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.

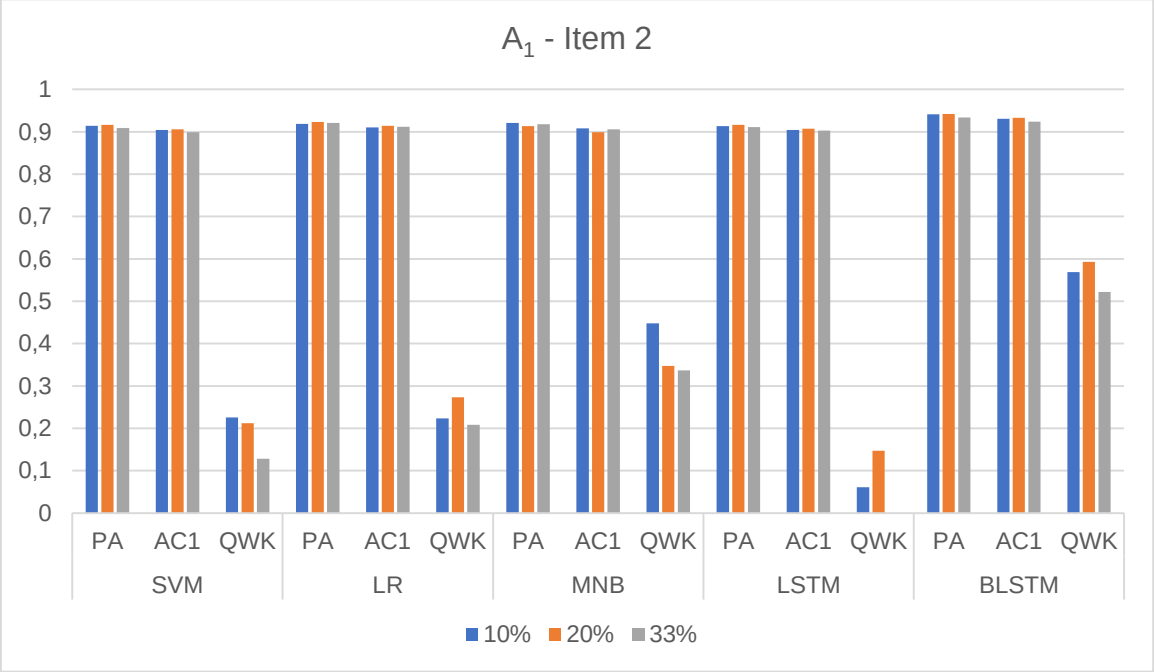

Figure 1. Graph showing Agreement Values for Item 2 in A_1 Booklet according to Automated Scoring Algorithms and Test Data Rates

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.

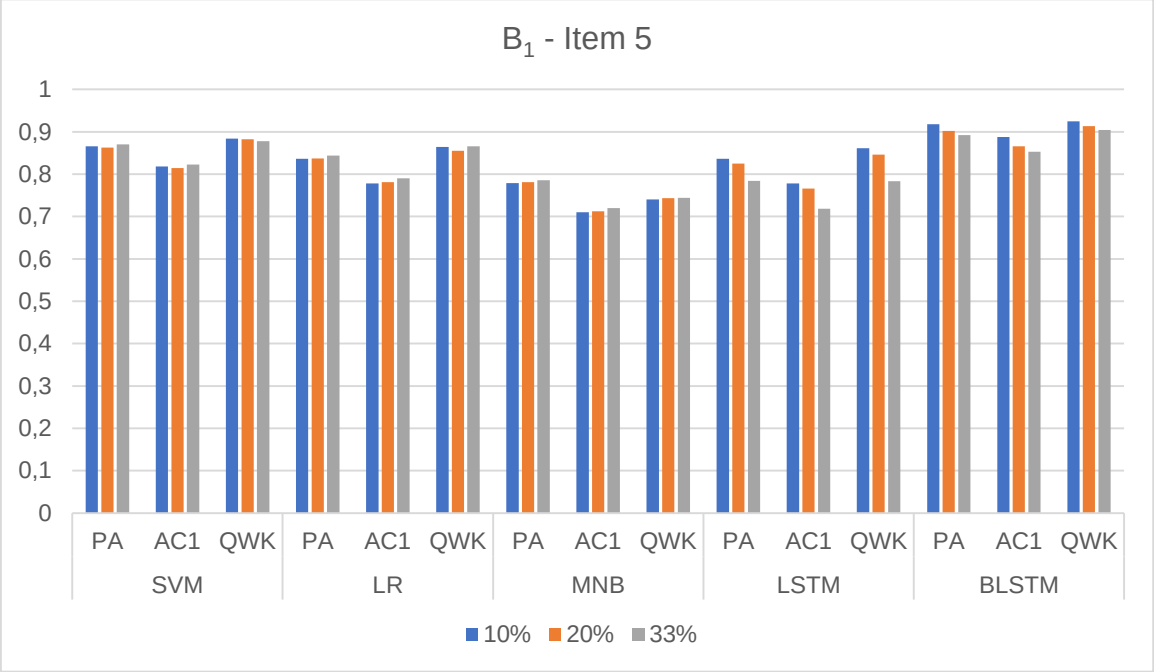

Figure 2. Graph showing Agreement Values for Item 5 in B_1 Booklet according to Automated Scoring Algorithms and Test Data Rates

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.

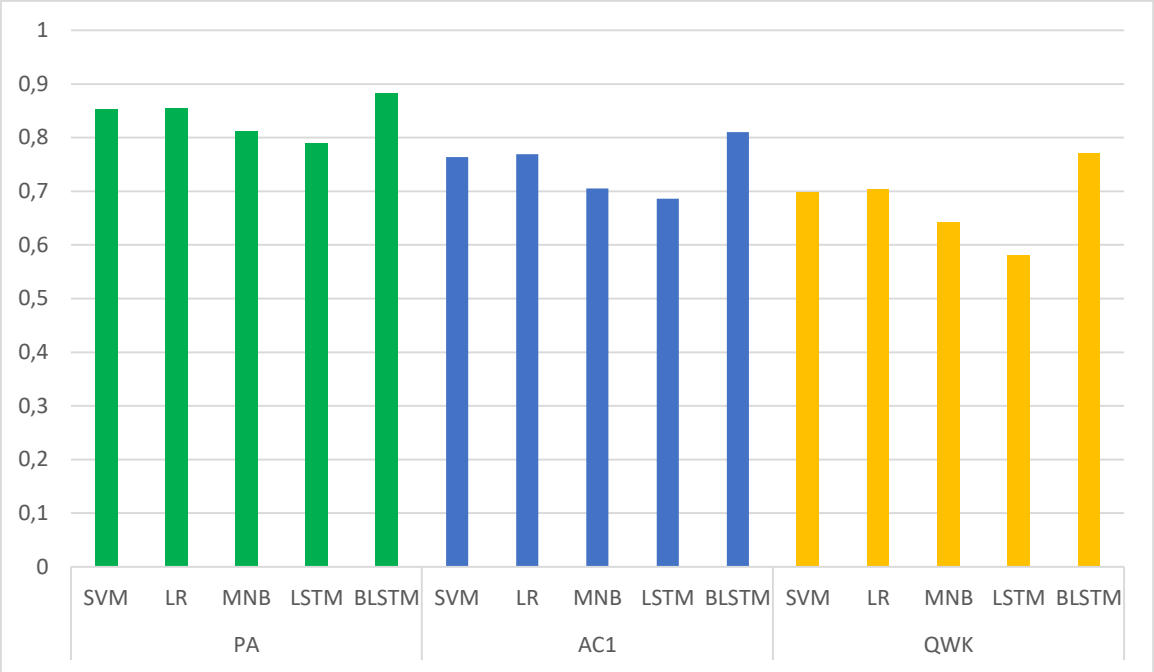

Figure 3. Chart Showing Average Performance of Automated Scoring Algorithms

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**