International Journal of Assessment Tools in Education
2021, Vol. 8, No. 4, 948–958

https://doi.org/10.21449/ijate.894851

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Examining the Discrimination of Binary Scored Test Items with ROC Analysis

Sait Cum [1,*]

[1]Ministry of National Education, Izmir Provincial Directorate, Turkey

ARTICLE HISTORY

Received: Mar. 11, 2021
Revised: Oct. 06, 2021
Accepted: Nov. 08, 2021

Keywords:
Item analysis,
Discrimination,
Roc analysis,
Test development,
Psychometrics.

Abstract: In this study, it was claimed that ROC analysis, which is used to determine to what extent medical diagnosis tests can be differentiated between patients and non-patients, can also be used to examine the discrimination of binary scored items in cognitive tests. In order to obtain various evidence for this claim, the 2x2 contingency table used in the ROC analysis was adapted in accordance with the logic of item discrimination. It was suggested in the article that the areas under the ROC curves (AUC) obtained by using the sensitivity and specificity values calculated with the adapted contingency table can be considered as a measure of item discrimination. The results of the statistical analyses made on the simulation data showed that the AUC values were positively and highly correlated with the D, r_{bis} and a parameter values of the items, and the AUC values from different sized samples were consistent. Additionally, ROC analysis was more stable against range narrowing than other methods. In this respect, it was concluded that very large groups were not needed to examine item discrimination with the proposed method.

1. INTRODUCTION

Osterlind (1990) defined the test item as a unit of measurement that includes a stimulus and a prescriptive response form created to examine mental attributes. Test items provide inferences about a number of psychological and cognitive structures related to the knowledge, ability, or personal characteristics of respondents based on their performance. In binary scored items, the value of "1" indicates that the item is answered correctly, and "0" indicates that the item is answered incorrectly or is left blank. These kinds of items are frequently encountered in achievement and ability tests, in which it is aimed to measure maximum performance. When writing items for purposes such as test development or item pooling, it is necessary to determine the psychometric properties of the items, which is important to obtain valid and reliable measurements. Psychometric properties that provide information about the aspects of items such as difficulty, discrimination, and probability of the item to be answered correctly with chance can be predicted by various statistical or mathematical techniques. Decisions on the using of the item in the test can be made based on these properties. The individuals who possess the knowledge/skill measured by an item are expected to answer that item correctly, and the

*CONTACT: Sait Cum ✉ saitcum@hotmail.com Ministry of National Education, Izmir Provincial Directorate, Turkey

e-ISSN: 2148-7456 /© IJATE 2021

individuals who don't have that knowledge are likely to respond to that item incorrectly. The power of the item to separate these two groups is defined as the item discrimination. It can be stated that the measurement can achive its goal as long as the groups are distinguished from each other. It can be emphasized that discrimination is an important item characteristic since if the measurement reaches its purpose, it is valid.

Prediction to determine the discrimination of the items can be made by various methods. It can be said that the commonly used classical approaches are to determine the biserial correlation coefficients between the item-total test scores (r_{bis}) and to determine the difference between the correct response rate in the upper group and the correct response rate in the lower group (D). These values, which are determined through these approaches and range from -1 to 1, are called item discrimination index.

The item discrimination index, which is calculated over small groups or homogeneous groups according to the ability levels of individuals, can provide misleading information due to the range narrowing (Ebel & Frisbie, 1991; Fulcher & Davidson, 2007). In this respect, Çüm, Gelbal and Tsai (2016) found in their study that item discrimination indexes calculated with the biserial correlation coefficient method showed considerable differences between different small samples.

Discrimination of items can also be examined based on the Item Response Theory (IRT). In IRT, the slope of the item characteristic curve is accepted as the item discrimination parameter (*a* parameter). Although it is stated that the parameter value theoretically changes in the $-\infty$ and $+\infty$ ranges, it usually takes the values between 0 and 2 (or 0 and 3) (DeMars, 2016; Hambleton & Swaminathan, 1985).

In this study, it was claimed that item discrimination can also be examined with the ROC analysis (Receiver Operating Characteristic Analysis). In this respect, a discrimination prediction method, which is not included in the literature, was discussed for the first time in the study.

ROC analysis (Receiver Operating Characteristic Analysis) provides the opportunity to evaluate the performance of medical tests, statistical classifiers, prediction models and algorithms. With the ROC curves created within the scope of the analysis, a graph showing the discrimination performance of a medical diagnostic test (0 = *no disease*, 1 = *disease*) is obtained (Zou et al., 2012). ROC curves are images created to summarize the accuracy of diagnostic predictions (1-0) and they can be used regardless of the source of these predictions. In addition, by comparing the generated ROC curves, the accuracy of different methods used for predictions might be compared (Gönen, 2007). ROC analysis is based on a 2x2 contingency table (Table 1).

Table 1. *The basis of the ROC analysis is a 2x2 contingency table.*

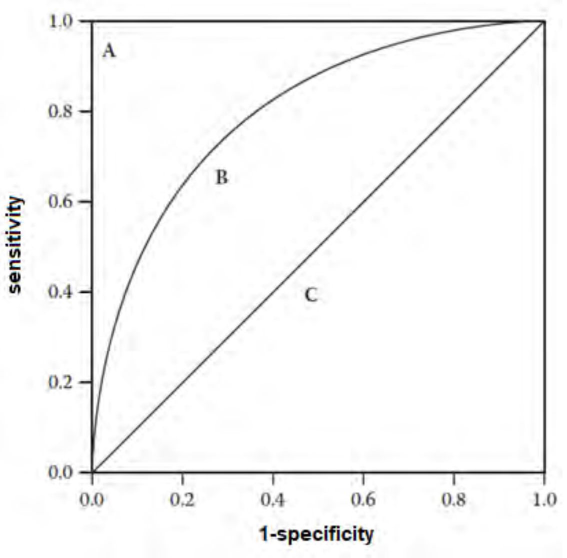
		True Status (Gold Standard)	
		Positive (+)	Negative (-)
Test Result (Result of Diagnosis)	Positive (+)	True Positive (TP)	False Positive (FP)
	Negative (-)	False Negative (FN)	True Negative (TN)
	Total	TP+FN	FP+TN

The variable given as the "test result" in the table, for example, refers to the decisions (the result of diagnosis) based on the scores (values) obtained as a result of a medical test whose effectiveness is examined. The values obtained from the result of the test are included in the analysis as a continuous variable. This continuous variable is then transformed into a two-category (positive, negative) variable, which separates values above and below a given cut-off score. The values in the table will change when the cut-off score changes. In ROC analysis, all possible cut-off points can be tested and optimal cut-off score can be determined by various statistical techniques.

The variable given as a true status in the table is a two-category variable obtained in usually from a more reliable reference test as a result of clinical follow-up or decisions made by a gold standard council, and it separates people into who are really positive or negative. Based on this 2x2 table, two important parameters are explained. The first one is the probability of the diagnostic test to classify a healty person (negative) as healthy, namely specificity; secondly, the probability of the test correctly classifying a patient person (positive) as a patient is sensitivity (Alonzo & Pepe, 2002; Krzanowski & Hand, 2009; Ruopp et al., 2008; Zou et al., 2012).

Considering the change of TP, FN, TN, FP values for each possible cut-off point, the sensitivity is calculated as TP / (TP + FN), and the specificity is calculated as TN / (FP + TN). The ROC curve is the graph obtained from the pairs of sensitivity and 1-specificity calculated from each of the possible cut-off points (Zou et al., 2012).

Figure 1. *ROC curve.*



In Figure 1, the ROC curve (B) is shown at a location in the area between point A and reference axis C. It can be stated that the diagnostic test distinguishes patients and non-patients as well as the ROC curve converges to point A. The C axis is obtained by connecting the points representing the randomness of this distinction. As the curve gets closer to this axis, the discriminative effectiveness of the test decreases. It can be stated that the area under the curve (AUC) is the measure that is generally used to summarize the analysis and provides the opportunity to evaluate the effectiveness of the test. Since the area under the curve will be equal to the area of the square when the curve reaches point A, the AUC value is maximum 1; When the curve coincides with the C axis, the area under curve will be equal to the area of the triangle, so, the AUC value will be 0.5 and this value expresses the randomness in identifying individuals (Krzanowski & Hand, 2009; van Erkel & Pattynama, 1998).

The method proposed in this study was started with the adaptation of the 2x2 contingency table that was taken as basis in the analysis in order to determine the item discrimination with ROC analysis and to use AUC values as the item discrimination measure. For this purpose, the modifications made on contingency table were shown in Table 2.

Table 2. *Contingency table adapted to predict item discriminations.*

		Item Score	
		True (1)	False (0)
Test Total Score	High-scoring Group	High-scoring Group True (HGT)	High-scoring Group False (HGF)
	Low-scoring Group	Low-scoring Group True (LGT)	Low-scoring Group False (LGF)
	Total	HGT+LGT	HGF+LGF

It can be said that a test item is discriminative to the extent that it can distinguish between individuals who have the attribute measured by item and those who do not. Based on the assumption that the item and the test measure the same attribute, in other words, the test is one-dimensional, individuals with high test total scores are expected to answer the item correctly, and individuals with low test total scores are expected to answer the item incorrectly. This underlies the logic of discrimination prediction based on internal criteria. The proposed method also provides an opportunity to examine the discrimination based on optimal internal criteria. The high and low scoring groups mentioned in the table are not the groups consisting of a definite and fixed number of individuals. Some individuals in these groups move to the other group at each cut-off score tested. The combination of these two groups forms the whole group in each case. In the adapted contingency table, the number of individuals in the high-scoring group who answered the item correctly to the HGT section, the number of individuals who answered the item incorrectly to the HGF section; the number of those who are in the low-scoring and who answered the item correctly is written in the LGT section, and the number of those who answered the item incorrectly is written in the LGF section by trying all possible cut-off points.

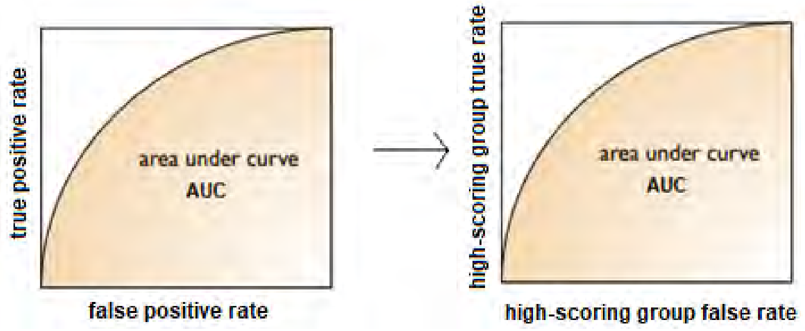
In this case, the sensitivity value gives the probability of an individual in the high-scoring group to answer the item correctly and is calculated as HGT / (HGT + LGT). The specificity value gives the possibility of an individual in the low-scoring group to answer the item incorrectly and is calculated as LGF / (HGF + LGF). The ROC curve, which will describe the item discrimination, is formed from the sensitivity and 1-specificity pairs obtained in the context of all possible cut-off points in accordance with the original analysis. The area under the curve (AUC) determines a measure of the item's discrimination. It is expected that the number of individuals who answered the item correctly from the high-scoring group will increase and the number of those who answered the item from the low-scoring group incorrectly will increase as close to the optimal cut-off score. The approach chosen to determine possible cut-off scores does not affect the basic logic of the analysis. For example, for a 10-item test, the starting point of the cut-off points is 1 point less than the score of the respondent who received the lowest score from the test; the endpoint can be determined to be 1 point more than the score of the respondent who got the highest score from the test. The cut-off points between them can be calculated as the average of each consecutive score pair. In the case where the lowest score obtained from the 10-item test exemplified is 1 and the highest score is 8, all possible cut-off points can be determined as follows:

$$0, \frac{1+2}{2}, \frac{2+3}{2}, \frac{3+4}{2}, \frac{4+5}{2}, \frac{5+6}{2}, \frac{6+7}{2}, \frac{7+8}{2}, 9$$

$$= 0, 1.5, 2.5, 3.5, 4.5, 5.5, 6.5, 7.5, 9$$

The estimation of the area under the ROC curve (AUC) for medical diagnostic tests has been formulated by some researchers (Bamber, 1975; Krzanowski & Hand, 2009; Pepe, 2003). In the method proposed in this study, it was envisaged that AUC values could be used as a measure of the discrimination of items and the formulas were arranged as follows based on the studies of the mentioned researchers (AUC was visualized in Figure 2).

Figure 2. *Area under curve.*



Medical diagnostic tests AUC formula includes test results randomly selected from patient and non-patient populations as variables. In the rearranged method, variables were changed as follows: the test result randomly selected from the high-scoring group was defined as the "H" variable and the test result randomly selected from the low-scoring group was defined as the "L" variable. Four classification possibilities arise from these variables.

S variable denotes test score and *t* variable denotes cut-off score:

1- The probability that an individual from population H will be correctly classified, high-scoring group true:

$$p(S>t|\mathrm{H}).$$

2- Probability of an individual from population L being misclassified, low-scoring group false:

$$p(S>t|\mathrm{L}).$$

3- The probability that an individual from population L will be correctly classified, low-scoring group true:

$$p(S\leq t|\mathrm{L}).$$

4- Likelihood that an individual from population H will be misclassified, high-scoring group false:

$$p(S\leq t|\mathrm{H}).$$

$$\forall x \in (0,1)$$

$$AUC = \int_0^1 ROC(x)dx$$

$$x \to 0 \; ast \to +\infty$$

$$x \to 1 \; ast \to -\infty$$

$$AUC = \int_{-\infty}^{+\infty} p(S > t|\ddot{U}, S = t|A)dt$$

As being unique to this study, the ROC table and the AUC formula used in this study were rearranged in order to examine the discriminations of binary scored test items for the first time.

However, it should be noted that these adaptation attempts were made with regard to item discrimination logic, they do not change mathematical basis of the analysis.

In the examinations made with the method suggested in this article, it was expected to determine the measures reflecting the item discrimination feature with less errors. Because it was thought that taking into consideration many possible cut-off scores in the calculations made with the proposed method would increase the precision of the measurements. The value for the area under the ROC curve, proposed as a measure of item discrimination, is a combination of observations for the functioning of the item under a number of different conditions. On the other hand, it was thought that the 2x2 contingency table, which was taken as the basis in ROC analysis and adapted to examine the item discrimination with this method, coincides logically and psychometrically with the concept of item discrimination. Comparison of the proposed method in this study with the currently used methods was important in terms of revealing the advantages and disadvantages of the approach. In addition, the sample size is also a matter of debate when it comes to choosing a method for determining the psychometric properties of the items. In this context, it was considered important to test the consistency of the proposed method between different samples in terms of size and score distribution. In the literature review, no study was found in which ROC analysis was used to examine item discrimination. In this sense, it can be stated that this study is important for the psychometrics literature. This study will pave the way for other advanced studies. The usage of the proposed method on the determination of the psychometric properties of the items can be advanced and extended by other psychometrists. It is also thought that ROC analysis can be easily carried out with many statistical software, especially SPSS, and it will provide convenience for test developers and test practitioners in terms of ease of calculation.

1.1. Purpose of the Study

The aim of this study was to compare the item discrimination measures obtained from different methods with simulation data and to examine the consistency of the measurements obtained from ROC analysis between samples of different sizes and different distribution characteristics. For this purpose, the answers to the following questions were sought in the study.

1- What are the correlations between the upper-lower groups item discrimination indexes (D), biserial correlation coefficients of the item-total test scores (r_{bis}), *a* parameters from IRT-2PLM, and the AUC values obtained from 20 items and 1000 respondents?

2- What are the correlations between the AUC values obtained from 20 items and 100, 200, 400, 1000 respondent groups, and is there a statistically significant difference between these AUC values?

3- To what extent are the values determined by different item discrimination prediction methods invariant in case of range narrowing?

2. METHOD

This study is a basic (pure) simulation research aimed at producing new information.

2.1. Data Set

Within the scope of the study, 20 binary scored test items were simulated. Values of the discrimination parameters (*a*) of these items vary between 0 and 2 and values of their difficulty parameters (b) ranged between -2 and 2. In addition, a group of 1000 respondents whose ability values (θ) ranged between -2 and 2 were simulated.

2.2. Data Analysis

The data handled within the scope of the study were produced in WinGen software and made ready for the analysis. To find the answer to the first research question, D indexes were calculated based on the correct response rates of the items in the upper-lower groups in terms of test scores, biserial correlation coefficients were calculated between the scores of each item and the total test scores, and finally, the areas under the ROC curve (AUC values) for each item were calculated (proposed method). Since the data were simulated based on the Item Response Theory, the generated *a* parameters were directly used. Correlations among the item discrimination measures obtained based on four different methods were determined by Spearman's rank-order correlation coefficient method.

In order to answer the second research question, AUC predictions were made by using randomly determined samples of 100, 200 and 400 respondents from the sample of 1000 respondents produced. Correlations between values obtained from samples of different sizes were examined by Spearman's rank difference correlation coefficient method. In addition, the significance of the differences between the predictions was examined with Kruskal Wallis-H. Same analyzes were made and reported with other methods in order to make comparisons.

In order to answer the third research question, the dataset was sorted in ascending order in terms of total test scores. The score range is narrowed by dividing the lowest-scoring 33% and the highest-scoring 33% of the group. The correlation coefficients between the item discrimination measures obtained from these narrowed-range groups and the full dataset were calculated by Spearman's rank difference correlation coefficient method. These analyzes were performed for each of the four different methods.

When using smaller datasets selected from the full dataset, analyzes based on IRT were performed with R (ShinyItemAnalysis) to obtain the *a* parameters. TAP and SPSS V23 statistical softwares were also used to analyze the research data. For the ROC analysis, the positive value of the real state variable was determined as “1” (items scored as 1-0). Sensitivity and specificity values were determined based on the assumption that larger test scores indicate more positive test results. Nonparametric approach was preferred for the predictions of the areas under the ROC curve.

3. RESULTS / FINDINGS

In order to find an answer to the first research question of the study, the values regarding the discrimination of the items were predicted based on the proposed method and the other three methods, and the findings were given in Table 3.

Table 3. *Values predicted by different methods regarding item discrimination.*

Item	1	2	3	4	5	6	7	8	9	10
D	0.470	0.250	0.770	0.560	0.710	0.350	0.460	0.500	0.320	0.720
r_{bis}	0.672	0.262	0.808	0.595	0.800	0.380	0.572	0.531	0.381	0.752
a	0.999	0.118	1.717	0.705	1.580	0.270	0.870	0.679	0.405	1.247
AUC	0.837	0.620	0.880	0.775	0.873	0.674	0.773	0.747	0.681	0.846

Item	11	12	13	14	15	16	17	18	19	20
D	0.350	0.740	0.420	0.670	0.620	0.660	0.200	0.400	0.610	0.510
r_{bis}	0.425	0.821	0.450	0.745	0.643	0.707	0.192	0.705	0.775	0.679
a	0.538	1.539	0.476	1.430	0.992	1.271	0.630	1.473	1.537	1.321
AUC	0.701	0.888	0.709	0.851	0.797	0.829	0.586	0.861	0.874	0.831

Hosmer and Lemeshow (2000) stated that if the AUC value is equal to 0.5, no discrimination can be mentioned, it is acceptable if the value is between 0.7 and 0.8, perfect if it is between 0.8 and 0.9, and an extraordinary distinction if it is greater than 0.9. Considering this view, it can be suggested that the AUC value should be above 0.7 for a good item discrimination. When Table 3 was analyzed, it was seen that items with AUC values below 0.7 (item no 2, 6, 9, 17) have low D and r_{bis} values. It was also determined that these items had very low or low discrimination in terms of *a* parameters. This determination was made according to Baker's (2001) criteria that the value of *a* parameters can be interpreted as very low discrimination in the range of 0.01 - 0.34, low discrimination in the range of 0.35 - 0.64, medium discrimination in the range of 0.65 - 1.34, high discrimination in the range of 1.35 - 1.69, and very high discrimination is greater than 1.70.

The correlation coefficients between the values in Table 3 were determined and the related findings were given in Table 4.

Table 4. *Correlations between predictions made by different methods.*

Method	D	r_{bis}	*a*	AUC
D	1			
r_{bis}	0.912*	1		
a	0.830*	0.970*	1	
AUC	0.838*	0.979*	0.977*	1

*Correlations are statistically significant at the 0.01 level.

Based on the findings, it can be interpreted that D, r_{bis}, *a*, and AUC values provide similar information on determining the item discrimination. All correlation coefficients showed a positive and high correlation between all pairs of prediction methods. In addition, the *a* parameters obtained based on the Item Response Theory showed the highest correlation with the AUC values obtained based on the ROC analysis among all other methods. This was noted because the Item Response Theory currently prevails among the test theories.

In order to find the answer to the second research question, the AUC values of the same items from 100, 200 and 400 groups determined randomly from the full data set of 1000 respondents were predicted and the correlations of the obtained values between the different groups were given in Table 5.

Table 5. *Correlations between AUC values obtained from different sized samples.*

Sample size	100	200	400	1000
100	1			
200	0.916*	1		
400	0.836*	0.930*	1	
1000	0.791*	0.912*	0.986*	1

*Correlations are statistically significant at the 0.01 level.

When Table 5 was examined, it was determined that there were positive high correlations between AUC values obtained from different sized samples. The lowest correlation coefficient (0.791) was among the samples consisting of 1000 and 100 respondents while the highest correlation coefficient (0.986) was among the samples consisting of 1000 and 400 respondents. In addition, the statistical significance of the differences between the AUC values obtained from different samples was examined with the Kruskal Wallis-H test and it was found that the p value of the test was 0.876. Findings showed that the predictions for the areas under the ROC curve were similar between different sized samples and the differences between the values are not statistically significant. In this regard, it can be inferred that large samples are not required to examine item discriminations with the proposed method.

Similar comparisons were also made between predictions made by other methods from different sized samples. The correlation coefficients between the D values obtained from four different sized samples ranged from 0.867 to 0.997. The correlation coefficients between r_{bis} values ranged from 0.783 to 0.977. Finally, correlation coefficients between the *a* parameters were in the range of 0.823 and 0.979. In addition, Kruskal Wallis-H test results showed that there was no statistically significant differences between the compared predictions. In this sense, it cannot be claimed that the AUC method provides an advantage over the other methods regarding this comparison.

The third research question was about examining the effect of range narrowing on the predictions. Accordingly, predictions were obtained from the lower 33% and upper 33% parts of the dataset in terms of the total test scores with different methods. Correlations of these predictions with each other and with the full dataset were determined for each method. Findings were given in Table 6.

Table 6. *Correlations between narrowed-range datasets and full dataset.*

	Lower 33%- Full Data	Upper 33%- Full Data	Lower 33%-Upper 33%
D	0.102	0.126	-0.624**
r_{bis}	0.640**	0.421	0.405
a	0.496*	0.257	0.691**
AUC	0.744**	0.586**	0.526*

** Correlations are statistically significant at the 0.01 level.
* Correlations are statistically significant at the 0.05 level.

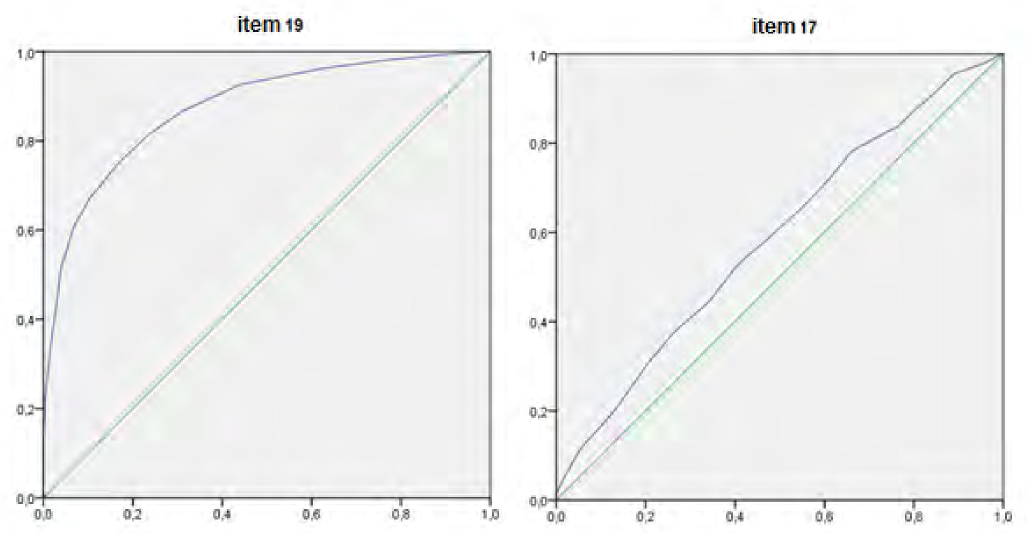
In the analyzes performed in terms of the invariance of item characteristics in case of range narrowing, it was determined that the most unstable indexes were the D indexes, which were calculated with the correct response rates of the upper and lower groups. This finding indicated that it would not be appropriate to use this method with homogeneous groups in terms of test scores. On the other hand, AUC values were the measures that showed the best performance compared to others, especially in terms of higher correlation between narrowed-range datas and full data values. Accordingly, it can be argued that the use of ROC analysis would be more appropriate than other methods when determining item discriminations with homogeneous groups.

4. DISCUSSION and CONCLUSION

As a result of the findings obtained from the comparisons made in this study, various evidence has been obtained for the claim that the ROC analysis, which is used to determine the degree of discrimination between patients and non-patients, especially with medical diagnostic tests, can also be used to examine item discrimination. In the proposed method, it was determined that the AUC values, which were accepted as a measure of item discrimination, positively and highly correlated with the D, r_{bis} and *a* parameter values of the items. In addition, it was interpreted that the items with AUC values below 0.7 were low or very low discriminative items based on the values of D, r_{bis} and *a* parameters.

Based on the aforementioned findings, it was concluded that the criteria proposed by Hosmer and Lemeshow (2000) for interpreting the area under the ROC curve can also be accepted if the analysis is used for the study of item discriminations. It can be stated that the AUC values obtained by the proposed method should take at least 0.7 in order for the discrimination of the items to be acceptable, and the discrimination of the items increases as this value gets closer to 1. In this study, it was concluded that AUC values were consistent between different sized samples and that large samples were not required to examine item discrimination with the proposed method. In addition, it was determined that the AUC values were affected less negatively compared to other methods if the score distributions in the group were homogeneous. This was noted as a very important advantage of determining item discriminations with ROC analysis. It should also be mentioned that, with the proposed method, item discriminations can be examined not only with AUC values, but also with ROC curve graphs.

Figure 3. *ROC curves showing the discrimination of two different items.*



It can be stated that as the ROC curve gets closer to the upper left corner of the graph, the discrimination of the examined item increases. As seen in Figure 3, the discrimination of item 19 is high, and the item 17 is low. It is thought that the method proposed in this study may also be advantageous in terms of ease of interpretation by providing visual expression of the discrimination of the items.

The ROC curves method adapted to item analysis can be recommended to test developers, test practitioners, and other researchers since it provides consistent predictions, and it does not require very large groups for these predictions. In this respect, proposed method can be added to the literature as an alternative method. Other researchers working in the field of psychometrics may develop or criticize the method from various aspects as well.

Declaration of Conflicting Interests and Ethics

The author declares no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author.

ORCID

Sait Cum https://orcid.org/0000-0002-0428-5088

5. REFERENCES

Alonzo, A.T., & Pepe, S. M. (2002). Distribution-free ROC analysis using binary regression tecniques. *Biostatistics*, *3*(3), 421-432. https://doi.org/10.1093/biostatistics/3.3.421

Baker, F.B. (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, *12*(4), 387–415. https://doi.org/10.1016/0022-2496(75)90001-2

Çüm, S., Gelbal, S., & Tsai, C-P. (2016). Examination of the consistency of the sato test theory item parameters obtained from different samples. *Journal of Measurement and Evaluation in Education and Psychology, 7(1)*, 170-181. https://doi.org/10.21031/epod.69276

DeMars, C. (2016). *Madde tepki kuramı [Item response theory]*. Nobel.

Ebel, R.L., & Frisbie, D.A. (1991). *Essentials of educational measurement*. Prentice-Hall Inc.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book.* Routledge.

Gönen, M. (2007). *Analyzing Receiver Operating Characteristic Curves with SAS*®. SAS Institute Inc.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Springer.

Hosmer, D.W., & Lemeshow, S. (2000). *Applied logistic regression*. John-Wiley & Sons, INC.

Krzanowski, W.J., & Hand, D.J. (2009). *ROC curves for continuous data.* Chapman and Hall/CRC Press.

Osterlind, S. J. (1990). Toward a uniform definition of a test item. *Educational Research Quarterly, 14(*4*)*, 2-5.

Pepe, M.S. (2003). *The statistical evaluation of medical tests for classification and prediction*. University Press, Oxford.

Ruopp, D. M., Perkins, J. N., Whitcomb, W. B., & Schisterman, F. E. (2008). Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biometrical Journal*, *3*, 419-430. https://doi.org/10.1002/bimj.200710415

Van Erkel, A.R., & Pattynama, P.M. (1998). Reciever operating characteristic analysis: Basic principles and applications in radiology. *European Journal of Radiology 27*, 88-94. https://doi.org/10.1016/S0720-048X(97)00157-5

Zou, H. K., Liu, A., Bandos, I.A., Onho-Machado, L., & Rockette, E. H. (2012). *Statistical evaluation of diagnostic performance topics in roc analysis.* CRC Press.