have plural suffixes, and gender is not specified in words. The differentiation of these features from other language families requires reviewing automated scoring studies in the Altaic language family. Jang et al. (2014) conducted research on the Korean language and Ishioka and Kameda (2006) on the Japanese language. In the two studies mentioned, algorithms in which properties are defined manually were used. The current research has originality since it was the first automated scoring attempt on the Turkish language.

METHOD

In this study, a correlational research method was adopted since the reliability of the scores of human raters and the reliability of the scores of automated scoring algorithms were compared. Creswell (2012) states that in correlational research, it is possible to see how the change in one variable affects the other variable.

The Development of the Software Used in Research

In the study, an automated scoring software developed by a team including the researcher was used. While the software was developed, the Turkish test's open-ended items with restricted responses in "Monitoring the Measurement and Evaluation Applications, Research and Development Project" applied by the Ministry of National Education (MoNE) were used. The Turkish test of "Monitoring the Measurement and Evaluation Applications, Research and Development Project" (ABIDE) is independent of the tests used in this stage. This test is for fifth-grade students and includes five open-ended items. While preparing the software, five open-ended items with restricted responses scored 0-1, and 0-1-2 were used. In this test, all student answers were graded by two raters, and when necessary, a final score was obtained by reaching the upper rater. Rubrics were used in scoring processes.

The results of two of the items used in the development of the software were presented as an example. The item with two categories (item 16) and the rubric is included in Appendix-A, the item with three categories (item 20) and the rubric is included in Appendix-B. Data of 303 students for the 16th item and 637 students for the 20th item in the Turkish test were used. Since item 20 was scored in three categories, more data were tried. An automated scoring system was created using the Python program on the Linux operating system, and trials were made. Five algorithms were used in automatic scoring: SVM, LR, MNB, LSTM, and BLSTM. Two libraries named Keras and scikit-learn were utilized in

the software. 90% of the data was used to train the system and 10% to test the system. The random sampling method was used with cross-validation. With 10-fold cross-validation, the test data and training data were changed ten times to be different from each other, and automated scoring was made as much as the number of data and the percentages of agreement were calculated over these scores. Thus, 303 scoring results were obtained in the trial conducted on 303 data, and 637 scoring results were obtained in the trial performed on 637 data. The usability of the software was investigated by examining the agreement between automated scoring and final scores of human raters. Table 1 includes the results of dichotomously scored (0-1) item 16 and polytomously scored (0-1-2) item 20.

Table 1. Percentages of Agreement Obtained While Creating the Software

	Data	Number of Categories	SVM (%)	LR (%)	MNB (%)	LSTM (%)	BLSTM (%)
Item 16	303	2	98.0	98.3	96.1	99.0	99.0
Item 20	637	3	85.5	82.4	75.1	87.3	88.7

Note: Percentages of agreement above 80% indicates an acceptable agreement. (Hartmann, 1977).

When Table 1 is examined, it is seen that the percentages of agreement obtained for item 16 are quite high. The algorithms showing the highest compliance percentage for the item 16 were LSTM and BLSTM. It was determined that the percentages of agreement obtained for item 20 were sufficient. The algorithm showing the best agreement for item 20 was BLSTM. The obtained results showed that the created system would be sufficient for scoring the structured answer items. Thus, an automated scoring process was started for ABIDE data sets within the scope of this research.

Research Data Source

The data source of the study consisted of 8th grades research of the Academic Skills Monitoring and Evaluation (ABIDE) Project implemented by MoNE in Turkey in 2016. In the tests aiming to examine students' higher-order thinking skills, multiple-choice and open-ended items with restricted responses are included together. The research was conducted on open-ended items with restricted responses in Turkish tests of A_1 and B_1 booklets. Nine items in the A_1 test and 10 items in the B_1 test are open-ended. The five open-ended items in the A_1 and B_1 tests are common. Open-ended items are scored as 0-1 and 0-1-2. The scoring process of open-ended items was made by two human raters. If there was no agreement between the scores, the answer was sent to the higher scorer. Thus, the final scores were obtained. Rubrics were used while scoring. It was stated that the Cramer's V coefficients of the open-ended items in the A_1 and B_1 booklets vary between .83-.98 and .87-.99, respectively. It is stated that the coefficients above .80 indicate that the consistency of the raters is high (MoNE, 2017a; MoNE, 2017b). Sample items and rubrics from ABIDE test are included in Appendix-C and Appendix-D.

Transfer of the Data to Computer Environment

First of all, the data described above were requested from the MoNE. Based on this request, 1000 data selected randomly among the data were shared with the researchers. In the data, there are score matrices of two different rater groups and final scores and student answers in jpeg format. Student answer sheets were entered into the computer environment manually. The reason for this is that student texts are difficult to read and due to the use of cursive handwriting, optical character recognition systems (OCR) cannot be adequately utilized. In addition, this eliminates errors caused by OCR programs. In order for the manually entered data to match the student answers, the data were checked by a study group of undergraduate students, and errors were corrected. Student responses were transferred directly and were not corrected.

Data Analysis

Before analyzing the research data, the data of 1000 students taken from the MoNE was examined. Data was entered based on the balanced distribution of the scores obtained from the open-ended items into the categories. This process was carried out to avoid the prevalence (imbalance in distribution to categories) problem of open-ended items in the data as much as possible. Nine open-ended items for the A_1 booklet and ten open-ended items for the B_1 booklet were taken into consideration, and 697 data from the A_1 booklet and 701 data from the B_1 booklet were entered. Then, students who answered half or more than half of the open-ended items in the test were selected. After this process, the missing data rate was calculated for each open-ended item. The data was cleaned so that the missing data rate remained below 5%. This process was carried out in order to prevent the coefficients of agreement from being higher than normal in automated scoring. While clearing the data, the distribution by categories was taken into account. Since there are few data in some categories, attention was paid not to exclude individuals that scored points in these categories as much as possible. The criteria mentioned above were considered and the data of 84 people from the A_1 booklet and 96 people from the B_1 booklet were cleared. Then, the scores given to the students by the human rater group 1 and the human rater group 2 were examined. A group of students was also excluded from the study because of the missing scores encountered here. A total of 6 people were excluded from the A_1 and B_1 booklets, respectively. Finally, the number of missing data in the multiple-choice items was evaluated, and the students who did not answer more than half of the total number of items in the test and more than half of the multiple-choice items were excluded from the study. Thereby, the missing data rate remained below 5%. No data was excluded from the A_1 booklet, and the data of 15 people were excluded from the B_1 booklet. Consequently, 90 people were from the A_1 booklet and 117 people from the B_1 booklet were excluded. Thus, the data preparation process was completed, and the automated scoring process was started with 607 data from the A_1 booklet and 584 data from the B_1 booklet.

Automated scoring of ABIDE open-ended data

In the automated scoring phase, the automated scoring system was trained by using some of the final scores. In this way, the automated scoring system was enabled to learn how to score from human raters, and scoring features were mapped to the system. Then, the data that were not used in the training of the system were scored automatically. There was no manual definition of any feature in the software. The data rate used in training/testing the system was a factor whose effect was examined in the research. The data rates used for the test were determined as 10%, 20%, and 33%. Therefore, the data rate used in training the system was 90%, 80%, and 67%, respectively. According to these values for the A1 booklet, 61, 121 and 200 data out of 607 data were used to test the system, and 546, 486 and 407 data out of 607 data were used to train the system, respectively. A similar calculation can be made for booklet B_1. When calculating the results, 10-fold cross-validation for 10% test data rate, 5-fold cross-validation for 20% test data rate and 3-fold cross-validation for 33% test data rate were used. In this way, the training and test data were differentiated and all 607 data for the A_1 booklet and all 584 data for the B_1 booklet were turned into test data. When comparing research results with other studies, data numbers rather than data rates should be used. The reason for indicating the result with the ratio is to increase the application of cross-validation and clarity.

For the evaluation of the automated scoring results, the consistency with the final scores of the human raters was calculated. The compatibility of the human rater group 1 and the human rater group 2 with the final scores was also examined in terms of making a comparison. Each item was examined separately.

Coefficients of agreement

While examining the agreement between raters, percentage of agreement (PA), quadratic weighted Kappa (QWK), and Gwet's AC1 (Gwet's AC1) coefficients were used. Detailed information is given below.

Percentage of Agreement: The percentage of agreement is a coefficient which can be understood and interpreted easily. Also, it can be calculated simply and quickly. Therefore, it was included in the research. In this method, the series of scores that the participants get from the first and second rater are compared, the ratio of the number of ratings that the raters fully agree on to the number of all ratings is calculated, and the result is stated as a percentage. The results obtained range from 0% to 100%. This coefficient is criticized as it does not take into account agreements that may occur by chance. Because this situation may lead to an excess of harmony. It also does not include the conflict between raters. This method can be used when all scale levels (nominal, ordinal, scale) and the number of score categories are two or more (Araujo & Born, 1985; Goodwin, 2001; Graham, Milanowski & Miller, 2012; Meyer, 1999). Although there is no certain rule, researchers have a consensus about the percentage of agreement should be above 80% (Hartmann, 1977).

Quadratic Weighted Kappa: Kappa coefficient is one of the most commonly used coefficients of agreement. The Kappa coefficient is a coefficient that takes into account the probability of agreements that may occur by chance between raters. But it does not take into account the possibility of disagreement between raters. For this reason, the Kappa coefficient has been weighted. When weighing the Kappa coefficient, weights are used according to the degree of mismatch. The two most commonly used weighting techniques are linear and quadratic. In linear weighting, weights are proportional to the standard deviation of the scores, while in quadratic weighting, weights are proportional to the square of the standard deviation of the scores (variance). Since it is easy to interpret, the use of quadratic-weighted Kappa (QWK) is quite common in practice. QWK is frequently used in automated scoring researches. Therefore, it was included in this research. This coefficient, which can be used when there are two or more score categories, can be misleadingly low if one of the scores is higher than the other or the others. This situation is defined as a prevalence problem in the literature and is the most reported problem related to the Kappa coefficient. Besides the prevalence, bias is also effective on the Kappa value. The bias problem arises when there is a difference between the frequencies of raters' evaluations about a situation (Byrt, Bishop & Carlin, 1993; Eugenio & Glass, 2004). The quadratic weighted Kappa can also be used to evaluate the agreement between automated scoring system scores and the human raters' scores agreed upon, and takes values ranging from 0 to 1. While the 0 coefficient indicates that there is no agreement between the raters, the one coefficient indicates a very good agreement between the raters. This value may drop below 0 when there is less agreement among the raters than the value that would arise by chance (Altman, 1991; Brenner & Kliebsch, 1996; Graham, Milanowski & Miller, 2012; Preston & Goodman, 2012; Sim & Wright, 2005; Vanbelle, 2016). Landis and Koch (1977) specified a criterion for the interpretation of the Kappa coefficient, and Altman (1991) adapted this criterion. Accordingly, the interpretation of values are as follows: <.20 as "poor", .21-.40 as "fair", .41-.60 as "moderate", .61-.80 as "good" and .81-1.00 as "very good" agreement. Williamson, Xi, and Breyer (2012) suggest that the agreement between human raters and automated scoring systems should be over .70. Equations used by Wang, Wei, Zhou, and Huang (2018) and Preston and Goodman (2012) were used to calculate the quadratic weighted Kappa value. Detailed information can be obtained from these sources.

Gwet's AC1 Coefficient: Gwet's AC1 coefficient (Gwet, 2008) emerged in line with the paradoxes encountered in Cohen's Kappa coefficient. The skewness (prevalence) in the distribution of the data into categories, the bias caused by the raters, the differentiation of the sensitivity and specificity of the raters reduce the capability of the Kappa value to determine the agreement between the raters (Eugenio & Glass, 2004; Gwet, 2008). The AC1 coefficient differs from the Kappa coefficient with the adjustment on the averages of marginal probability for each category and the expected ratio of chance agreement. Thus, comparing with the Kappa value, it is less affected by paradoxes, and it is more stable against the skewness between categories, that is, the variability between categories (Hoek & Scholman, 2017).

When there are imbalance and lack of symmetry in the categories, the AC1 coefficient is more efficient at detecting the agreement between raters (Shankar & Bangdiwala, 2014). Gwet's AC1 coefficient can be used in categorical data regardless of the number of raters (Wongpakaran, Wongpakaran, Wedding & Gwet, 2013). AC1 coefficient takes lower values than the percentage of agreement and higher than

the Kappa coefficient (Lacy, Watson, Riffe & Lovejoy, 2015). Gwet's AC1 coefficient can be interpreted through the criteria defined by Landis and Koch (1977) for the Kappa coefficient (Senay, Delisle, Raynauld, Morin & Fernandes, 2015; Siriwardhana, Walters, Rait, Bazo-Alvarez & Weerasinghe, 2018). Hoek and Scholman (2017) recommend researchers to use the AC1 value along with the Kappa value in their research. In addition, Haley (2007) states that the AC1 coefficient is an efficient way to evaluate the automated scoring systems. Therefore, this coefficient was included in the current study. The equation used to calculate Gwet's AC1 coefficient can be found in Gwet's research (2016).

When interpreting the coefficients of agreement, the prevalence of scores and the bias of raters are crucial. Therefore, the prevalence and bias indexes are calculated. Byrt, Bishop, and Carlin (1993) state that its essential to take into consideration the prevalence and bias indexes so that the Kappa coefficient is not misleading. Even though the prevalence index varies between -1 and 1, it can be stated that since the absolute value is used, being close to 1 of the coefficients obtained will decrease the Kappa value. On the other hand, the absolute value of the bias index varies between 0 and 1, and it can be stated that the increase in the bias coefficients will also increase the Kappa value (Byrt, Bishop & Carlin, 1993). The prevalence and bias coefficients of all structured answer items in A_1 and B_1 booklets were examined. The prevalence coefficient of item 2, item 7, item 14, and item 19 in the A_1 booklet; item 3 and item 5 in the B_1 booklet are high, and consequently, it is predicted that the QWK value in these items may be lower than the real agreement value. It is predicted that items 10 and 11 in the A_1 booklet, item 8, item 9, and item 18 in the B_1 booklet are the items with the lowest prevalence coefficient, and therefore the QWK value will be closer to the real agreement. The bias values of all of the items in the A_1 and B_1 booklets are very low, and therefore it is very unlikely of the QWK value's being higher than the real agreement value.

While calculating the percentage of agreement, QWK and AC1 coefficients; the "irr" (Gamer, Lemon, Fellows & Singh, 2010), "rel" (LoMartire, 2017) and "Metrics" (Hamner & Frasco, 2018) packages in the R program (R Core Team, 2018) were used, respectively. The performances of the algorithms were compared by averaging all items for the coefficients of agreement. In addition, the performance of the algorithms was reviewed by averaging the data rates used in testing the system.

FINDINGS

The coefficients of agreement related to the open-ended items in the A_1 booklet were first calculated between the human raters group 1 and 2 and the final scores of the human raters. Then, the consistency between five different automated scoring algorithms and the final scores was examined by changing the data rates used in testing the automated scoring system. The results are shown in Table 2 for the A_1 booklet. A sample of the interpretation of an item (item 2) in the A_1 booklet is given. The sample item is about a situation where there is a prevalence problem. The results related to other items in the A_1 booklet can be evaluated in Table 2. In Table 2, three coefficients with the highest agreement values are shown in bold, and three coefficients with the lowest agreement values are shown in italic for each type of agreement coefficient.

When the values belonging to item 2 in table 2 are examined, it is seen that the percentage of agreement between the first human raters group and the final scores was .980, the AC1 index was .976, and the QWK value was .880. The percentage of agreement between the second human raters group and the final scores was .979, the AC1 index was .975, and the QWK value was .862.

When the agreement between the automated scoring and the final scores of the human raters is examined with a 10% test data rate, it is seen that the highest percentage of agreement was obtained as .941 with the BLSTM algorithm, followed by the .921 with MNB algorithm. The lowest percentage of agreement was obtained with .913 in the LSTM algorithm. When the percentages of agreement are examined, it was concluded that the values were close to each other and at acceptable levels (>.80). When the AC1 index is examined, the algorithm with the highest agreement was the BLSTM algorithm with .931, followed by the LR algorithm with .910. The lowest AC1 value was in the SVM and LSTM algorithms with a value of .904. It was observed that AC1 values were close to each other and had a

very good agreement (>.80) for all algorithms. The highest QWK value was found as .569 with the BLSTM algorithm, followed by the MNB algorithm with .448. The lowest QWK value was in the LSTM algorithm with .061, and this value was followed by the LR algorithm with .223. It was concluded that the QWK values varied considerably among the algorithms, the range was .508, and it differed from the AC1 index and the percentage of agreement. When the QWK value is evaluated as a whole, it can be stated that the BLSTM and MNB algorithms were moderate (<.60 Λ >.40), the LR and SVM algorithms (<.40 Λ >.20) were fair, and the LSTM algorithm was poor (<.20).

With 20% test data rate, the BLSTM algorithm showed the highest percentage of agreement with .942, while the MNB algorithm showed the lowest percentage of agreement with .913. It is seen that the percentages of agreement in all algorithms were very close to each other and at an acceptable level (>.80). When the agreement was evaluated in terms of the AC1 index, the highest agreement was found in the BLSTM algorithm with .933, and the lowest with .899 in the MNB algorithm. It can be stated that the AC1 index values were generally close, and all of them showed very good agreement (>.80). When the QWK values are examined, it can be stated that the algorithm with the highest agreement was the BLSTM algorithm with .593 and the algorithm with the lowest agreement was the LSTM algorithm with .147. The second algorithm with the lowest agreement was SVM with .212. As it can be seen, at a 20% test data rate, similar to the 10% test data rate, QWK values were low, and there were differences between algorithms. The range of QWK values at a 20% test data rate was .446. When the QWK values were examined in general, it is seen that the BLSTM algorithm showed moderate agreement (<.60 Λ >.40), the MNB, LR, and SVM algorithms showed a fair agreement (<.40 Λ >.20), and the LSTM algorithm indicated a poor agreement (<.20).

For the 33% test data rate, the highest percentage of agreement is the BLSTM algorithms with .934. The algorithm with the lowest percentage of agreement is the SVM with .909. Generally, the percentages of agreement were high, close to each other, and acceptable (>.80). In addition to the fact that AC1 indexes are generally high, the highest agreement is in the BLSTM algorithm with .924, and the lowest agreement is in the SVM algorithm with .899. The values obtained for all algorithms are close to each other and show very good agreement (>.80). When the QWK values were evaluated, the highest agreement was obtained in the BLSTM algorithm with .522, and the lowest two agreements were obtained in the SVM algorithm with .128 and in the LSTM algorithm with .000. At 33% test data rate, the QWK values were low, varied widely between algorithms, and its range was .522. When the values obtained were examined, it was seen that the BLSTM algorithm had moderate agreement (<.60 Λ >.40), MNB and LR algorithms had fair agreements (<.40 Λ >.20), and LSTM and SVM algorithms had poor agreements (<.20).

Table 2. Coefficients of Agreement between Human Rater Groups, Automated Scoring Algorithms and Final Scores for Open-Ended Items in A_1 Booklet

Item Code	Agreement Between Human Rater Group and Final Scores				Test data selection method	Agreement Between Automated Scoring Algorithms and Final Scores (Agreed by Human Raters)													
						SVM			LR			MNB			LSTM			BLSTM		
		PA	AC1	QWK		PA	AC1	QWK	PA	AC1	QWK	PA	AC1	QWK	PA	AC1	QWK	PA	AC1	QWK
Item 2	P_1-P_F	.980	.976	.880	CV %10	.914	*.904*	.226	.919	.910	.223	.921	.908	.448	*.913*	*.904*	*.061*	**.941**	**.931**	**.569**
	P_2-P_F	.979	.975	.862	CV %20	.916	.906	.212	.923	.914	.273	*.913*	*.899*	.347	.916	.907	.147	**.942**	**.933**	**.593**
					CV %33	*.909*	*.899*	*.128*	.921	.912	.208	.918	.906	.337	*.911*	*.903*	*.000*	**.934**	**.924**	**.522**
Item 7*	P_1-P_F	.979	.970	.974	CV %10	.845	.782	.862	.822	.752	.836	.735	.642	*.720*	*.720*	*.629*	*.683*	**.881**	**.833**	**.884**
	P_2-P_F	.970	.958	.971	CV %20	**.855**	**.796**	.859	.815	.743	.832	.731	.639	*.720*	.735	.647	.744	**.881**	**.833**	**.892**
					CV %33	.827	.756	.825	.822	.752	.832	*.722*	*.625*	*.705*	*.728*	*.638*	.726	**.875**	**.823**	**.877**
Item 8*	P_1-P_F	.997	.995	.997	CV %10	.928	.894	.910	.936	.906	**.915**	.896	.849	.859	*.779*	*.687*	*.701*	**.957**	**.937**	**.937**
	P_2-P_F	.987	.981	.985	CV %20	.936	.906	**.917**	.931	.899	.911	.901	.856	.868	*.776*	*.683*	*.684*	**.946**	**.921**	.899
					CV %33	.931	.899	.909	.931	.899	.896	.875	.819	.839	*.771*	*.676*	*.672*	**.942**	**.916**	.912
Item 10*	P_1-P_F	.944	.891	.885	CV %10	.837	.682	.665	**.845**	**.699**	**.681**	.827	.667	*.641*	.840	.688	.672	**.863**	**.733**	**.720**
	P_2-P_F	.947	.897	.892	CV %20	.840	.689	.672	**.842**	.693	**.675**	.835	.681	.660	.829	.662	.652	**.842**	**.695**	.673
					CV %33	*.817*	*.642*	*.626*	*.819*	*.649*	*.626*	.830	.673	.648	*.824*	*.657*	*.637*	.835	.680	.660
Item 11*	P_1-P_F	.985	.972	.968	CV %10	.870	.755	.723	.875	.769	.726	*.843*	*.720*	*.648*	.924	.860	.835	**.956**	**.917**	**.904**
	P_2-P_F	.985	.972	.968	CV %20	.873	.761	.730	.881	.779	.744	*.835*	*.708*	*.626*	.934	.879	.855	**.962**	**.929**	**.918**
					CV %33	.871	.757	.727	.865	.748	.708	*.825*	*.693*	*.600*	.870	.759	.717	**.946**	**.898**	**.883**

* Common items in A_1 and B_1 booklets.

Note 1: P_1: First rater group, P_2: Second rater group, P_F: Final scores

Note 2: PA: Percentage of Agreement, AC1: Gwet's AC1 Coefficient, QWK: Quadratic Weighted Kappa

Note 3: CV: Cross validation, 10%, 20% and 33% shows test data rate.

Table 2 (continued). Coefficients of Agreement between Human Rater Groups, Automated Scoring Algorithms and Final Scores for Open-Ended Items in A_1 Booklet

Item Code	Agreement Between Human Rater Group and Final Scores				Test data selection method	Agreement Between Automated Scoring Algorithms and Final Scores (Agreed by Human Raters)													
						SVM			LR			MNB			LSTM			BLSTM		
		PA	AC1	QWK		PA	AC1	QWK	PA	AC1	QWK	PA	AC1	QWK	PA	AC1	QWK	PA	AC1	QWK
Item 14	P_1-P_F	.975	.959	.937	CV %10	.901	.839	.744	.911	.857	.764	.890	.828	.695	*.792*	*.709*	*.318*	**.929**	**.884**	**.818**
	P_2-P_F	.969	.948	.921	CV %20	.895	.829	.724	.904	.847	.747	.881	.817	.667	*.873*	*.807*	*.635*	**.928**	**.880**	**.816**
					CV %33	.893	.825	.725	.906	.849	.752	.876	.811	.646	*.792*	*.710*	*.315*	**.916**	**.864**	**.781**
Item 15*	P_1-P_F	.972	.960	.971	CV %10	.708	.585	.683	.720	**.603**	.686	.687	.563	.613	*.560*	*.428*	*.224*	**.766**	**.666**	**.714**
	P_2-P_F	.960	.943	.943	CV %20	.717	.595	.678	.712	.593	.664	.672	.544	.589	*.539*	*.415*	*.137*	**.740**	**.628**	**.707**
					CV %33	.677	.539	.656	.690	.562	.625	.680	.557	.564	*.516*	*.397*	*.000*	**.741**	**.628**	**.711**
Item 18	P_1-P_F	.997	.995	.997	CV %10	.956	.937	**.952**	.924	.893	.914	.867	.811	.790	*.718*	*.616*	*.517*	**.970**	**.958**	**.961**
	P_2-P_F	.998	.998	.994	CV %20	.941	.916	.937	.921	.888	.904	.868	.813	.796	*.761*	*.672*	*.599*	**.965**	**.951**	**.952**
					CV %33	.924	.893	.912	.923	.891	.906	.863	.807	.756	*.671*	*.544*	*.515*	**.960**	**.944**	**.947**
Item 19	P_1-P_F	.997	.996	.997	CV %10	.919	.892	.900	**.936**	**.915**	**.918**	.815	.752	.807	*.802*	*.739*	*.749*	**.939**	**.918**	**.922**
	P_2-P_F	.995	.993	.996	CV %20	.914	.886	.897	.931	.908	.909	.822	.762	.820	*.797*	*.736*	*.720*	**.937**	**.916**	**.936**
					CV %33	.918	.890	.904	.921	.895	.899	.820	.760	.800	*.778*	*.719*	*.624*	.919	.891	**.918**

* Common items in A_1 and B_1 booklets.
Note 1: P_1: First rater group, P_2: Second rater group, P_F: Final scores
Note 2: PA: Percentage of Agreement, AC1: Gwet's AC1 Coefficient, QWK: Quadratic Weighted Kappa
Note 3: CV: Cross validation, 10%, 20% and 33% shows test data rate.

Figure 1 shows the agreement values obtained for item 2 in A_1 booklet according to automated scoring algorithms and test data rates.

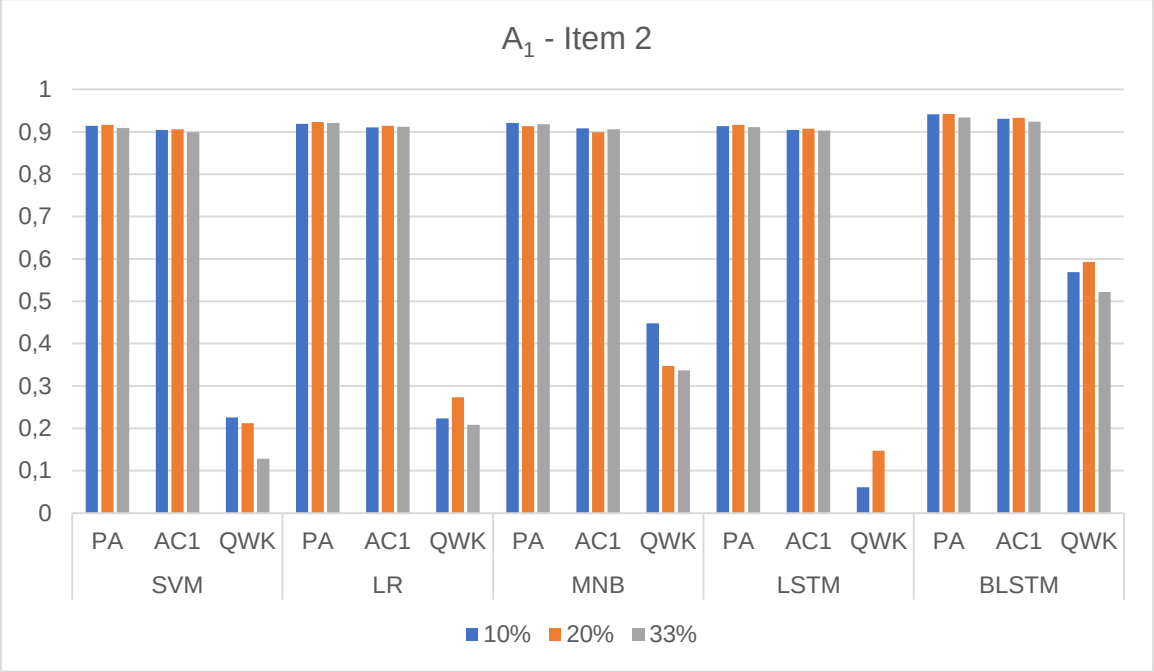

Figure 1. Graph showing Agreement Values for Item 2 in A_1 Booklet according to Automated Scoring Algorithms and Test Data Rates

When figure 1 is examined, for item 2, in all the test data rates and automated scoring algorithms, the QWK coefficient was considerably lower than the AC1 values and percentage of agreement. The reason for the low values encountered in all of the QWK coefficients and the coefficient's being close to .000 under some circumstances was the prevalence problem. Therefore, QWK was not taken into consideration. This was one of the situations predicted in the research. When a comparison was made by considering all test data rates and automated scoring algorithms, it was observed that the agreement values were slightly higher at 20% test data rate and slightly lower at 33% test data rate. However, the differences between them were very small. The agreement percentages were above .80, which is the acceptable limit in all conditions. The AC1 index indicated a very good agreement in all conditions (>.80). AC1 values were evaluated in the same direction as the Kappa coefficient. Accordingly, all AC1 coefficients were higher than the expected agreement value (>.70, Williamson et al., 2012) between automated scoring and human raters. When all the conditions for item 2 in table 2 were considered, the highest percentage of agreement (.942) and the highest AC1 value (.933) were obtained in the BLSTM algorithm with a 20% test data rate. These values were close to the percentage of agreement and AC1 value between the human rater groups and the final scores. Due to the prevalence problem encountered in item 2, the QWK values calculated between the human raters and the final scores were also low. This situation has reflected on machine learning more negatively.

The coefficients of agreement for open-ended items in the B_1 booklet were calculated in the same way as in the A_1 booklet. The results are shown in Table 3. The interpretation of an item (item 5) in the B_1 booklet is given as an example. Results related to the other items in the B_1 booklet can be evaluated in table 3. In table 3, three coefficients with the highest agreement values are shown in bold, and the three coefficients with the lowest agreement values are shown in italics according to each type of coefficient of agreement.

When the values in item 5 in table 3 are examined, it is seen that the percentage of agreement between the first human rater group and the final scores was .971, the AC1 index was .960, and the QWK value was .972. The percentage of agreement between the second human rater group and the final scores was .979, the AC1 index was .972, and the QWK value was .979.

When the agreement between automated scoring and final scores was examined at a 10% test data rate, the highest agreement percentage was obtained as .918 with the BLSTM algorithm. This percentage of agreement was followed by the SVM algorithm with .866. The lowest agreement percentage was obtained with .779 in the MNB algorithm. When the percentages of agreement were examined in general, it is seen that acceptable values (>.80) were reached for SVM, LR, LSTM, and BLSTM algorithms. When the AC1 index was examined, the algorithm with the highest agreement was the BLSTM algorithm with .888. The lowest AC1 value was in the MNB algorithm with .710, followed by LR and LSTM algorithms with .778. AC1 values were found to indicate very good agreement (>.80) for BLSTM and SVM algorithms, and good agreement (>.60 Λ <.80) for LR, LSTM, and MNB algorithms. The highest QWK value was found to be .925 with the BLSTM algorithm, followed by the SVM algorithm with .884. The lowest QWK value was in the MNB algorithm with .740. It was seen that the QWK values were greater than the AC1 indexes. The QWK value demonstrated very good agreement (>.80) for SVM, LR, LSTM, and BLSTM algorithms and good agreement (>.60 Λ <.80) for MNB algorithm.

At a 20% test data rate, the BLSTM algorithm showed the highest percentage of agreement with .902, and the MNB algorithm showed the lowest percentage of agreement with .781. According to the percentage of agreement, the BLSTM, LR, LSTM, and SVM algorithms showed acceptable agreement (>.80), while the MNB algorithm did not. In terms of the AC1 index, the highest agreement was obtained in the BLSTM algorithm with .866, and the lowest one was obtained in the MNB algorithm with .712. It can be stated that AC1 index values indicated very good agreement (>.80) for BLSTM and SVM algorithms, and good agreement (<.80 Λ >.60) for LR, LSTM, and MNB algorithms. When the QWK values are examined, it can be stated that the algorithm with the highest agreement was the BLSTM algorithm with .913 and the algorithm with the lowest agreement was the MNB with .743. The second algorithm with the lowest QWK value was LSTM with .846. As it is seen, in terms of QWK, good agreement (<.80 Λ >.60) for MNB and very good agreement for BLSTM, LR, LSTM, and SVM algorithms (>.80) were achieved. It is seen that the QWK values were greater than the AC1 indexes at a 20% test data rate.

For the 33% test data rate, the highest agreement percentage was the BLSTM algorithm with .892. The algorithm with the lowest percentage of agreement was the LSTM with .784. The percentage of agreement was acceptable (>.80) in all algorithms except in LSTM and MNB algorithms. According to the AC1 indexes, the highest agreement was in the BLSTM algorithm with .853. The lowest agreement was in the LSTM algorithm with .718 and this algorithm was followed by the MNB algorithm with .720. In terms of AC1 indexes, it is seen that very good agreement (>.80) was achieved for BLSTM and SVM algorithms, and good agreement (<.80 Λ >.60) for LR, LSTM, and MNB algorithms. According to the QWK coefficient, the highest agreement was obtained in the BLSTM algorithm with .904 and the lowest two agreements were obtained in the MNB algorithm with .744 and in LSTM algorithm with .783. QWK values indicated very good agreement (>.80) for BLSTM, LR, and SVM algorithms, good agreement (<.80 Λ >.60) for LSTM and MNB algorithms. It is seen that the QWK values were also greater than the AC1 indexes at 33% test data rate.

Table 3. Coefficients of Agreement between Human Rater Groups, Automated Scoring Algorithms and Final Scores for Open-Ended Items in B_1 Booklet

Item Code	Agreement Between Human Rater Group and Final Scores				Test data selection method	Agreement Between Automated Scoring Algorithms and Final Scores (Agreed by Human Raters)													
						SVM			LR			MNB			LSTM			BLSTM		
		PA	AC1	QWK		PA	AC1	QWK	PA	AC1	QWK	PA	AC1	QWK	PA	AC1	QWK	PA	AC1	QWK
Item 3	P_1-P_F	.966	.952	.877	CV %10	.911	.879	.665	.913	.882	.667	*.906*	*.871*	.653	.913	.880	.678	**.923**	**.894**	**.719**
	P_2-P_F	.973	.962	.900	CV %20	.914	.883	.683	.911	.879	.665	*.904*	*.869*	*.642*	**.921**	**.891**	**.716**	.913	.879	.686
					CV %30	**.916**	**.885**	**.688**	*.906*	.872	*.644*	*.901*	*.865*	*.623*	.911	.878	.671	.911	.878	.671
Item 5*	P_1-P_F	.971	.960	.972	CV %10	.866	.818	.884	.836	.778	.864	*.779*	*.710*	*.740*	.836	.778	.861	**.918**	**.888**	**.925**
	P_2-P_F	.979	.972	.979	CV %20	.863	.814	.882	.837	.781	.855	*.781*	*.712*	*.743*	.825	.766	.846	**.902**	**.866**	**.913**
					CV %30	.870	.823	.878	.844	.790	.866	.786	.720	*.744*	*.784*	*.718*	.783	**.892**	**.853**	**.904**
Item 6*	P_1-P_F	.991	.988	.981	CV %10	.942	.915	.909	**.954**	**.933**	**.924**	.884	.833	.861	*.740*	*.628*	*.654*	**.959**	**.940**	**.939**
	P_2-P_F	.993	.990	.995	CV %20	.945	.920	.919	.947	.923	.915	.873	.819	.848	*.752*	*.649*	*.645*	.949	.925	.923
					CV %30	.937	.908	.916	.947	.923	.906	.846	.781	.832	*.719*	*.593*	*.682*	**.952**	**.930**	**.926**
Item 8*	P_1-P_F	.950	.902	.899	CV %10	.827	.659	.649	.818	.645	.629	.820	.649	.632	.834	.673	.663	**.854**	**.713**	**.704**
	P_2-P_F	.957	.916	.913	CV %20	.812	.629	.618	*.800*	*.608*	*.591*	.832	.673	.656	*.805*	.618	.601	**.858**	**.719**	**.713**
					CV %30	.820	.646	.634	*.793*	*.593*	*.578*	.827	.662	.646	*.793*	*.590*	*.582*	**.842**	**.691**	**.679**
Item 9*	P_1-P_F	.985	.971	.967	CV %10	.846	.711	.670	*.836*	.696	.642	*.796*	*.637*	*.538*	**.877**	**.772**	**.732**	**.885**	**.788**	**.751**
	P_2-P_F	.993	.987	.985	CV %20	.844	.706	.668	.844	.714	.658	*.796*	*.641*	*.533*	.873	.767	.722	**.882**	**.779**	**.746**
					CV %30	.849	.716	.679	*.837*	.698	.647	*.796*	*.643*	*.531*	.868	.760	.707	.872	.766	.716

* Common items in A_1 and B_1 booklets.

Note 1: P_1: First rater group scores, P_2: Second rater group scores, P_F: Final scores

Note 2: PA: Percentage of Agreement, AC1: Gwet's AC1 Coefficient, QWK: Quadratic Weighted Kappa

Note 3: CV: Cross validation, 10%, 20% and 33% shows test data rate.

Note 4: Item 5, item 6, item 8 and item 9 in this table correspond to item 7, item 8, item 10 and item 11 in the A_1 booklet, respectively.

Table 3 (continued). Coefficients of Agreement between Human Rater Groups, Automated Scoring Algorithms and Final Scores for Open-Ended Items in B_1 Booklet

Item Code	Agreement Between Human Rater Group and Final Scores				Test data selection method	Agreement Between Automated Scoring Algorithms and Final Scores (Agreed by Human Raters)													
						SVM			LR			MNB			LSTM			BLSTM		
		PA	AC1	QWK		PA	AC1	QWK	PA	AC1	QWK	PA	AC1	QWK	PA	AC1	QWK	PA	AC1	QWK
Item 11	P_1-P_F	.986	.981	.987	CV %10	.918	.886	.912	.911	.876	.902	*.861*	*.807*	*.867*	.882	.838	.887	**.940**	**.916**	**.925**
	P_2-P_F	.990	.986	.989	CV %20	.902	.865	.893	.913	.878	.900	*.863*	*.810*	*.863*	*.878*	.833	.880	**.943**	**.920**	**.929**
					CV %30	.904	.867	.901	.914	.881	.899	*.861*	*.808*	*.860*	.885	.843	.894	**.930**	**.901**	**.927**
Item 12	P_1-P_F	.949	.923	.932	CV %10	.736	.606	.667	.757	.637	.719	.707	.566	*.606*	*.654*	*.490*	.663	**.793**	**.690**	**.749**
	P_2-P_F	.938	.908	.937	CV %20	.759	.640	.718	.764	.647	**.740**	.682	.528	*.559*	*.649*	*.481*	.674	**.784**	**.677**	**.741**
					CV %30	.755	.634	.718	.755	.635	.719	.683	.531	*.573*	*.634*	*.467*	.654	**.774**	**.662**	.738
Item 17*	P_1-P_F	.974	.963	.966	CV %10	.707	.580	.653	.693	.565	.631	.635	.492	.522	*.541*	*.393*	*.171*	**.743**	**.634**	**.705**
	P_2-P_F	.978	.968	.974	CV %20	**.729**	**.612**	**.675**	.678	.543	.609	.610	.456	.488	*.545*	*.391*	*.302*	**.716**	**.595**	**.671**
					CV %30	.680	.543	.617	.700	.575	.637	.616	.471	.478	*.575*	*.430*	*.339*	.697	.567	.644
Item 18	P_1-P_F	1.000	1.000	1.000	CV %10	*.712*	*.425*	*.429*	.748	.497	.497	.740	.480	.485	**.784**	**.568**	**.571**	**.786**	**.572**	**.572**
	P_2-P_F	.995	.990	.990	CV %20	*.711*	*.421*	*.425*	.741	.483	.483	.726	.453	.458	.759	.517	.520	.767	.535	.534
					CV %30	*.719*	*.439*	*.442*	.731	.463	.462	.731	.463	.466	.755	.510	.512	**.769**	**.538**	**.538**
Item 20	P_1-P_F	.969	.945	.929	CV %10	.818	.687	.569	.817	.685	.563	*.760*	*.562*	*.471*	**.834**	.703	**.623**	**.839**	**.717**	**.627**
	P_2-P_F	.969	.946	.929	CV %20	.815	.681	.562	.830	**.708**	.597	*.750*	*.544*	*.447*	.820	.683	.585	**.837**	**.710**	**.629**
					CV %30	.789	.640	.495	.810	.674	.545	*.740*	*.527*	*.421*	.793	.630	.529	.820	.691	.572

* Common items in A_1 and B_1 booklets.

Note 1: P_1: First rater group scores, P_2: Second rater group scores, P_F: Final scores

Note 2: PA: Percentage of Agreement, AC1: Gwet's AC1 Coefficient, QWK: Quadratic Weighted Kappa

Note 3: CV: Cross validation, 10%, 20% and 33% shows test data rate.

Note 4: Item 17 in this table correspond to item 15 in the A_1 booklet.

Figure 2 shows the agreement values obtained for item 5 in B_1 booklet according to automated scoring algorithms and test data rates.

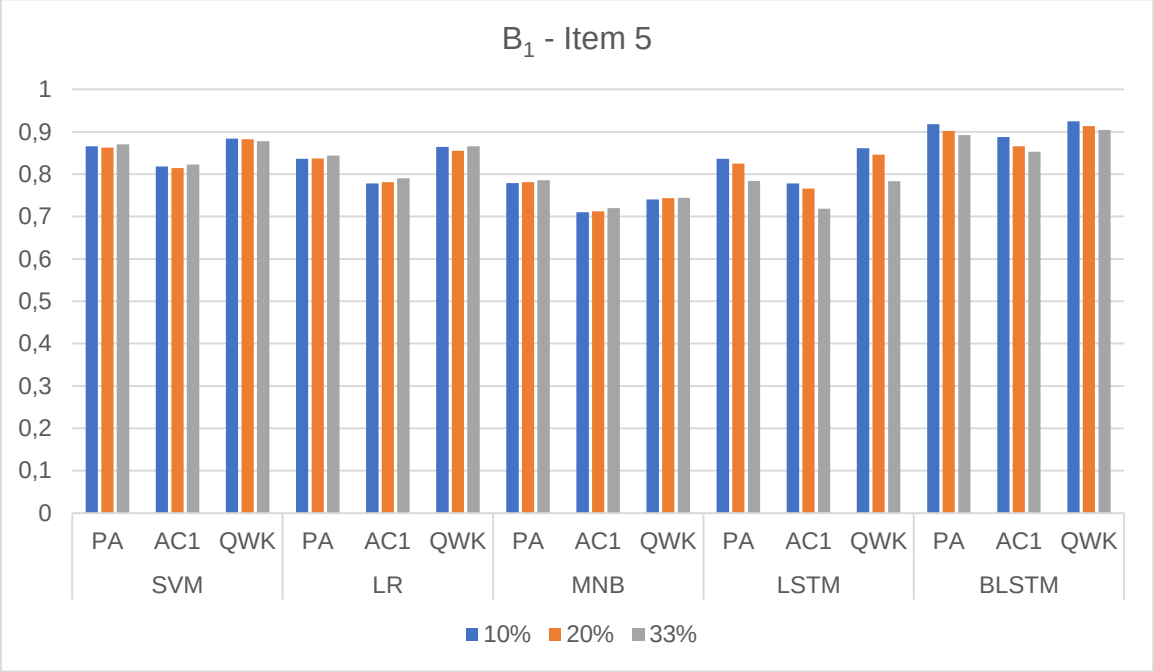

Figure 2. Graph showing Agreement Values for Item 5 in B_1 Booklet according to Automated Scoring Algorithms and Test Data Rates

When Figure 2 is examined, in all conditions, the coefficients of agreement of the MNB algorithm are lower than the coefficients of agreement of the other algorithms, while the coefficients of agreement of the BLSTM algorithm are higher than the coefficients of agreement of the other algorithms. QWK value indicated very good agreement in all test data rates for BLSTM, LR, and SVM algorithms and at 10% and 20% test data rates for LSTM algorithm (>.80). It also showed good agreement in all test data rates for the MNB algorithm and at 33% test data rate for the LSTM algorithm (<.80 Λ >.60). In all conditions, AC1 values showed very good agreement (>.80) for BLSTM and SVM algorithms and good agreement (<.80 Λ >.60) for LR, MNB, and LSTM algorithms. All AC1 coefficients for item 5 were lower than QWK coefficients. Percentage of agreement showed acceptable values in all test data rates for the BLSTM, LR, and SVM algorithms and at 10% and 20% test data rates for the LSTM algorithm. The QWK values were acceptable in all algorithms and test data rates according to Williamson, Xi, and Breyer's (2012) criteria that the Kappa coefficient of agreement between human raters and automated scoring should be at least .70. When the same criteria were used for the AC1 coefficient, acceptable values were achieved in all algorithms and test data rates. For item 5, the highest percentage of agreement (.918), AC1 value (.888) and QWK coefficient (.925) were obtained in BLSTM algorithm at 10% test data rate. These values are close to the values of AC1, QWK, and the percentage of agreement between the human rater groups and the final scores.

In order to make a general comparison between the automated scoring algorithms, the performance of the algorithms in each item was averaged. Table 4 shows the performances of the automated scoring algorithms in different test data rates and the averages of these performances. In Table 4, the coefficients showing the highest agreement in each test data rate and average performance in all coefficients of agreement are shown in bold, and the coefficients showing the lowest agreement are shown in italic.

Table 4. Average Performance of Automated Scoring Algorithms

Coefficients of Agreement	Automated Scoring Algorithm	%10	%20	%33	Mean
	SVM	.855	.855	.848	.853
	LR	.857	.856	.851	.855
PA	MNB	.816	.810	.807	.811
	LSTM	*.794*	*.799*	*.775*	*.789*
	BLSTM	**.889**	**.883**	**.874**	**.882**
	SVM	.768	.767	.756	.764
	LR	.773	.771	.762	.769
AC1	MNB	.712	.704	.700	.705
	LSTM	*.694*	*.698*	*.665*	*.686*
	BLSTM	**.822**	**.810**	**.798**	**.810**
	SVM	.705	.704	.689	.699
	LR	.710	.710	.692	.704
QWK	MNB	.658	.640	.627	.642
	LSTM	*.583*	*.612*	*.545*	*.580*
	BLSTM	**.782**	**.775**	**.755**	**.771**

When the percentages of agreement for each test data rate are examined in Table 4, it is seen that the values were close to each other, but there was a slight decrease in the values at the 33% test data rate. All algorithms, except the LSTM algorithm, showed acceptable values in terms of percentage of agreement. But the LSTM algorithm showed close values to the acceptable agreement.

When AC1 values are examined, it is seen that there was a slight decrease at 33% test data rate, and the average performances of SVM, LR, MNB, and LSTM algorithms indicated good agreement. The BLSTM algorithm showed very good agreement at 10% and 20% test data rates and good agreement at 33% test data rate.

When the QWK values are examined, it is seen that there was a decrease in the test data rate of 33% similar to the AC1 and the percentage of agreement, besides, close values were obtained in all test data rates. In terms of QWK value, SVM, LR, MNB, and BLSTM algorithms indicated good agreement. On the other hand, the LSTM algorithm showed good agreement at 20% test data rate, and moderate agreement at 10% and 33% test data rates.

When the averages of all test data rates are examined in terms of each automated scoring algorithm and coefficient of agreement, it is seen that the algorithm with the highest percentage of agreement and highest AC1 and QWK values is BLSTM. Along with the BLSTM algorithm had an acceptable percentage of agreement, it showed very good agreement according to the AC1 coefficient and good agreement according to the QWK coefficient. SVM, LR, and MNB algorithms indicated good agreement according to the acceptable percentage of agreement, the AC1 coefficient, and the QWK coefficient. The LSTM algorithm did not have an acceptable percentage of agreement, but it indicated good agreement in terms of the AC1 index and moderate agreement in terms of the QWK coefficient. As a result of both the evaluation of the item averages and the evaluations made within the scope of the item, the best three automated scoring conditions were determined as the BLSTM algorithm at 10% test data rate, the BLSTM algorithm at 20% test data rate and the BLSTM algorithm at 33% test data rate. Figure 3 shows the average of the algorithms taken according to the test data rates.

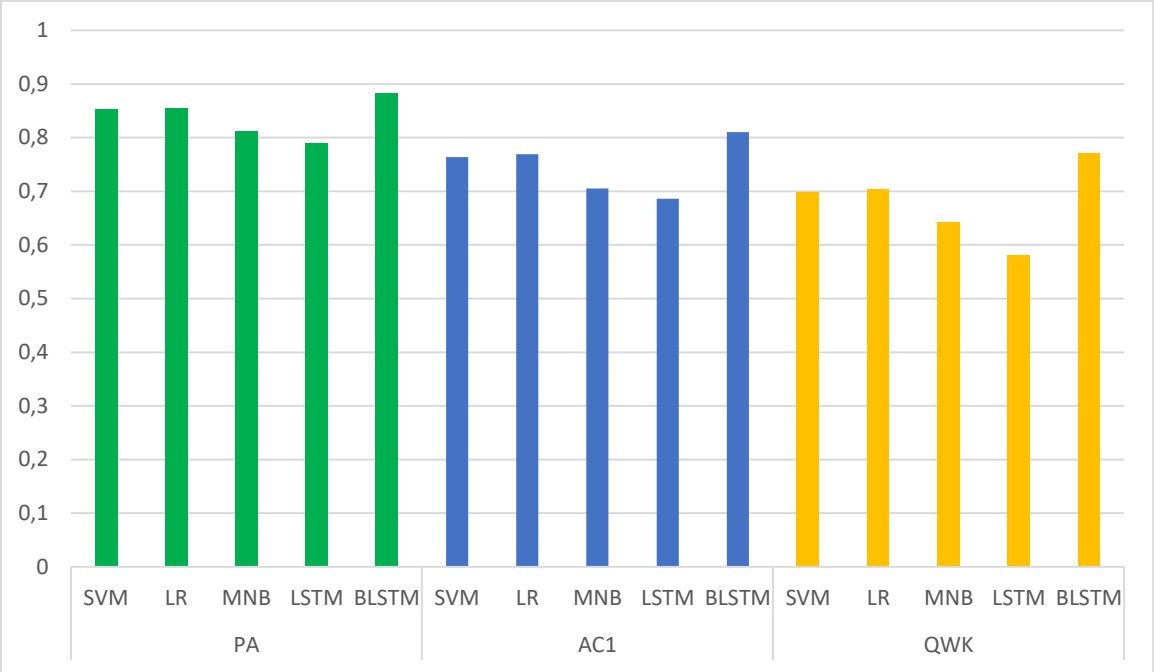

Figure 3. Chart Showing Average Performance of Automated Scoring Algorithms

When Figure 3 is examined, it was determined that MNB and LSTM algorithms performed slightly less than other algorithms. The lowest performance was observed in the LSTM algorithm and the highest performance was observed in the BLSTM algorithm.

RESULTS AND DISCUSSION

The research compared automated scoring algorithms with changes made on data rates used in testing the system. For this purpose, SVM, LR, MNB, LSTM, and BLSTM algorithms were compared with each other according to 10%, 20%, and 33% test data rates. When comparing the algorithms, the consistency of human raters with the final scores was taken into account. Thus, the difference between human raters and automated scoring was determined. Considering the ABIDE data, the results showed that the best automated scoring was achieved with the BLSTM algorithm. LSTM and MNB algorithms had lower agreement values than SVM, LR, and BLSTM algorithms. In their previous experiments on various classification algorithms, Kumar and Rama Sree (2014) determined that Naive Bayes algorithm had lower percentages of agreement than LR and SVM algorithms. This result supports the research findings. Gierl et al. (2014) stated that the QWK value was very good in the automated scoring process performed with the SVM algorithm. In the current study, it was determined that the SVM algorithm indicated good agreement. Taghipour and Tou Ng (2016) found that the algorithm with the highest QWK value (.746) was LSTM in their study in which they compared the recurrent neural networks in the automated scoring process. In the same study, the closest QWK value was obtained in the BLSTM algorithm (.699). Similarly, in the current study, the QWK value of the BLSTM algorithm indicated good agreement. However, in the current study, it was determined that the LSTM algorithm showed a medium level of agreement according to the QWK value. The reason for this situation may be that the one-way analysis of sentences in LSTM algorithm and two-way analysis of sentences in BLSTM algorithm may differ in the Turkish language. Even though the comparisons made according to the test data rates showed that the coefficients of agreement slightly decreased at 33% test data rate, SVM, LR, MNB, and BLSTM algorithms indicated good or very good agreement in all conditions.

When the comparison was made according to the lowest acceptable agreement for automated scoring, it was determined that the LR and BLSTM algorithms were at the desired level, and the SVM algorithm was very close to the desired level. When the percentage of agreement of the system created with this

current research was taken into account, it can be stated that this system performed better than the unsupervised machine learning-based method prepared by Adesiji et al. (2016). Thus, it was concluded that open-ended items in the Turkish language could be scored automatically by selecting the appropriate automated scoring algorithm based on supervised machine learning in the Turkish language. Although automated scoring systems developed in languages that have similar features to the Turkish language are not based on supervised machine learning, they can be used similarly. Ishioka and Kameda (2006) and Jang et al. (2014) determined that there was a high level of correlation between the automated scoring system and human scores in the Japanese language and the Korean language, respectively.

The automated scoring system created in the Turkish language can be used in large-scale tests. It was also stated that the automated scoring system created in Korean, which is a similar language to Turkish, can be used in large-scale tests (Jang et al., 2014). Based on the findings obtained as a result of the research, the recommendations for researchers and practitioners are as follows:

1. Automated scoring, which is tried for the first time in the Turkish language and seems to be usable, can be used in large-scale tests by developing the system and pilot scheme, and exam costs can be reduced, and the results can be explained more quickly.

2. Among the automated scoring algorithms, BLSTM and LR algorithms can be preferred for data having similar characteristics to the data used in this study.

3. In automated scoring, it can be suggested that MNB and LSTM algorithms should not be used in data having characteristics similar to the data used in this study.

4. This research reflects automated scoring results with at least 400 training data. In future studies, the effect of this situation on the coefficients of agreement can be evaluated by making automated scoring with less training data. Moreover, after the automated scoring process with a large number of training data in large samples (>1000 or >3000), the effect of this situation on automated scoring can be examined by gradually reducing the training data.

5. Automated scoring results obtained in cases where the spelling errors in the data are corrected or not corrected in subsequent studies can be compared.

6. In subsequent studies conducted on paper-pencil tests, the results obtained by data entry via OCR systems and manual data entry can be compared.

7. Within the scope of the research, items with two and three categories were studied. In case of an increase in the number of categories in later studies, the results of automated scoring systems can be examined.

REFERENCES

Adesiji, K. M., Agbonifo, O. C., Adesuyi, A. T., & Olabode, O. (2016). Development of an automated descriptive text-based scoring system. *British Journal of Mathematics & Computer Science, 19*(4), 1-14. doi: 10.9734/BJMCS/2016/27558

Altman, D. G. (1991). *Practical statistics for medical research*. Boca Raton: CRC.

Araujo, J., & Born, D. G. (1985). Calculating percentage agreement correctly but writing its formula incorrectly. *The Behavior Analyst, 8*(2), 207-208. doi: 10.1007/BF03393152

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v.2. *Journal of Technology, Learning, and Assessment, 4*(3). Retrieved from http://www.jtla.org.

Berg, P-C., & Gopinathan, M. (2017). *A deep learning ensemble approach to gender identification of tweet authors* (Master's thesis, Norwegian University of Science and Technology). Retrieved from https://brage.bibsys.no/xmlui/handle/11250/2458477

Brenner, H., & Kliebsch, U. (1996). Dependence of weighted Kappa coefficients on the number of categories. *Epidemiology, 7*(2), 199-202. https://doi.org/10.1097/00001648-199603000-00016

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and Kappa. *Journal of Clinical Epidemiology, 46*(5), 423-429. doi: 10.1016/0895-4356(93)90018-V

Chen, H., Xu, J., & He, B. (2014). Automated essay scoring by capturing relative writing quality. *The Computer Journal, 57*(9), 1318-1330. doi:10.1093/comjnl/bxt117

Cohen, Y., Ben-Simon, A., & Hovav, M. (October, 2003). *The effect of specific language features on the complexity of systems for automated essay scoring*. Paper presented at the International Association of Educational Administration, Manchester.

Cohen, Y., Levi, E., & Ben-Simon, A. (2018). Validating human and automated scoring of essays against "True" scores. *Applied Measurement in Education, 31*(3), 241-250. https://doi.org/10.1080/08957347.2018.1464450

Creswell, J. W. (2012). *Educational research: Planning, conducting and evaluating quantitative and qualitative research* (4th ed.). Boston: Pearson.

Downing, S. M. (2009). Written tests: Constructed-response and selected-response formats. In S. M. Downing & R. Yudkowsky (Eds.), *Assessment in health professions education* (pp. 149-184). New York, NY: Routledge.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Eugenio, B. D., & Glass, M. (2004). The Kappa statistic: A second look. *Computational Linguistics, 30*(1), 95-101. https://doi.org/10.1162/089120104773633402

Gamer, M., Lemon, I., Fellows, J., & Singh, P. (2010). *irr: Various coefficients of interrater reliability and agreement* (Version 0.83) [Computer software]. https://CRAN.R-project.org/package=irr

Geisinger, K. F., & Usher-Tate, B. J. (2016). A brief history of educational testing and psychometrics. In C. S. Wells, M. Faulkner-Bond (Eds.), *Educational measurement from foundations to future* (pp. 3-20). New York: The Guilford.

Gierl, M. J., Latifi, S., Lai, H., Boulais, A. P., & Champlain, A. D. (2014). Automated essay scoring and the future of educational assessment in medical education. *Medical Education, 48*, 950-962. doi: 10.1111/medu.12517

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercise Science, 5*(1), 13-34. https://doi.org/10.1207/S15327841MPEE0501_2

Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. Report of the Center for Educator Compensation Reform. Retrieved from https://files.eric.ed.gov/fulltext/ED532068.pdf

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61*(1), 29-48. doi: 10.1348/000711006X126600

Gwet, K. L. (2016). Testing the difference of correlated agreement coefficients for statistical significance. *Educational and Psychological Measurement, 76*(4), 609-637. doi: 10.1177/0013164415596420

Haley, D. T. (2007). *Using a new inter-rater reliability statistic* (Report No. 2017/16). UK: The Open University.

Hamner, B., & Frasco, M. (2018). *Metrics: Evaluation metrics for machine learning* (Version 0.1.4) [Computer Software]. https://CRAN.R-project.org/package=Metrics

Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis, 10*(1), 103-116.

Hoek, J., & Scholman, M. C. J. (2017). *Evaluating discourse annotation: Some recent insights and new approaches*. In H. Bunt (Ed.), ACL Workshop on Interoperable Semantic Annotation (pp. 1-13). https://www.aclweb.org/anthology/W17-7401

Ishioka, T., & Kameda, M. (2006). *Automated Japanese essay scoring system based on articles written by experts*. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, Sydney, 44, 233-240. doi: 10.3115/1220175.1220205

Jang, E-S., Kang, S-S., Noh, E-H., Kim, M-H., Sung, K-H., & Seong, T-J. (2014). *KASS: Korean automatic scoring system for short-answer questions*. Proceedings of the 6th International Conference on Computer Supported Education, Barcelona, 2, 226-230. doi: 10.5220/0004864302260230

Kumar, C. S., & Rama Sree, R. J. (2014). An attempt to improve classification accuracy through implementation of bootstrap aggregation with sequential minimal optimization during automated evaluation of descriptive answers. *Indian Journal of Science and Technology, 7*(9), 1369-1375.

Lacy, S., Watson, B. R., Riffe, D., & Lovejoy, J. (2015). Issues and best practices in content analysis. *Journalism and Mass Communication Quarterly, 92*(4), 1-21. doi: 10.1177/1077699015607338

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Lilja, M. (2018). *Automatic essay scoring of Swedish essays using neural networks* (Doctoral dissertation, Uppsala University). Retrieved from http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1213688&dswid=9250

LoMartire, R. (2017). *rel: Reliability coefficients* (version 1.3.1) [Computer software]. https://CRAN.R-project.org/package=rel

Messick, S. (1993). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and porffolio assessment* (pp. 61-73). New Jersey: Lawrence Erlbaum Associates, Inc.

Meyer, G. J. (1999). Simple procedures to estimate chance agreement and Kappa for the interrater reliability of response segments using the rorschach comprehensive system. *Journal of Personality Assessment, 72*(2), 230-255. doi: 10.1207/ S15327752JP720209

Ministry of National Education (MoNE). (2017a). *Akademik becerilerin izlenmesi ve değerlendirilmesi (ABİDE) 2016 8. sınıflar raporu*. Erişim Adresi: https://odsgm.meb.gov.tr/meb_iys_dosyalar/2017_11/30114819_iY-web-v6.pdf

Ministry of National Education (MoNE). (2017b). *İzleme değerlendirme raporu 2016*. Erişim Adresi: http://odsgm.meb.gov.tr/meb_iys_dosyalar/2017_06/23161120_2016_izleme_degYerlendirme_raporu.pdf

Page, E. B. (1966). The imminence of grading essays by computers. *Phi Delta Kappan, 47*(5), 238–243. Retrieved from http://www.jstor.org/stable/20371545

Powers, D. E., Escoffery, D. S., & Duchnowski, M. P. (2015). Validating automated essay scoring: A (modest) refinement of the "gold standard". *Applied Measurement in Education, 28*(2), 130-142. doi: 10.1080/08957347.2014.1002920

Preston, D., & Goodman, D. (2012). *Automated essay scoring and the repair of electronics.* Retrieved from https://www.semanticscholar.org/

R Core Team. (2018). *R: A language and environment for statistical computing* (version 3.5.2) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing, 18*(1), 25-39. https://doi.org/10.1016/j.asw.2012.10.004

Senay, A., Delisle, J., Raynauld, J. P., Morin, S. N., & Fernandes, J. C. (2015). Agreement between physicians' and nurses' clinical decisions for the management of the fracture liaison service (4iFLS): The Lucky Bone™ program. *Osteoporosis International, 27*(4), 1569-1576. doi: 10.1007/s00198-015-3413-6

Shankar, V., & Bangdiwala, S. I. (2014). Observer agreement paradoxes in 2x2 tables: Comparison of agreement measures. *BMC Medical Research Methodology, 14*(100). Advance online publication. https://doi.org/10.1186/1471-2288-14-100

Shermis, M. D. (2010). Automated essay scoring in a high stakes testing enviroment. In V. J. Shute, B. J. Becker (Eds.), *Innovative assessment for the 21st century* (pp. 167-185). New York: Springer.

Shermis, M. D., & Burnstein, J. (2003). *Automated essay scoring*. Mahvah, New Jersey: Lawrence Erlbaum Associates.

Sim, J., & Wright, C. C. (2005). The Kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy, 85*(3), 257-268. https://doi.org/10.1093/ptj/85.3.257

Siriwardhana, D. D., Walters, K., Rait, G., Bazo-Alvarez, J. C., & Weerasinghe, M. C. (2018). Cross-cultural adaptation and psychometric evaluation of the Sinhala version of Lawton Instrumental Activities of Daily Living Scale. *Plos One, 13*(6), 1-20. https://doi.org/10.1371/journal.pone.0199820

Taghipour, K., & Tou Ng, H. (2016). *A neural approach to automated essay scoring*. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, 1882-1891. doi: 10.18653/v1/D16-1193

Vanbelle, S. (2016). A new interpretation of the weighted Kappa coefficients. *Psychometrika, 81*(2), 399-410. https://doi.org/10.1007/s11336-014-9439-4

Wang, J., & Brown, M. S. (2007). Automated essay scoring versus human scoring: A comparative study. *Journal of Technology, Learning, and Assessment, 6*(2). Retrieved from http://www.jtla.org.

Wang, Y., Wei, Z., Zhou, Y., & Huang, X. (2018, November). Automatic essay scoring incorporating rating schema via reinforcement learning. In E. Reloff, D. Chiang, H. Julia & T. Jun'ichi (Eds.), *Empirical methods in natural language processing* (pp. 791-797). Brussels, Belgium: Association for Computational Linguistics.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2-13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Medical Research Methodology, 13*(61), 1-9.

Appendix A. 2-Category Scored Sample Item Used in the Development of the Software

GÜZEL ATLAR ÜLKESİ: KAPADOKYA

Kapadokya neresidir? Bir şehir, bir ülke yoksa bir bölge midir? Neden her yıl binlerce insan orayı ziyaret eder, yüzlerce kilometre öteden görmeye gelir, dağları geçer, denizleri aşar? Peki, Kapadokya'da ilk önce nereyi ziyaret etmek gerekir? Ne güzel sorular bunlar değil mi! İnsan, öğrenmeye merak etmekle başlar. Sorular sorar, araştırır, bulur, öğrenir. Öğrendikçe de daha bilgili, daha cesur, daha güvenli olur.

Kapadokya, Anadolu ya da Mezopotamya gibi bir bölgenin adı. Nevşehir ilinin sınırları içinde, çok geniş bir alan. 25.000 kilometrekare. Yalnız, oldukça ilginç bir bölge. Bu sebeple binlerce insan her yıl oraya geliyor. Öyle bir bölge ki tarihi "Yontma Taş Devri"ne kadar uzanıyor. Sırasıyla Hititler, Persler, Bizanslılar, Selçuklular ve Osmanlılar yaşamış Kapadokya'da.

Birinci paragraftaki soruların hangisinin cevabı ikinci paragrafta yoktur?

Madde No	16
Bağlam Adı	Güzel Atlar Ülkesi: Kapadokya
Doğru Yanıt (1 Puan) Açıklama	"Kapadokya'da ilk önce nereyi ziyaret etmek gerekir?" sorusuna atıfta bulunan cevaplar doğru cevap olarak kabul edilecektir.
Yanlış Yanıt (0 Puan) Açıklama	Boş cevap ve "Kapadokya'da ilk önce nereyi ziyaret etmek gerekir?" sorusuna atıfta bulunan cevapların haricindeki tüm cevaplar yanlış olarak kabul edilecektir.
Örnek Doğru Yanıtlar	- Peki Kapadoyada en önce nereyi ziyaret etmek gerekir - Kapadokya'da ilk önce nereyi ziyaret etmek gerekir? sorusunun cevabı yoktu? - Kapadokyayı ziyarete gelen ilk önce nereye gider?
Örnek Yanlış Yanıtlar	- Kapadokya neresidir? Sorusunun cevabı yok - NEDEN Binlerce insan orayı ziyaret eder? Peki Kapa dokyada ilk nereyi ziyaret etmek gerekir? - Bir şehirmi yoksa bir ülkemidir

Appendix B. 3-Category Scored Sample Item Used in the Development of the Software

BESLENME

Beslenme çantamda;
Bir dilim ekmek,
Az peynir,
İki bilye, bir topaç
Bir de masal kitabı var.

Gülmeyin arkadaşlar!
Ruhum da doymalı,
Karnımın doyduğu kadar.

Şiire göre, çocuk ruhunu nasıl doyurmaktadır?

Madde No	20
Bağlam Adı	Beslenme
Doğru Yanıt (2 Puan) Açıklama	Çocuğun ruhunu; oyun oynayarak ve kitap okuyarak doyurduğunu ifade eden tüm cevaplar doğru kabul edilir.
Kısmi Doğru Yanıt (1 Puan) Açıklama	Oyun oynar ve kitap okur ifadelerinden sadece birini içeren cevaplar kısmi cevap olarak kabul edilir.
Yanlış Yanıt (0 Puan) Açıklama	Yanlış, ilgisiz ve metinden aynen alınan ifadeler.
Örnek Doğru Yanıtlar	- İki bilyeyi ve bir tane topacı oynayıp, bir masal kitabı okuyarak doyurmaktadır. - 1 bilye bir topaç birde masal kitab okuyup oyunayı Ruhudoyar - Beslenerek, eğlenerek ve okuyarak.
Örnek Kısmi Doğru Yanıtlar	- okuyarak ruhunu doyurma isteğiyle - eğlenerek doyuruyo - Kitap okuyarak, kendini kitabın içine koyarak, ruhunu geliştirip, hissederek.
Örnek Yanlış Yanıtlar	-. iki bilye bir topaç birde masal Kitabi ruhunu doyurmuştur - bir dilim ekmek ,az peynir, iki bilye, bir topaç birde masal kitapı var. - Çocuk ruhunu masal kitabıyla doyurur.

Appendix C. ABIDE 2016 Turkish Test Sample Item Group 1

İSTANBUL DEĞİŞİYOR

İstanbul'da beklenmedik bir şekilde nüfusun artması; gecekonduların çoğalmasına, altyapının kurulmasında sorunlar yaşanmasına neden olmaktadır. Kentlerin dokusunda ise önemli değişmeler görülmektedir.

İstanbul'un eski semtleri olan Beyoğlu, Sirkeci, Eminönü ve Beyazıt'ta ara sokaklarda taş veya ahşap binalar, birbirini kesen dar sokaklar ve caddeler yer almaktadır. Bakırköy, Caddebostan, Etiler, Nişantaşı, Levent gibi yeni semtlerde çoğu kez doğrusal uzanış gösteren ve birbirini dik kesen cadde ve sokaklar vardır. Ataköy, Bahçeşehir gibi planlı olarak kurulan semtlerde ise daha düzenli caddeler yer almakta, çok katlı binalar yapılmaktadır.

7 - 9. soruları yukarıdaki metne göre yanıtlayınız.

7. **Nüfusun olağan dışı artması beraberinde hangi sorunları getirmektedir? Yazınız.**

8. **Metni göz önünde bulundurduğunuzda fotoğrafta görülen yer İstanbul'un hangi semti olabilir? Gerekçesiyle yazınız.**

9. **Metinde altı çizili sözcükle anlatılmak istenen aşağıdakilerden hangisidir?**

A) Yapı
B) Büyüklük
C) Kapladığı alan
D) Gelişmişlik düzeyi

"İSTANBUL DEĞİŞİYOR" Bağlamına Ait Puanlama Anahtarı

Soru No:	**5**
Soru Kodu:	**T-2016-0007**
Bağlam Adı:	**İSTANBUL DEĞİŞİYOR**
DOĞRU YANIT-(2 PUAN)Açıklama	Gecekonduların çoğalması VE altyapı problemlerinin artması sorunlarının her ikisine birden vurgu yapan YA DA bu sorunları genelleyen ifadeleri içeren yanıtlar

Appendix C (continued). ABIDE 2016 Turkish Test Sample Item Group 1

Örnek Yanıtlar	Çarpık kentleşme ve imar sorunları
	Gecekonduların artması ve altyapı problemleri
	Gecekonduların artması ve yapılan yolların yeterli olmaması
KISMI DOĞRU-(1 puan) Açıklama	Metinde geçen iki sorundan "gecekonduların çoğalması" YA DA "alt yapı problemlerinin artması" ifadelerinden sadece birini içeren yanıtlar
YANLIŞ YANIT-(0 Puan) Açıklama	Yetersiz ve belirsiz yanıtlar verir
Örnek Yanıtlar	Kentlerin dokusunda önemli değişmeler görülmektedir
BOŞ-Açıklama	Yanıt kâğıdında soruya ilişkin alanda hiçbir karalamanın ya da işaretlemenin olmadığı yani alanın tamamen boş olduğu durumlar.

Soru No:	**6**
Soru Kodu:	**T-2016-0008**
Bağlam Adı:	**İSTANBUL DEĞİŞİYOR**
DOĞRU YANIT-(2 PUAN)Açıklama	"Beyoğlu, Sirkeci, Eminönü, Beyazıt semtlerinden birinin, birkaçının veya hepsinin adını içeren, gerekçe olarak "Ara sokaklarda taş veya ahşap binalar bulunur." YA DA "Birbirini kesen dar sokaklar ve caddeler bulunur." ifadelerinden birini içeren yanıtlar
Örnek Yanıtlar	Beyoğlu çünkü evler ahşap.
	Sirkeci, Eminönü çünkü ara sokaklarda taş veya ahşap binalar bulunur.
KISMİ DOĞRU-(1 puan) Açıklama	Sadece semt adını içeren ancak gerekçenin yazılmadığı yanıtlar
Örnek Yanıtlar	Beyoğlu
	Eminönü, Beyazıt
	Beyoğlu, Sirkeci, Eminönü, Beyazıt
YANLIŞ YANIT- (0 Puan) Açıklama	Yetersiz ve belirsiz yanıtlar
BOŞ-Açıklama	Yanıt kâğıdında soruya ilişkin alanda hiçbir karalamanın ya da işaretlemenin olmadığı yani alanın tamamen boş olduğu durumlar.

Appendix D. ABIDE 2016 Turkish Test Sample Item Group 2

Soru No:	**7**
Soru Kodu:	**T-2016-0009**
Bağlam Adı:	**İSTANBUL DEĞİŞİYOR**
Doğru Yanıt	**A**

BASINDA OBEZİTE

10.01.2015

12 Yaş Altı Çocuklarda Mobil Cihazların Kullanımının Yasaklanması İçin Bir Sebep: Obezite

Video oyunları ve televizyon, obezitenin artması ile ilişkilidir. Odasında bu tür cihazları kullanmasına izin verilen çocuklarda obezite görülme sıklığı %30 oranında artmaktadır. Obez olan çocukların %30'unda diyabet ortaya çıkmakta, kalp krizi ve erken felç riski artmakta ve ortalama yaşam süresi kısalmaktadır.

15.12.2014
Çocukluk Döneminde Risk: Obezite

Anne ve babanın obez olması, çocuğun yeme alışkanlığı bakımından anne ve babasını örnek alması, çocukların televizyon ve bilgisayar başında çok zaman geçirmesi, stres, kaygı gibi unsurlar çocukluk döneminde obezitenin oluşmasına neden olmaktadır.

10.11.2014
Çocukları Obez Olan Ailelere Para Cezası Geliyor!

Porto Riko'da hükümet, obeziteyle mücadele amaçlı, çocukları fazla kilolu olan anne ve babalara 800 dolara kadar para cezası verilmesini planlıyor. Gelecek nesillerin daha sağlıklı olması için bu uygulamanın yararlı olacağını düşünenlerin sayısı ülkede oldukça fazla.

10 - 12. soruları yukarıdaki metne göre yanıtlayınız.

10. **Gazetelerde obeziteyle ilgili haberlere sıklıkla yer verilmesinin nedeni nedir? Bir ya da iki cümleyle yazınız.**

11. **Mobil cihazların kullanımı obeziteyi neden arttırır? Bir ya da iki cümleyle yazınız.**

12. **Gazete haberlerine göre aşağıdakilerden hangisi söylenebilir?**

A) Obezite ve diyabet birbirleriyle ilişkilidir.
B) Televizyon izlemeyen çocuklar obeziteye yakalanmıyor.
C) Porto Riko'daki para cezası birçok ülkeye örnek olmuştur.
D) Obezite yalnızca çocukluk döneminde ortaya çıkan bir sorundur.

"BASINDA OBEZİTE" Bağlamına Ait Puanlama Anahtarı

Soru No:	**10**
Soru Kodu:	**T-2016-0010**
Bağlam Adı:	**BASINDA OBEZİTE**

Appendix D (continued). ABIDE 2016 Turkish Test Sample Item Group 2

DOĞRU YANIT-(2 PUAN)Açıklama	Obezite ile ilgili bilinçlendirmeye vurgu yapan yanıtlar
Örnek Yanıtlar	"Obezitenin yaygınlaşmasını önlemek için."
	"Halkı bilinçlendirmek için."
	"Obezitenin bir hastalık olduğuna dikkat çekmek."
	"Halkı uyarmak için."
	"Aileleri bilinçlendirmek için."
	"Anne ve babaların önlem almasını sağlamak için." vb.
YANLIŞ YANIT- (0 Puan) Açıklama	Yetersiz ve belirsiz yanıtlar
Örnek Yanıtlar	Para cezasını haber vermek için
BOŞ-Açıklama	- Yanıt kâğıdında soruya ilişkin alanda hiçbir karalamanın ya da işaretlemenin olmadığı yani alanın tamamen boş olduğu durumlar.

Soru No:	**11**
Soru Kodu:	**T-2016-0011**
Bağlam Adı:	**BASINDA OBEZİTE**
DOĞRU YANIT- (1 PUAN) Açıklama	"Uzun süre hareketsiz kalma, çocukların televizyon ve bilgisayar başında çokça vakit geçirmesi" ifadelerini içeren yanıtlar
Örnek Yanıtlar	" Çocukların bilgisayar ve televizyon başında çok zaman geçirmesi."
	"Çocukların bilgisayar başında çok zaman geçirmesinden dolayı hareketsiz kalması."
YANLIŞ YANIT- (0 Puan) Açıklama	Yetersiz ve belirsiz yanıtlar
BOŞ-Açıklama	Yanıt kâğıdında soruya ilişkin alanda hiçbir karalamanın ya da işaretlemenin olmadığı yani alanın tamamen boş olduğu durumlar.

Soru No:	**12**
Soru Kodu:	**T-2016-0012**
Bağlam Adı:	**BASINDA OBEZİTE**
Doğru Yanıt	**A**

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(1); 54-70

The Achievement Gap between Schools and Relationship between Achievement and Socioeconomic Status in Turkey

H. Eren SUNA* Mahmut ÖZER **

Abstract
Nowadays, the performances of education systems are monitored through national and international large-scale studies. In these studies, besides the academic performance of the countries, their status regarding equality in education is also considered. In large-scale studies the relationship between the socioeconomic status and academic achievement and the achievement gap between schools are emphasized. The achievement gap between schools is considered a chronic problem of Turkey, and socioeconomic differences are also considered to be one of the elements of this problem. In this study, the achievement gap between schools and the relationship between socioeconomic status and academic achievement in Turkey were examined through data regarding the last three cycles of TIMSS. For this purpose, multilevel regression analysis was used. The findings showed that although the mean score of Turkey increased between the 2011 and 2019 cycles, the relationship between socioeconomic characteristics and achievement remained at a similar level, with a partial decrease in 2019. These results show that despite the significant increase in Turkey's TIMSS performance in the last cycles, the share of socioeconomic levels on this performance does not increase simultaneously. Another result showed that the achievement gap between schools increased in the last cycle at both grades. Although the relationship between socioeconomic status and achievement does not increase, the widening of the achievement gap between schools may indicate that within-school factors might have stronger relations with achievement. Results revealed that the performance of Turkey in TIMSS increased significantly at 8th grade, and the relationship between socioeconomic status and achievement decreased partially; however, the alleviation of the achievement gap between schools remained a development area for Turkey. Although the relationship between socioeconomic status and academic achievement decreased in the 2019 cycle, the current socioeconomic status role increased the importance of compensating students' socioeconomic disadvantages through educational support programs.

Key Words: Achievement gap, socioeconomic status, equality in education, TIMSS, academic achievement

INTRODUCTION

Education is prominent in ensuring the human and economic development of countries (Brown & Lauder, 1991; United Nations Educational, Scientific and Cultural Organization-UNESCO-UIS, 2018). This power of education in society increases the importance of evaluating the educational process and efficiency (Ross and Jürgens Genevois, 2006). Many different criteria have been used in the evaluation of education systems for many years. Traditional criteria include variables such as access to education, the number of students, teachers and administrators, the average number of students in the classroom, the ratio of students per teacher, and these criteria describe the general structure of the education systems. Following the massification in education, the number and diversity of individuals accessing education increased remarkably, enriching the criteria used in evaluation (Ainscow, 2016; Opertti, 2014). Especially since the 1990s, the criteria for equality in education have been emphasized in educational discussions and evaluations (Beaton, Postlethwaite, Ross, Spearritt, & Wolf, 1999; UNESCO-UIS, 2018). These new criteria focus on the performance of students participating in education and equality in education and the relationship between the characteristics of education systems (Beaton et al., 1999; European Commission, 2020; UNESCO-UIS, 2018).

* PhD., Ministry of National Education, Ankara-Turkey, herensuna@gmail.com, ORCID ID: 0000-0002-6874-7472
** Prof. PhD., Ministry of National Education, Ankara-Turkey, mahmutozer2002@yahoo.com, ORCID ID: 0000-0001-8722-8670

To cite this article:
Suna, H. E., & Özer, M. (2021). The Achievement Gap between Schools and Relationship between Achievement and Socioeconomic Status in Turkey. *Journal of Measurement and Evaluation in Education and Psychology, 12*(1), 54-70. doi: 10.21031/epod.860431

Received: 13.01.2021
Accepted: 28.03.2021

Countries constantly monitor their education systems and make policy changes based on these new performance criteria (Fischman, Topper, Silova, Holloway, & Goebel, 2017, Organization for Economic Development and Cooperation-OECD, 2004). In this regard, countries also benefit from international large-scale assessment outcomes (Beaton et al., 1999, Özer, 2020a, Ross & Jürgens Genevois, 2006). In these studies, the performance of education systems can be compared with other countries and can be examined in a longitudinal way (Mullis, Martin, Foy, Kelly, & Fishbein, 2020; OECD, 2004). Along with the performance of countries in these studies, one of the most frequently focused areas is equality in education (Mullis et al., 2020, OECD, 2019). These studies mainly investigate the effects of out-of-school factors and differences between schools on the academic achievement of students (Mullis et al., 2020; OECD, 2019).

The achievement gap between schools is a problem area for all countries to a diverse extent (OECD, 2004). Considering its stakeholders and teaching processes, each school has different characteristics and it is an expected result that there are small differences in school outcomes. In this context, there are many factors that create the achievement differences between schools, and students' socioeconomic status (SES) is one of these factors. The problem is that students' socioeconomic backgrounds become one of the determining factors in their achievement, and consequently, a significant achievement gap may arise between schools. If a significant achievement gap arisen between schools, then a hierarchy is formed between schools in terms of academic outcomes, and the achievement of the students becomes more dependent on the their school (Ainscow, 2016; Gür, Çelik ve Coşkun, 2013; Önder & Güçlü, 2014). Therefore, a student's academic achievement is more closely related to his/her school (OECD, 2005, 2019). In this case, student groups with more access to high-achieving schools may be more advantageous than others (Suna, Gür, Gelbal, & Özer, 2020b; Willms, 1992). Therefore, the achievement gap between schools indicates negativity for equality in education (OECD, 2008, 2019).

Another criterion evaluated in the context of equality in education is the relationship between socioeconomic characteristics and academic achievement. Numerous studies showed that these characteristics, which are not under the control of students, have a significant relationship with educational performance (Broer, Bai, & Fonseca, 2019; Mullis et al., 2020; OECD, 2019). The differences in the students' socioeconomic status may lead to differences in outcomes from the same education process. In addition, the failure to compensate for the effects of socioeconomic level differences causes these effects to increase the achievement gap between schools (Akyüz, 2014; Alacacı & Erbaş, 2010; OECD, 2019). A clear indication is that in countries with a large achievement gap between schools, these differences are significantly based on differences in students' socioeconomic background (OECD, 2019).

In Turkey, studies on the achievement gap between schools are evaluated on the basis of international large-scale studies and high-stake test results used in national transition systems. Studies focus on the achievement gaps between school types, especially using PISA and TIMSS data (Alacacı & Erbaş, 2010; Berberoğlu & Kalender, 2005; Dinçer & Uysal Kolaçin, 2009; Suna, Tanberkan & Özer, 2020; Yavuz, Demirtaşlı, Yalçın, & Dibek, 2017). For example, in PISA 2003, it was shown that more than 60% of the variance in mathematics literacy scores in Turkey was explained by the achievement gap between schools, and this rate was more than twice the rate in OECD countries (OECD, 2004). The rate was calculated as 61.8% in PISA 2012 (OECD, 2015). These results show that between-school differences are one of the most important determinants of the students' literacy.

The achievement gap between schools and the relationship between socioeconomic status and achievement have been debatable issues in Turkey for many years. Studies focus on the relationship between socioeconomic status and academic achievement show that the strength of relationship is mostly at moderate level, and their relationship with academic achievement is stronger than other variables compared (Acar Güvendir, 2014; Akyüz, 2014; Arifoğlu, 2019; Berberoğlu & Kalender, 2005; Dinçer & Uysal Kolaçin, 2009; Ebrar Yetkiner Özer, Özel & Thompson, 2013; Erdoğan & Acar Güvendir, 2019; Gümüş & Atalmış, 2012; Gür, Çelik & Coşkun, 2013; Kalender, 2004; Karbeyaz, 2019; Koç, 2018; Önder & Güçlü, 2014; Özdemir, 2015; Suna, Tanberkan, Gür, Perc & Özer, 2020a, Suna, Gür, Gelbal & Özer, 2020b). In particular, the achievement gap between schools is considered as

one of the chronic problems of the education system which maintains its effect for a long time despite the changes in the system structure (Berberoğlu & Kalender, 2005; Gür, Çelik & Coşkun, 2013). However, the achievement gap is mostly examined at the secondary education level, and the negativities in this regard are associated with the secondary education level. On the other hand, the results of international large-scale studies show that this problem started in earlier years, and its visibility has increased in secondary education level (Betts, Zau & Rice, 2003; Broer, Bai & Fonseca, 2019; Crenna-Jennings, 2018; Garcia & Weiss, 2017; Mullis et al., 2020; Opdenakker & van Damme, 2006; Shin, Lee & Kim, 2009; Suna et al., 2020a).

The school tracking, which is implemented in the last year of secondary school, is main the reason why these differences have become increasingly visible at the secondary education level (Bölükbaş & Gür, 2020; Özer, 2020a; Özer & Perc, 2020). The tracking of students into school types based on their academic achievement further strengthens the achievement gap between schools. Discussions on the achievement gap between high school types in transition to higher education also increase this visibility (ÖSYM, 2018). However, both studies and international large-scale studies show that the achievement gap between schools started in the first years of education (Cansız, Ozbaylanlı & Çolakoğlu, 2019; Mullis et al., 2020; Suna et al., 2020a). Therefore, the differences in students' access to preschool education and the differences in their socioeconomic background lead achievement differences. Failure to compensate for these differences in the first years of education through various interventions makes the problem more permanent. In other words, the initial advantage increases the later advantage while the disadvantage increases the disadvantage.

In Turkey, many studies have been performed on the relationship between students' socioeconomic background and their achievement (Acar Güvendir, 2014; Ebrar Yetkiner Özel, Özel & Thompson, 2013; Erdoğan & Acar Güvendir, 2019; Karaağaç Cingöz & Gür, 2020; Gelbal, 2008; Özer Özkan & Acar Güvendir, 2014; Suna et al., 2020a; 2020b). The common finding of these studies is that one of the important determinants of student achievement in Turkey is socioeconomic characteristics. In addition, it is shown that the socioeconomic composition of students in schools is also associated with school achievements (Dinçer & Uysal Kolaçin, 2009). This finding indicates that when the socioeconomic characteristics of students are considered at the school level, they become one of the factors that determine the school achievement.

Therefore, student socioeconomic characteristics become one of the main factors in the formation of achievement gaps between schools. In other words, the achievement gap between schools, which is a chronic problem in Turkey, is related to the students' socioeconomic differences in their early years of education. It is very important to focus on these relations correctly to determine the most rational approach and time to implement the support programs. However, studies focus on the difference ap between schools, and the relationship between socioeconomic characteristics and student achievement over time is limited. In the studies conducted, the indicators of the socioeconomic level vary according to the years and the data sets, and a single grade is considered mostly. This study has been structured in a way to consider the change in the relationship between the achievement and these variables over time, at different grade levels. TIMSS data set was chosen because it is a curriculum-based large-scale study and provides information at different grade levels. The approach proposed by Broer, Bai, and Fonseca (2019) was used to ensure the comparability of indicators related to socioeconomic status over time. Thus, it has been made possible to compare students on similar socioeconomic indicators over the years within the scope of TIMSS cycles. Therefore, this study examines the achievement gap between schools and the relationship between socioeconomic status and achievement by using TIMSS 2011, 2015, and 2019 data. In addition, based on the findings, suggestions are made for steps to alleviate the relationship of these out-of-school features with achievement.

Purpose of the research

In this study, it was aimed to determine the achievement gaps between schools and the relationship between socioeconomic status and academic achievement in mathematics and science by using data

from the 2011, 2015, and 2019 TIMSS cycles. For this purpose, answers to the following questions were sought.

1. Does the variance explained by the students' socioeconomic status in mathematics and science achievement change between TIMSS 2011, 2015, and 2019 cycles?

    1.a. Does the variance explained by the socioeconomic status of students at 4th grade level in mathematics and science achievement change between TIMSS 2011, 2015, and 2019 cycles?

    1.b. Does the variance explained by the socioeconomic status of students at 8th grade level in mathematics and science achievement change between TIMSS 2011, 2015, and 2019 cycles?

2. Does the between-school variance in students' mathematics and science achievement change between TIMSS 2011, 2015, and 2019 cycles?

    2.a. Does the between-school variance in mathematics and science achievement of students at 4th-grade change between TIMSS 2011, 2015, and 2019 cycles?

    2.b. Does the between-school variance in mathematics and science achievement of students at 8th-grade change between TIMSS 2011, 2015 and 2019 cycles?

METHOD

Research Design

In this study, the relationship between socioeconomic characteristics and academic achievement between school achievement differences was examined through a correlational design. The relationship between variables were examined without interfering with the nature of the process. In correlational studies, it is aimed to determine whether the characteristics of interest change concurrently without any intervention to the variables and the process, and a relationship exists, the direction and strength of this relationship is determined (Creswell, 2014; Karasar, 2011; Privitera, 2019).

Population and Sample

Based on the fact that Turkey has participated in the TIMSS 2011, 2015, and 2019 cycles with diverse samples in 4th grade, the student population must be defined in two different ways. The student population in 8th grade consists of students continuing formal education in Turkey in 2011, 2015, and 2019. The student population in 4th grade includes students continuing formal education in the 4th-grade level in 2011 and 2015 in Turkey. Additionally, Turkey has participated in a cycle with a 5th grade sample for the first time in TIMSS 2019. In this manner, the student population in the 4th grade level of TIMSS 2019 includes students continuing formal education in the 5th grade level in 2019 in Turkey.

The sample, on the other hand, can be described at two different levels as in the definition of the population. The distribution of students in the sample by years and the socioeconomic characteristics are given in Table 1.

Table 1: Socioeconomic Distribution of Students in Turkish Sample in 2011, 2015 and 2019 TIMSS Cycles*

	TIMSS 2011		TIMSS 2015		TIMSS 2019	
	Grade 4	Grade 8	Grade 4	Grade 8	Grade 5	Grade 8
Number of Students	7.479	6.928	6.456	6.079	4.028	4.077
Number of Schools	257	239	242	218	180	181
Number of Books in the House						
0-10	1.789 (23.9%)	1.301 (18.8%)	1.400 (21.7%)	979 (16.1%)	825 (20.5%)	633 (15.5%)
11-25	2.493 (33.3%)	2.574 (37.2%)	2.162 (33.5%)	2.114 (34.8%)	1.301 (32.3%)	1.266 (31.1%)
26-100	1.927 (25.8%)	1.895 (27.4%)	1.804 (27.9%)	1.835 (30.2%)	1.154 (28.6%)	1.265 (31.0%)
101-200	576 (7.7%)	691 (10%)	525 (8.1%)	622 (10.2%)	391 (9.7%)	514 (12.6%)
More than 200	430 (5.7%)	430 (6.2%)	350 (5% , 4)	475 (7.8%)	231 (5.7%)	355 (8.7%)
Owning a computer / tablet	4.295 (57.4%)	4.035 (58.2%)	3.625 (56.1%)	3.349 (55.1%)	2.961 (73.5%)	2.941 (72.1%)
Owning Work Desk	4984 (66.6%)	4.520 (65.2%)	4.424 (68.5%)	4.433 (72.9%)	2.814 (69.9%)	3.110 (76.3%)
Education Level of Parents						
Primary school or below	-	3.315 (47.8%)	2.575 (39.9%)	1.266 (20.8%)	1.143 (28.4%)	690 (16.9%)
Secondary	-	977 (14.1%)	781 (12.1%)	1.789 (29.4%)	844 (18.5%)	1.192 (29.2%)
High school	-	1.582 (22.8%)	1.662 (25.7%)	1.669 (27.5%)	1.079 (26.8%)	1.094 (26.8%)
Associate Degree	-	314 (4.5%)	394 (6.1%)	321 (5.3%)	256 (6.4%)	274 (6.7%)
University or higher	-	498 (7.2%)	732 (11.3%)	752 (12.4%)	529 (13.1%)	507 (12.4%)

** Information about the education level of parents at the 4th-grade level is collected through the home questionnaire. In the TIMSS 2011, the home survey was conducted only in countries participating in both TIMSS and PIRLS. In this cycle, Turkey did not participate in PIRLS 2011, and there is no information about the education level of parents' in this cycle.*

As seen in Table 1, there are remarkable changes in Turkish samples between 2011, 2015, and 2019 TIMSS cycles in terms of socioeconomic characteristics. First of all, from the TIMSS 2011 cycle to the 2019 cycle, there are significant improvements in the socioeconomic characteristics of the students in the sample. This improvement is clearly seen at the education level of parents. Rates of parents in lower education levels decreased significantly in the 2019 cycle. The second change is that Turkey has participated in TIMSS 2019 cycle 4th grade with a sample of 5th-grade students and declared that the average age of 5th-grade students is more appropriate and comparable with the international average (Ministry of National Education, 2020).

Measurement Tools

Student questionnaire, home questionnaire, and achievement tests in TIMSS 2011, 2015, and 2019 cycles are the measurement tools used in this study. Achievement tests are developed based on TIMSS assessment frameworks and in collaboration between item development experts from participating countries and experts from the TIMSS international center. As a result of the quality control and pilot study, the items to be included in the final tests are determined and 14 booklets are prepared with equal psychometric qualities. The booklets are equated with item response theory-based scaling methods. In mathematics tests for the 4th-grade level, numbers, measurement and geometry and data areas are considered. In the 4th-grade mathematics tests, algebra, geometry, data and probability are considered as subject areas. In science tests, life sciences, physical sciences and earth science are considered at the 4th-grade level; biology, chemistry, physics and earth science are assessed at 8th grade (Mullis et al., 2020).

Within the scope of the study, the criteria used to determine the socioeconomic status (SES) were obtained from the student questionnaire and home questionnaire. In the study, the approach suggested by Broer, Bai, and Fonseca (2019) was used to compare socioeconomic status in different TIMSS cycles. Broer, Bai, and Fonseca (2019) stated that as an indicator of socioeconomic status in TIMSS research, the home educational resources index (HER) is not comparable between cycles, and the elements of this index have increased in the recent TIMSS cycles. In their study, they showed that the socioeconomic indicators that did not change in the twenty years of TIMSS were the number of books at home, the owning of a computer or tablet, having a desk, and the education level of the parents. In addition, in order to make the answer categories of these variables comparable, they created an index with a maximum of 10 points by creating the common categories given in Table 2.

Table 2.Comparable Socioeconomic Indicators Between TIMSS Cycles *

Variable	Level	Score
Education level of parents	Below secondary school	0
	Secondary school level	1
	High school level	2
	Associate degree and equivalent level	3
	University or higher level	4
House facilities	None none	0
	Computer / tablet	1
	Desk	2
Number of books at home	0-10 books	0
	11-25 books	1
	26-100 books	2
	101-200 books	3
	More than 200 books	4

* *Broer, Bai and Fonseca (2019)*

Data Collection and Analysis

The data for the 2011 and 2015 TIMSS cycles used in this study were obtained from the TIMSS database provided by IEA. Student data regarding the TIMSS 2019 cycle were used with the approval number E-65739364-605.01-18900584 of the General Directorate of Measurement and Evaluation of the Ministry of National Education (MoNE).

The achievement differences between schools are examined frequently through *between-schools variance* in international large-scale studies. In social sciences, mostly multilevel modeling (multilevel modeling, hierarchical modeling) approach is used to determine the between schools variance. In this approach, estimates can be made to lower errors in accordance with the nested structure of education (Woltman et al., 2012). In this study, a two-level regression analysis was performed using HLM 8 software. Before performing the multilevel regression analysis, the assumptions were tested and are given in Annex-1. First of all, to test the normal distribution assumption, the skewness and kurtosis indexes of the plausible values at both grade levels were calculated, and it is determined that all the values change between -1 and 1. Similarly, the SES index values formed within the scope of the study changed between -1 and 1. In order to test the linearity assumption, the distribution of SES and plausible values and the distribution of residual values were examined with scatter diagrams. Diagrams show that the relationships between variables are in a linear pattern.

A multilevel nested structure has been designed in which the academic achievement of students at the first level and the characteristics of the schools at the second level. Within the scope of multilevel regression analysis, the *intraclass correlation-ICC* was used. This coefficient allows the variance in student achievement to be divided into two parts: between-schools variance and within-school variance (Brunner et al., 2018; Konstantopoulos, 2007; Raudenbush & Bryk, 2002).

$$ICC_{bs} = \frac{\sigma_B^2}{\sigma_T^2} = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}$$

σ_B^2= Between-school variance (difference in achievement at the mean score level between schools)

σ_W^2= Within-school variance (individual achievement differences between students in schools)

σ_T^2= Total variance (sum of between schools variance and within schools variance).

In international large-scale assessments, students' academic performance is generally determined not with a single indicator, but using plausible values (OECD, 2019; Wu, 2005). Since plausible values are predictions based on students' response patterns, they may have different values about students'

performance. For this reason, it is recommended to make a multivariate analysis that considers all these values to avoid bias (Wu, 2015). In other words, approaches such as choosing only one of the possible values or reducing it to a single value such as average are not recommended because they may cause information loss and biased results (Arıkan, Özer, Şeker, & Ertaş., 2020; Rutkowski, Gonzalez, Joncas and von Davier., 2010; Tat, Koyuncu, & Gelbal, 2019). In the study, HLM 8 software was used to calculate the intraclass correlation and to perform multilevel regression analysis. All five plausible values in science and mathematics were analyzed together to yield unbiased results.

Using the sample weights is another important factor in international large-scale studies (Rutkowski et al., 2010). The samples in these studies consist of students who were selected by weighting in a way to represent students in that country or economy (Rutkowski et al., 2010). Therefore, the number or percentage of students represented in the population by each student in the sample may differ from one another. Similarly, the schools sampled in these studies are selected to represent certain particular school types. In this context, the number or percentage of schools represented in the population by the schools selected as sampling similar to students may also vary. In this manner, sampling weights should be used in order for the unbiased estimates to represent the population (Arıkan et al., 2020; Rutkowski, 2010; Tat, Koyuncu, & Gelbal, 2019). In the TIMSS, both students and schools are weighted and selected by a two-stage sampling methodology (Rutkowski et al., 2010; Mullis et al., 2020). In this study, sampling weights for students (HOUWGT) and schools (SCHWGT) were used in multilevel analysis. HOUWGT is a weighting index developed to weigh the national student sample in the target group (Foy, 2013; Harmouch, Khraibani, & Atrissi, 2017). The HOUWGT, produced by a transformation from the frequently used TOTWGT, and is less affected by sample size differences (Harmouch, Khraibani, & Atrissi, 2017). HOUWGT is preferred in analysis because the Turkish sample sizes show significant changes between TIMSS cycles in each grade level. SCHWGT is the only weighting index commonly used in different TIMSS cycles in weighting schools.

In this study, students' socioeconomic status in Turkish sample was calculated to vary between 0 and 10 through the approach Broer, Bai and Fonseca (2019). In order to determine the students' socioeconomic status at school level, the average socioeconomic level index of the students in each school was taken into account as the average socioeconomic status of that school. Then, the average socioeconomic level of the school was added to the analysis as a second-level explanatory variable in the multilevel regression model.

In the TIMSS, some of the variables regarding students' socioeconomic characteristics are collected through the home questionnaire. In cycles where TIMSS and PIRLS are conducted in the same year, this questionnaire is applied only in the countries participating in both studies. In 2011, Turkey participated only in TIMSS 2011, and some of the socioeconomic variables could not be collected. Therefore, data on the TIMSS 2011 cycle at the fourth-grade level could not be used.

RESULTS

In this section, firstly, results regarding the relationship between socioeconomic status and academic achievement are given. Then, results related to the relationship between school achievement differences and student achievement are presented.

Changes in the Relationship between Students' Socioeconomic Status and Mathematics and Science Achievements in Recent TIMSS Cycles

The variance of mathematics and science achievement explained by the students' socioeconomic status in diverse TIMSS cycles are given in Figure 1.

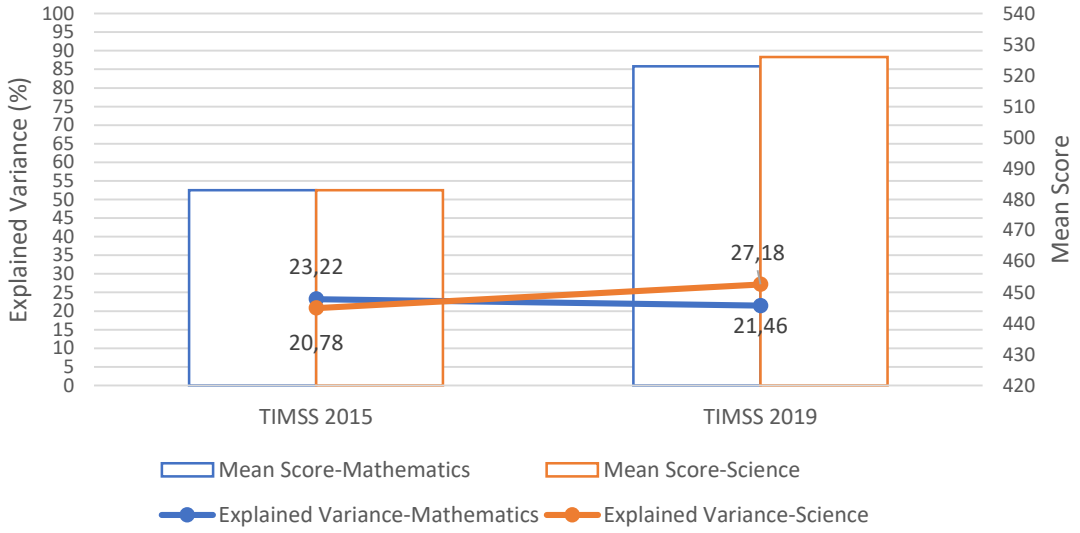

Figure 1. The Variance of Achievement Explained by the Students' Socioeconomic Status at the 4th Grade Level

Figure 1 shows that there may be changes in the relationship between students' socioeconomic status and their academic achievement in diverse TIMSS cycles. The variance explained by socioeconomic status in academic achievement varies between 21.46% and 23.22% in mathematics and between 20.78% and 27.18% in science. Since the sample of 4th-grade level changes in these two cycles, the change in question has been evaluated only with a descriptive perspective.

The fact that Turkey has participated in TIMSS 2019 4th grade level with the 5th-grade sample might lead to changes in the relationship between achievement and socioeconomic status. The impact of this possible factor will be evaluated together with the results of 8th grade, where the grade of sampling did not change.

Another important finding is that socioeconomic status explains a remarkable rate of variance in achievement in the early stages of education, especially in the last year of primary school and the first year of secondary school. In other words, approximately one-fourth of the change in students' achievement in this early period is explained by their socioeconomic status.

The variance explained in the academic achievement of the socioeconomic status of 8th-grade students is given in Figure 2.

As seen in Figure 2, explained variance of mathematics and science achievement by students' socioeconomic status change partially over time. The explained variance varies between 16.93% and 17.91% in mathematics, and 16.54% and 18.08% in science. The other important finding is that the relationship between socioeconomic characteristics and achievement has partially weakened in the transition from 2015 to 2019. Therefore, the variance explained by socioeconomic characteristics in achievement in the TIMSS 2019 cycle has decreased to close to the rates in TIMSS 2011. This finding also indicates that the relationship between out-of-school factors and achievement does not accompany the increase while students' performance increases.

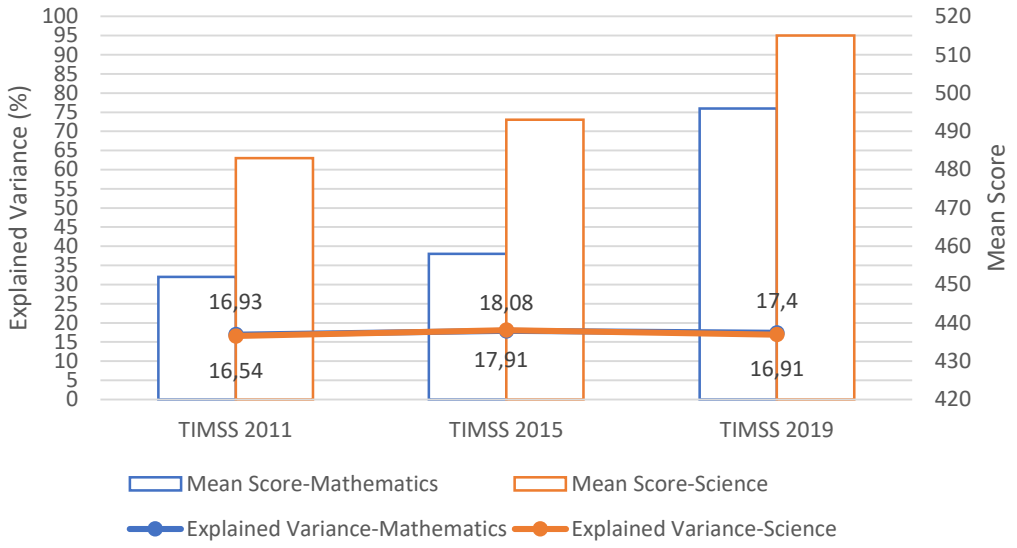

Figure 2. The Variance of Achievement Explained by the Students' Socioeconomic Status at the 8th Grade Level

Explained variances at both grade levels are considered together. It is clearly seen that the relationship between socioeconomic characteristics and academic achievement is higher at the 4th-grade level. This finding is valid for both the 4th-grade sample in 2015 and the 5th-grade sample in 2019. When the findings from different grades are compared, it is predicted that the results of the 4th-grade level may be partially related to the sample change. The relationship between socioeconomic status and academic achievement at the 8th-grade level is relatively weak and shows partial changes between cycles; however, these changes were larger at the 4th-grade level. The results in the next TIMSS cycles will provide detailed information about the impact of participation with the 5th-grade sample.

Between School Variances within Mathematics and Science Achievements in Recent TIMSS Cycles

The between-schools variance explained is analyzed by intra-class correlation and the results regarding the 4th-grade level are given in Figure 3.

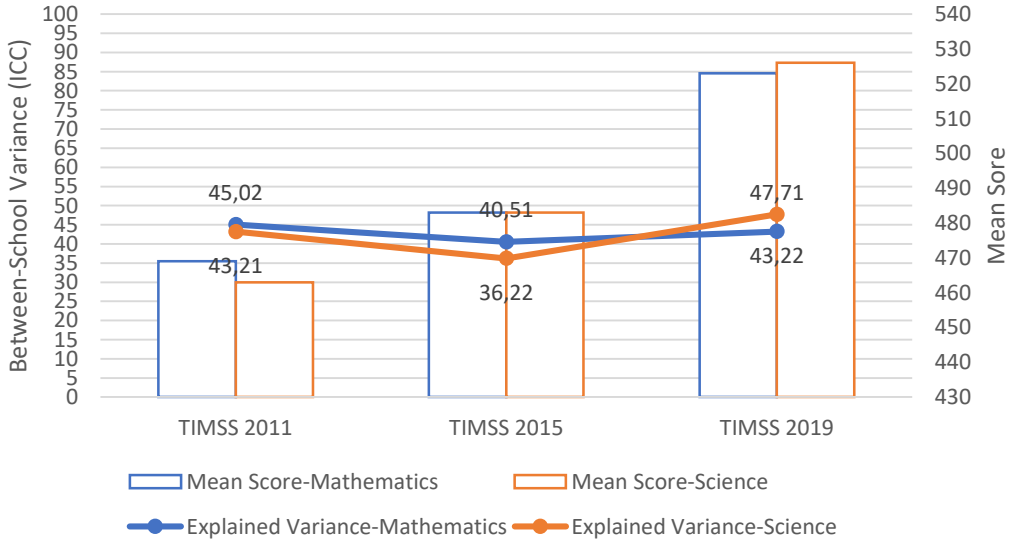

Figure 3. The Between-School Variances at 4th Grade Level in Recent TIMSS Cycles

As shown in Figure 3, the between-school variances change between 40.51% and 45.02% in mathematics, and 36.22% and 47.71% in science in diverse TIMSS cycles. From 2011 to 2015, the cycles that Turkey participated in the study with 4th-grade sample, the achievement gap between schools decreased slightly. On the other hand, between-school variance increased relatively in the TIMSS 2019 cycle, when Turkey participated in the study with a 5th grade sample. In 2011, Turkey's performance in mathematics was at the lowest level although the level of variance explained the inter-school achievement differences were relatively higher. Within the scope of science, the between-school variance decreased from 2011 to 2015 despite the fact that the mean performance of Turkey has increased significantly. An important finding in science is that the between-school variance of science achievement is higher than mathematics achievement in the 2019 cycle when Turkey has participated in the study with the 5th-grade sample.

Findings related to the between-school variance of mathematics and science achievement at 8th-grade students are given in Figure 4.

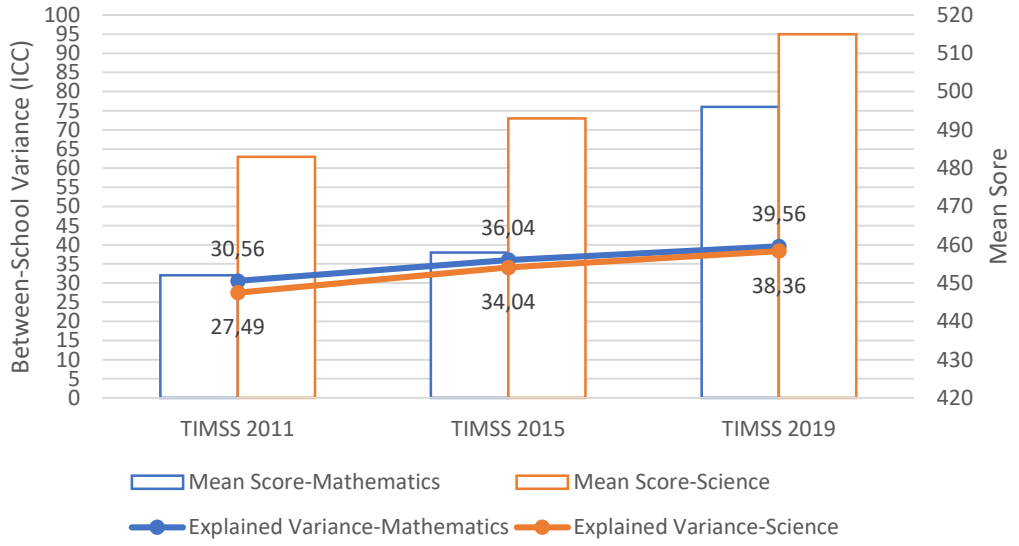

Figure 4. The Between-School Variances at 8th Grade Level in Recent TIMSS Cycles

Figure 4 shows that between-school variance varies between 30.56% and 39.56% in mathematics, and 27.49% and 38.36% in science. Since 2011, the mean score of Turkey has increased in both mathematics and science in 8th-grade; the between-school variance accompanied this increase. The first increase was higher in 2015 (5.48% in mathematics, 6.55% in science), and the second increase was relatively lower (3.52% in mathematics, 4.32% in science) in 2019. However, it is important to emphasize that the between-school variance at the 8th-grade level is lower than the 4th-grade level.

The findings at the 8th-grade level are also important that it provides a reference to the findings at the 4th-grade level. Turkey has participated in TIMSS 2011, 2015, and 2019 cycles with 8th grade, and sampled grade has not changed in this level. The results indicate that the between-school variance at this grade level also increased in the 2015 and 2019 cycles. As seen in Figure 3, this result might indicate that the remarkable change at the 4th-grade level may be partially related to the change in the sample (participation at the 5th grade level in the 2019 cycle).

DISCUSSION and CONCLUSION

The performance of Turkey has increased significantly in international large-scale studies such as PISA and TIMSS since the beginning of the 2000s with several educational indicators. However, the

achievement gap between schools and the effects of non-school factors on educational outcomes has become effective to a diverse extent. These two criteria are also emphasized in international large-scale studies as a performance indicator of the education systems (Mullis et al., 2020, OECD, 2019).

The achievement gap between schools is one of the chronic problems of the education system in Turkey. School tracking (academic segregation) in transition to secondary education also strengthens this gap (Suna et al., 2020a). This result often leads to the illusion that this problem arises in secondary education. However, the results of studies show that the achievement gap between schools begin in the first years of education and have a significant relationship with achievement even in these years (Akyüz, 2014; Mullis, 2020; Önder & Güçlü, 2014; Suna et al., 2020a). The differences between students' socioeconomic status have also become a factor in the achievement gap between schools.

A number of studies have performed on the achievement gap between schools and the relationship between socioeconomic status and academic performance in Turkey (Alacacı & Erbaş, 2010; Berberoğlu & Kalender, 2005; Dinçer & Uysal Kolaçin, 2009; Suna et al., 2020a, 2020b). However, it is seen that the studies conducted mostly focus on the secondary education level or focus on a particular learning area. Therefore, it is important to determine the longitudinal change of the relationship between these variables and academic achievement. This study examines the achievement gap between schools and the relationship between socioeconomic level and achievement in mathematics and science achievement based on the 2011, 2015 and 2019 TIMSS cycles.

In the first research question, the relationship of socioeconomic characteristics on student achievement was examined via the approach suggested by Broer, Bai, and Fonseca (2019). This approach allows comparable socioeconomic status measures across different TIMSS cycles. The results showed that the relationship between the socioeconomic status and academic achievement at the 8th-grade level was similar in the 2011, 2015, and 2019 cycles, and it partially decreased in the 2019 cycle. At the 4th grade level, due to the sample change between 2015 and 2019, the rates are given descriptively. It was found that the relationship between socioeconomic status and academic achievement was stronger in science in 2019, the cycle that the 5th-grade sample participated in. Considering the results at the 8th grade, one of the possible reasons for this change in 2019 was the change in the sample (participation with the 5th grade). The results of future TIMSS cycles will provide reliable and comparable information on the impact of sample change. From 2011 to 2019, Turkey's mean performance increased significantly at the 8th grade, and it is important to show that the relationship between socioeconomic status and academic achievement did not get stronger in this period. This result shows that the increase in mean performance cannot be directly associated with out-of-school factors, but it might more closely related to in-school factors.

The findings regarding the between-school variance showed that the achievement gap in both the 4th and 8th grade explains a significant variance rate in student achievement. The results yield that the between-school variance at the 4th grade is higher than the 8th grade in both 4th and 5th-grade samples. This is important to indicate that the achievement gap between schools has become observable at this early stage. On the other hand, at the end of four years, the between-school variance maintains its existence significantly in the 8th grade. In the 2011 and 2015 cycles, the rates regarding between-school variances in science are consistent with Karbeyaz (2019). In addition, it was shown that the between-school variance was higher in mathematics than in science. These findings indicate that the achievement gap between schools arises before the secondary education level, and the explained variance by socioeconomic status is relatively high.

The findings on the change of between schools variance show that the achievement gap between schools increased in 2015 and 2019. The variance partially increased at the 8th-grade level. In the last two cycles, the increase in the achievement gap is important for indicating that the heterogeneity between schools has increased. However, the findings show that the achievement gap between schools may be more closely related to within-school processes. While the achievement gap between schools increases, the relationship between socioeconomic status and achievement does not accompany this increase. Therefore, it seems more reasonable to associate the reason for the achievement gaps between schools with the within-school factors. For example, in a study by Alacacı and Erbaş (2010), it was found that 55% of the students' achievement differences in PISA 2006 were due to differences in between-school

variance. The important finding of the study is that two-thirds of the variance regarding the achievement gap between schools is explained by the time allocated to mathematics education, the processes in student selection, gender, geographical region, and students' socioeconomic characteristics. Sevgi (2009) showed that the factors that create the gap between schools in TIMSS 2007 are differences in students' socioeconomic levels, the ratio of parents attending school programs, school resources for teaching mathematics and the school climate. Therefore, many factors such as the management of the school, educational resources, the region, educational processes, communication with parents, climate, and the perception of safety and discipline become important factors in the achievement gap between schools. Policies for educational equality need to consider these factors that are shown to be effective on student achievement will also serve to reduce the achievement gaps.

Improving the performance in international large-scale studies is clearly an important achievement for Turkey, with an education system that is more than the total population of many countries. The fact that there is a steady increasing trend in the 2011, 2015, and 2019 TIMSS cycles and that is a clear indicator of this performance increase. Another positive result regarding this increase in performance is that the relationship between socioeconomic status and student achievement remained at a similar level in the 2011, 2015, and 2019 cycles. The study findings show that the increase in performance cannot be directly related to the relationship between students' socioeconomic status and achievement. However, the relationship between the socioeconomic status and academic achievement at the 4th grade in both sampling groups is still stronger than 8th grade. In the early years of education, socioeconomic status explains a remarkable rate of variance in academic achievement, increasing the risks for socioeconomically disadvantaged students. For this reason, interventions to be made in the early stages against disadvantages both solve the problem at an early stage and reduce the intervention cost (Heckman, 2006). It is shown that the dissemination of preschool education and the implementation of academic support programs in the early period provide significant benefits for socioeconomically disadvantaged children (Magnuson, Meyers, Ruhm, & Waldfogel, 2004; Waldfogel, 2015).

On the other hand, numerous projects to mitigate the achievement gap between schools successfully implemented over the years to compensate for the lack of students' learning in Turkey (Ozer, Gençoğlu and Suna, 2020). In these programs, students are provided with multi-dimensional support and the academic deficiencies are compensated. Especially with the Remedial Education & Support Programme in Primary Education (İYEP), which is implemented at primary education, provides an opportunity to alleviate the achievement gap (Gençoğlu, 2019). Increasing the prevalence of İYEP and improving its scope will also strengthen this opportunity. On the other hand, Support and Training Courses (DYK) continue to provide opportunities to compensate for their shortcomings of students from lower- and upper secondary education. It is necessary to expand and increase the diversity of studies that focus on providing multi-dimensional improvement by including especially disadvantaged schools, such as the 1.000 Schools in Vocational Education Project implemented in 2020 (Özer, 2021).

Finally, Covid-19 made it much more likely to increase inequalities in education by the disadvantages of distance education (Özer & Suna, 2020; Özer et al., 2020a). The fact that home resources become more important during the pandemic also increases the possibility that socioeconomic status will be more determinant in student performance in the long term (Özer & Suna, 2020; Özer et al., 2020a). Therefore, the implementation of a comprehensive remedial program as an addition to current support programs has become even more critical for the future.

Limitations

In order to make the socioeconomic status variable comparable over time in the study, the number of sub-criteria has been reduced. Although HER index gives more information about the socioeconomic status of students in TIMSS cycles, fewer criteria were taken into account to maintain comparability between cycles. Another limitation of the study is interpreting the 2019 data descriptively based on the fact that Turkey has participated in this cycle with 5th grade. While comparing the findings of the 2019 cycle with the previous cycles, the possible effect of the sample change was taken into account. The

effect of sample change will be evaluated in detail in future TIMSS cycles. Therefore, the findings regarding TIMSS 2019 were analyzed and interpreted with a descriptive approach.

REFERENCES

Acar Güvendir, M. (2014). Öğrenci başarılarının belirlenmesi sınavında öğrenci ve okul özelliklerinin Türkçe başarısı ile ilişkisi. *Eğitim ve Bilim Dergisi, 39*(172), 163-180, Geniş Ölçekli Test Uygulamaları Özel Sayısı.

Ainscow, M. (2016). Diversity and equity: A global education challenge. *NZ J Educ Stud,* 51**,** 143–155. https://doi.org/10.1007/s40841-016-0056-x

Akyüz, G. (2014). The effects of student and school factors on mathematics achievement in TIMSS 2011. *Eğitim ve Bilim, 39*(172), 150-162.

Alacacı, C., & Erbaş, A. K. (2010). Unpacking the inequality among Turkish schools: Findings from PISA 2006. *International Journal of Educational Development, 30,* 182–192.

Arıkan, S., Özer, F., Şeker, V., & Ertaş, G. (2020). The importance of sample weights and plausible values in large-scale assessments. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 11*(1), 43-60.

Beaton, A. E., Postlethwaite T. N., Ross, K. N., Spearritt, D., & Wolf, R. M. (1999). *The benefits and limitations of international educational achievement studies.* Paris: UNESCO.

Berberoğlu, G., & Kalender, İ. (2005). Öğrenci başarısının yıllara, okul türlerine, bölgelere göre incelenmesi: ÖSS ve PISA analizi. *Eğitim Bilimleri ve Uygulama, 4*(7), 21-35.

Betts, J. R., Zau, A. C., & Rice, L. A. (2003). *Determinants of student achievement: New evidence from San Diego*. San Francisco: Public Policy Institute of California.

Bölükbaş, S., & Gür, B. S. (2020). Tracking and inequality: The results from Turkey. *International Journal of Educational Development*, 78, DOI: 10.1016/j.ijedudev.2020.102262

Broer, M., Bai, Y., & Fonseca, F. (2019). *Socioeconomic inequality and educational outcomes: Evidence from twenty years of TIMSS*. IEA. Springer open Access publication. Retrieved from https://link.springer.com/content/pdf/10.1007%2F978-3-030-11991-1.pdf

Brown, P., & Lauder, H. (1991). Education, economy and social change. *International Studies in Sociology of Education, 1,* 3-23.

Brunner, M., Keller, U., Wenger, M., Fischbach, A., & Lüdtke, O. (2018). Between-school variation in students' achievement, motivation, affect, and learning strategies: Results from 81 countries for planning group-randomized trials in education. *Journal of Research on Educational Effectiveness, 11*(3), 452-478.

Cansız, M., Ozbaylanlı, B., & Çolakoğlu, M. H. (2019). Okul türünün öğrenci başarısı üzerindeki etkisi. *Eğitim ve Bilim, 44*(197), 275-314.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th Edition). California, USA: Sage Publications.

Dinçer, M. A., & Uysal Kolaşin, G. (2009). *Türkiye'de öğrenci başarısında eşitsizliğin belirleyicileri.* İstanbul: Eğitim Reformu Girişimi.

Ebrar Yetkiner Özel, Z., Özel, S., & Thompson, B. (2013). Türkiye'deki sosyoekonomik seviyeye bağlı matematik başarı farklılıklarının Avrupa Birliği ülkeleri ile karşılaştırılması. *Eğitim ve Bilim, 38*(170), 179-193.

European Commission/EACEA/Eurydice (2020). *Equity in school education in Europe: Structures, policies and student performance.* Eurydice report. Luxembourg: Publications Office of the European Union.

Erdoğan, E., & Acar Güvendir, M. (2019). Uluslararası Öğrenci Değerlendirme Programında öğrencilerin sosyoekonomik özellikleri ile okuma becerileri arasındaki ilişki. *Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi. 20*(Özel Sayı), 493-523.

Foy, P. (2013). *TIMSS and PIRLS 2011 user guide for the fourth grade combined international database*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College

Fischman, G. E., Topper, A. M., Silova, I., Holloway, J. L., & Goebel, J. (2017). *An examination of the ınfluence of ınternational large scale assessments and global learning metrics on national school reform policies.* Arizona State University & Center for Advanced Studies in Global Education (CASGE). Retrieved from https://education.asu.edu/sites/default/files/casge_working_papers_2_updated.pdf.

Garcia, E., & Weiss, E. (2017). *Education inequalities at the school starting gate*. Economic Policy Institute. Washington, DC: Economic Policy Institute.

Gelbal, S. (2008). Sekizinci sınıf öğrencilerinin sosyoekonomik özelliklerinin Türkçe başarısı üzerinde etkisi, *Eğitim ve Bilim, 33*(150), 1-13.

Gençoğlu, C. (2019). Millî bir destekleme ve yetiştirme sistemi modeli: İlkokullarda yetiştirme programı (İYEP). *Milli Eğitim, 48*(1), 853-881.

Gümüş, S., & Atalmış, E. H. (2012). Achievement gaps between different school types and regions in Turkey: Have they changed over time?. *Mevlana International Journal of Education, 2*(2), 50-66.

Gür, B. S., Çelik, Z., & Coşkun, İ. (2013). *Türkiye'de ortaöğretimin geleceği: Hiyerarşi mi, eşitlik mi?.* SETA Analiz Raporu No:69. Ankara: SETA.

Harmouch, T., Khraibani, Z., & Atrissi, T. (2017). A multilevel analysis to analyse the TIMSS data: A comparison of the Lebanese and Singapore. *International Journal of New Technology and Researc, 3*(9), 87-102.

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science, 312*(5782), 1900-1902.

Kalender, İ. (2004). *Bir yapısal denklem modellemesi çalışması: ÖBBS 2002 verilerine dayalı olarak sınıf düzeyleri ve okul türlerine göre fen başarısını etkileyen faktörler.* Yayımlanmamış Yüksek Lisans Tezi, Orta Doğu Teknik Üniversitesi, Ankara.

Karaağaç Cingöz, Z., & Gür, B. S. (2020). Ekonomik, sosyal ve kültürel statünün akademik başarıya etkisi PISA 2015 ve TEOG 2017 sonuçlarının karşılaştırması. *İnsan ve Toplum, 10*(4), 247-288.

Karasar, N. (2011). *Bilimsel araştırma yöntemi.* Ankara: Nobel Yayınları.

Karbeyaz, A. S. (2019). *İlkokullarda okullar arası başarı farkı ve nedenlerinin incelenmesi.* Yayımlanmamış Yüksek Lisans Tezi. Marmara Üniversitesi Eğitim Bilimleri Enstitüsü, İstanbul.

Konstantopoulos, S. (2007). *A comment on variance decomposition and nesting effects in two- and three-level designs*. IZA Discussion Paper No. 3178. Retrieved from https://d-nb.info/986455970/34.

Magnuson, K. A., Meyers, M. K., Ruhm, C. J., & Waldfogel, J. (2004). Inequality in preschool education and school readiness. *American Educational Research Journal, 41*(1), 115-157.

Ministry of National Education (2020). *TIMSS 2019 Türkiye ön raporu.* Eğitim Analiz ve Değerlendirme Raporları Serisi No: 15. Ankara: MEB.

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

Organization for Economic Development and Cooperation. (2004). *What makes school systems perform? Seeing school systems through the prism of PISA.* Paris: OECD Publishing.

Organization for Economic Development and Cooperation. (2005). *School factors related to quality and equity: Results from PISA 2000.* Paris: OECD Publishing.

Organization for Economic Development and Cooperation. (2008). *Ten steps to equity in education.* Paris: OECD Publishing.

Organization for Economic Development and Cooperation. (2019). *PISA 2018 results: Where all students can succeed (Volume II).* Paris: OECD Publishing.

Opdenakker, M. C., & Van Damme, J. (2000). Effects of schools, teaching staff and classes on achievement and well-being in secondary education: similarities and differences between school outcomes. *School Effectiveness and School Improvement, 11*(2), 165-196.

Opertti, R., Walker, Z., & Zhang, Y. (2014). Inclusive education: From targeting groups and schools to achieving quality education as the core of EFA. In L. Florian (Ed.), *The SAGE handbook of special education*. London: SAGE.

Ölçme Seçme ve Yerleştirme Merkezi (2018). *YKS değerlendirme raporu.* Değerlendirme Raporları Serisi No: 9. Ankara: ÖSYM.

Önder, E., & Güçlü, N. (2014). İlköğretimde okullar arası başarı farklılıklarını azaltmaya yönelik çözüm önerileri. *Eğitim Bilimleri Dergisi, 40*, 109-132.Ölçme, Seçme ve Yerleştirme Merkezi (2018). *YKS değerlendirme raporu.* Değerlendirme Raporları Serisi No: 9. Ankara: ÖSYM.

Özer Özkan, Y., & Acar Güvendir, M. (2014). Socioeconomic factors of students' relation to mathematic achievement: Comparison of PISA and ÖBBS. *International Online Journal of Educational Sciences, 6*(3), 776-789.

Özer, M. (2020). What does PISA tell us about performance of education systems?. *Bartın University Journal of Faculty of Education, 9*(2), 217-228.

Özer, M., & Perc, M. (2020). Dreams and realities of school tracking and vocational education. *Palgrave Communications, 6*, 34.

Özer, M., Gençoğlu, C., & Suna, H. E. (2020). Türkiye'de eğitimde eşitsizlikleri azaltmak için uygulanan politikalar. *Ondokuz Mayıs Üniversitesi Eğitim Fakültesi Dergisi, 39*(2), 294-312.

Özer, M., Suna, H. E., Çelik, Z., & Aşkar, P. (2020). Covid-19 salgını dolayısıyla okulların kapanmasının eğitimde eşitsizlikler üzerine etkisi. *İnsan ve Toplum, 10*(4), 217-246.

Özer, M., &Suna, H. E. (2020). COVID-19 salgını ve eğitim. M. Şeker, A. Özer ve C. Korkut (Ed.). *Küresel salgının anatomisi: İnsan ve toplumun geleceği içinde* (ss. 172-192). Ankara: TÜBA.

Özer, M. (2021). A new step towards narrowing the achievement gap in turkey: "1,000 Schools in Vocational Education and Training" project. *Bartın Üniversitesi Eğitim Fakültesi Dergisi, 10*(1), 97-108.

Özdemir, C. (2015). *Relationship between equity and excellence in education: Multilevel analysis of international student assessment data with a focus on Turkey.* Yayımlanmamış Doktora Tezi. Orta Doğu Teknik Üniversitesi, Ankara.

Privitera, G. J. (2019). *Research methods for the behavioral sciences* (3rd Edition). California, USA: Sage Publications.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Application and data analysis methods.* (2nd Edition). Newbury Park, CA: Sage.

Ross, K. N., & Jürgens Genevois, I. J. (Eds.) (2006). *Cross-national studies of the quality of education: planning their design and managing their impact.* UNESCO/IIEP/Inwent. Paris: International Institute for Educational Planning.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). *Educational Researcher, 39*(2), 142–151.

Shin, J., Lee, H, & , Kim. Y. (2009). Student and school factors affecting mathematics achievement: International comparisons between Korea, Japan and the USA. *School Psychology International, 30*, 520. doi:10.1177/0143034309107070

Sevgi, S. (2009). *The connection between school and student characteristics with mathematics achievement in Turkey.* Yayımlanmamış Yüksek Lisans Tezi, Orta Doğu Teknik Üniversitesi, Ankara.

Suna, H. E., Tanberkan, H., & Özer, M. (2020). Türkiye'de öğrencilerin okuryazarlık becerilerinin yıllara ve okul türlerine göre değişimi: Öğrencilerin PISA uygulamalarındaki performansı. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 11*(1), 76-97.

Suna, H. E., Tanberkan, H., Gür, B. S., Perc, M., & Özer, M. (2020a). Socioeconomic status and school type as predictors of academic achievement. *Journal of Economy Culture and Society, 61,* 41–64.

Suna, H. E., Gür, B. S., Gelbal, S., & Özer, M. (2020b). Fen lisesi öğrencilerinin sosyoekonomik arkaplanı ve yükseköğretime geçişteki tercihleri. *Yükseköğretim Dergisi, 10*(3), 356-370.

Tat, O., Koyuncu, İ. & Gelbal, S. (2019). The influence of using plausible values and survey weights on multiple regression and hierarchical linear model parameters. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 10*(3), 235-248.

Waldfogel, J. (2015). *The role of preschool in reducing inequality: Preschool improves child outcomes, especially for disadvantaged children*. IZA Disussion Paper. Retrieved from https://wol.iza.org/uploads/articles/219/pdfs/role-of-preschool-in-reducing-inequality.pdf.

Willms, J. D. (1992). *Monitoring school performance: A guide for educators.* London: Falmer.

Woltman, H., Feldstain, A., MacKay, C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology, 8*(1), 52-69.

United Nations Educational, Scientific and Cultural Organization (2018). *Handbook on measuring equity in education.* Montreal: UNESCO Institute for Statistics.

Yavuz, H. C., Demirtaşlı, N., Yalçın, S., & Dibek, M. I. (2017). Türk öğrencilerin TIMSS 2007 ve 2011 matematik başarısında öğrenci ve öğretmen özelliklerinin etkileri. *Eğitim ve Bilim, 42*(189), 27-47.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(1); 54-70

Appendix A. Test of Normality Results for Multilevel Regression Analysis 4th-Grade

TIMSS 2015

	1. Plausible Value Maths	2. Plausible Value Maths	3. Plausible Value Maths	4. Plausible Value Maths	5. Plausible Value Maths	1. Plausible Value Science	2. Plausible Value Science	3. Plausible Value Science	4. Plausible Value Science	5. Plausible Value Science	SES
N	6456	6456	6456	6456	6456	6456	6456	6456	6456	6456	6454
Mean	482,47	482,10	482,79	481,71	482,38	483,75	481,51	481,46	480,40	483,70	3,87
Standard Deviation	95,51	95,66	95,91	96,61	96,09	91,28	92,70	92,99	94,08	92,69	2,43
Skewness	-0,45	-0,45	-0,48	-0,46	-0,45	-0,45	-0,53	-0,51	-0,55	-0,53	0,41
Kurtosis	0,06	0,10	0,14	0,19	0,09	0,12	0,24	0,17	0,32	0,28	-0,54
Minimum	113,06	116,98	72,69	82,47	85,18	149,76	110,22	133,37	70,36	104,52	0,00
Maximum	771,30	765,95	773,32	867,94	784,35	752,90	745,09	746,43	845,11	780,63	10,00

TIMSS 2019

N	4028	4028	4028	4028	4028	4028	4028	4028	4028	4028	4024
Mean	522,00	522,77	521,67	522,36	521,29	526,98	525,37	525,77	525,81	528,25	4,29
Standard Deviation	99,03	98,97	98,33	99,00	99,36	89,11	89,73	90,20	90,06	89,72	2,41
Skewness	-0,39	-0,36	-0,36	-0,33	-0,36	-0,63	-0,65	-0,66	-0,64	-0,66	0,33
Kurtosis	0,00	-0,05	-0,16	-0,06	-0,18	0,39	0,44	0,42	0,40	0,46	-0,61
Minimum	114,79	151,40	135,13	113,22	206,66	168,29	145,69	134,32	150,94	107,55	0,00
Maximum	837,09	785,13	844,85	821,08	791,23	775,04	759,25	768,81	786,97	748,26	10,00

Appendix B. Test of Normality Results for Multilevel Regression Analysis 8th-Grade

TIMSS 2011

	1. Plausible Value Maths	2. Plausible Value Maths	3. Plausible Value Maths	4. Plausible Value Maths	5. Plausible Value Maths	1. Plausible Value Science	2. Plausible Value Science	3. Plausible Value Science	4. Plausible Value Science	5. Plausible Value Science	SES
N	6928	6928	6928	6928	6928	6928	6928	6928	6928	6928	6924
Mean	449,58	448,84	448,05	447,79	448,83	478,48	478,83	479,07	479,57	478,95	3,72
Standard Deviation	109,11	110,59	112,20	111,07	111,13	100,69	101,74	101,02	101,33	101,24	2,44
Skewness	0,16	0,14	0,13	0,14	0,14	-0,08	-0,11	-0,13	-0,10	-0,11	0,49
Kurtosis	-0,16	-0,21	-0,07	-0,14	-0,21	-0,06	-0,13	-0,04	-0,14	-0,16	-0,42
Minimum	105,73	93,32	59,20	44,54	95,53	123,44	113,66	114,58	119,20	88,34	0,00
Maximum	839,23	845,22	875,19	917,68	840,44	882,34	831,76	860,61	818,65	806,75	10,00
TIMSS 2015											
N	6079	6079	6079	6079	6079	6079	6079	6079	6079	6079	6055
Mean	455,85	456,28	455,52	453,28	456,27	490,14	490,92	491,15	489,80	490,85	4,36
Standard Deviation	103,45	103,98	105,03	107,70	105,67	95,85	96,46	96,28	97,30	95,37	2,39
Skewness	0,00	0,03	0,01	0,03	0,02	-0,24	-0,25	-0,27	-0,27	-0,24	0,35
Kurtosis	-0,23	-0,24	-0,19	-0,21	-0,21	-0,10	-0,04	0,01	-0,05	-0,01	-0,48
Minimum	77,86	31,00	55,22	69,84	54,99	115,34	99,30	76,35	124,45	86,73	0,00
Maximum	772,25	780,23	807,90	794,71	784,90	798,05	782,30	772,77	787,02	777,34	10,00
TIMSS 2019											
N	4048	4048	4048	4048	4048	4048	4048	4048	4048	4048	4048
Mean	490,95	492,22	491,64	489,46	490,69	511,07	512,12	511,01	510,17	511,32	4,70
Standard Deviation	107,00	107,19	108,62	109,97	107,83	97,42	96,93	95,63	98,15	96,74	2,43
Skewness	0,06	0,07	0,06	0,00	0,01	-0,19	-0,19	-0,19	-0,19	-0,21	0,22
Kurtosis	-0,16	-0,21	-0,22	-0,19	-0,26	-0,12	-0,10	-0,12	-0,02	-0,08	-0,55
Minimum	128,39	115,80	100,62	91,89	117,26	158,32	103,27	137,28	163,16	100,51	0,00
Maximum	871,37	866,83	888,37	862,05	838,33	815,61	807,82	819,90	815,05	801,05	10,00