

Received: April 18, 2018

Revision received: August 28, 2018

Accepted: October 31, 2018

Copyright © 2019 ESTP

www.estp.com.tr

DOI 10.12738/estp.2019.1.0116 ♦ February 2019 ♦ 19(1)♦3-21

Article

Anxiety in Prospective Teachers: Determining the Cut-off Score with Different Methods in Multi-Scoring Scales*

D. Bahar Şahin Sarkın¹
Istanbul Okan University

H. Deniz Gülleroğlu²
Ankara University

Abstract

This study contributes to validity studies by determining the cut-off point of an inventory measuring university students' anxiety levels with Angoff, ROC, and Borderline methods and by examining high/low anxiety levels according to these methods point. The study is regarded as a basic research due to the newly-added data in a multi-scoring inventory which validity and reliability studies were previously performed. This study included 290 participants who studied at the Education Faculty of a state university and were planning to take the Public Personnel Selection Examination. Trait Anxiety and Test Anxiety Scales were applied to the students. Findings obtained from this study show that the cut-off scores (48 and 48.5) calculated by Angoff and ROC analyses are very close to each other. The consistency coefficient among the judges was examined to prove the reliability of the cut-off points determined by Angoff and Borderline methods. According to the AUC (area under the curve) value obtained in the study, it is seen that the Trait Anxiety Inventory could correctly classify high anxious and low anxious individuals at the rate of 73% (moderate level). It is shown that the scale is quite likely to determine anxiety levels when it is used for screening rather than diagnosis. As a conclusion, the study indicates that when ROC analysis is used to determine the cut-off score of a multi-scoring psychological measuring instrument, such characteristics as sensitivity, specificity, and positive/negative predictive values that other methods do not have provide more detailed and objective information.

Keywords

Multi-scoring • Cut-off Score • Angoff • ROC • Borderline

*This study is based on a summary of the doctoral dissertation entitled "Determining the Cut-off Score with Angoff, Roc and Border Group Methods in Multi-Scoring Scales" prepared in the Educational Measurement and Evaluation Program, Ankara University.

¹Correspondence to: D. Bahar Şahin Sarkın (PhD), Department of Educational Sciences, Istanbul Okan University, 34959 Tuzla Istanbul Turkey. Email: bahar.sarkin@okan.edu.tr

²Department of Measurement and Evaluation in Education, Ankara University, 06590 Cebeci Ankara Turkey. Email:denizgulleroglu@yahoo.com

Citation: Şahin Sarkın, D. B. & Gülleroğlu, H. D. (2019). Determining the Cut-off Score with Angoff, Roc and Border Group Methods in Multi-Scoring Scales. *Educational Sciences: Theory & Practice*, 19(1), 3-21. <http://dx.doi.org/10.12738/estp.2019.1.0116>

Subjectivity of psychological variables that are not directly observable but can be predicted through observable behaviours is carried to objectivity by means of measurement instruments. In the set of rules regulating the relationship between the measurer and the measured; the measurement instrument is the observing body, and the responses are the observed body. In psychological context where physical stimuli are replaced by psychological stimuli, theories and techniques of measuring have been introduced resulting in development of psychometrics (Guilford, 1954). Psychometrics is defined as an approach that focuses on identification and statistical analysis of personality traits, intelligence and abilities of individuals with various tests in psychology (Nunnally, 1970). In psychometric approach, it is possible to observe the qualifications sought in an individual by determining the magnitude of a certain trait by means of psychological tests which are objective and standard measures of a behavioural sample (Anastasi, 1988). Still, it must be remembered that tests are not mechanisms that make decisions about people on their own, rather they are tools developed to understand individuals and provide important information (Cronbach, 1984).

Psychological tests are used in many fields of social sciences, especially in education, psychology, sociology and psychiatry, which study human behaviours. In Turkey, researchers working in these fields use such tests in mental health centres, psychiatry clinics, counseling centres, industry and workplaces as much as for training and scientific research (Öner, 1997). Regardless of the affective trait under scrutiny, generalizability and functionality of the findings from psychological tests are in parallel with reliability and validity of the measurement instrument used. Thus, although each stage of research process is important, measurement and evaluation of variables that are not directly observable depends on the nature of the measurement instruments used. The most neglected part of scale development or adaptation studies is the lack of clear reporting of how to interpret scores obtained from scales (Erkuş, 2007). Scale users face questions such as how to evaluate the person with the scale score, whether it is possible to obtain a total score or subscales must be considered separately, whether the increase in scores corresponds to an increase or decrease in the variance, and in what way total scores should be interpreted. In this regard, it is important to know how the criterion is created and how levels of individuals' responses differ in order to be able to evaluate the scores obtained. Departing from this, the need to examine methods for determining the cut-off score rose in order to be able to evaluate individuals' reactions in multiple scoring psychological tests and classify their levels.

Cut-off score is the threshold which groups individuals according to their measured levels and which helps interpret these levels (Crocker and Algina, 1986; Irwin, 2007). Each test has its own standards and these standards may change over time. During the determination of cut-off score, the knowledge of the relation of the judges and psychological structure to variables that can be directly or indirectly observed distinguished the standard setting and Border value detection methods.

Standard Setting Methods

There are many methods to be used in standard setting process. Jeager (1989) defined the methods on the basis of testing as "test-centered" methods, and those based on students as "student-centered" methods.

In test-centered methods, judges try to determine the cut-off score for the minimum level in classification discrimination of a particular response level, ie, affective trait, by examining the items in the test. The cut-off

score to be determined is calculated according to the response percentage of individuals at the border level relevant to each trait measured. In this study, Angoff method was preferred as a test-centered method.

Angoff method

Each judge responds to the items by considering what response levels belonging to the items could distinguish individuals at the threshold level relevant to the measured trait. The sum of the points taken by the person at the threshold level from each item gives the score the person will take from the test (Cizek and Bunch, 2007). The minimum test score of each judge is calculated by collecting the results of each judge for all items in the test. The average of minimum test scores of judges gives the final cut-off score for this method (Livingston and Zieky, 1982).

In student-centred methods, judges perform evaluation according to the students at minimum level about a topic; in other words, by thinking like students. In this study, Border Group method was preferred from student-centred methods.

Border group method

Judges are asked to define individuals as having low and high response rates according to the measured trait. The individuals at the minimum level of high anxiety which includes individuals who are in the middle of the two classifications, that is, those who are able to produce neither low nor high response level, are considered as the border group. The cut-off score for this method is obtained by determining the median of the scale scores of judges (Livingston and Zieky, 1982).

In this study, also a third method, Receiver Operating Characteristics Curve-ROC, was used, which is often preferred for creating border values in fields such as medicine, veterinary medicine, health sciences, and so on.

Methods Based on Border Value Detection

In such methods, the value regarded suitable as a cut-off score is determined with statistical decision theory. Shapiro (1999) divided Border value detection methods into three as Receiver Operating Characteristics Curve (ROC), interval estimation ($\pm 1,96xS$), and standard setting according to percentage values.

Receiver operating characteristics curve (ROC)

ROC analysis allows determining the discriminative power of a test by using the sensitivity and specificity values calculated based on positive-negative ratio values, and comparing its diagnostic power to different tests with the same characteristic (Dirican, 2001). ROC analysis is a method used to detect the performance of diagnostic tests applied in different clinical situations, and to evaluate accuracy of statistical models such as logistic models and linear classification analysis (Obuchowski, 2005). In ROC analysis performed with the help of the ROC curve, a reference is needed to determine the appropriate cut-off score.

In this study, whereas cut-off score detection was performed on the Trait Anxiety Scale, the Test Anxiety Scale developed by Spielberger (1980) and adapted to Turkish by Öner and Le Compte (1983) was accepted as a criterion. Calculations of sensitivity, specificity, positive-negative ratio and positive-negative predictive values required for ROC analysis are displayed in Table 1 (Akçay and Demirel, 2011; Çamlıca and Dişçi, 2008).

Table 1
Values for ROC Analysis

Test Result	Trait in Reality (Criterion Test Result)			Total
	High (Positive) Low (Negative)	High (Positive) A(DP) C(YN) A+C	Low (Negative) B(YP) D(DN) B+D	
				A+B C+D A+B+C+D

(True Positive, DP): Number of students with high levels of anxiety according to the Trait Anxiety Scale but also high levels of anxiety in reality according to the Test Anxiety Scale

B (False Positive, YP): Number of students with high levels of anxiety according to the Trait Anxiety Scale but low levels of anxiety in reality according to the Test Anxiety Scale

C (False Negative, YN): Number of students with low levels of anxiety according to the Trait Anxiety Scale but high levels of anxiety in reality according to the Test Anxiety Scale

D (True Negative, DN): Number of students with low levels of anxiety according to the Trait Anxiety Scale but also low levels of anxiety in reality according to the Test Anxiety Scale

Individuals' anxiety states were classified according to the criterion scale and the scale whose cut-off score values is determined, and numbers of individuals with high and low level of anxiety were calculated. These values are also the coordinates used to determine the sensitivity and specificity ratios required when constructing the receiver operating characteristics curve. The discriminative power of the test can be explained by the ROC curve. The better the discriminative power of the investigated test is, the more upwards the curve shifts (the region of high sensitivity). As the curve approaches the $y = x$ function, the discriminative power of the test decreases and it is thought that the rate of false values is high (Eröz, 2010; Obuchowski, 2005; Sasse, 2002; Uluç, 2007). The ROC curves obtained from three different tests are shown in Figure 1.

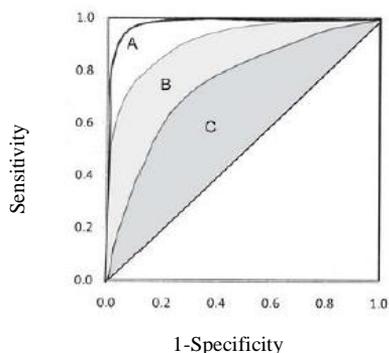


Figure 1. Classification of Different ROC Curves

In Figure 1; the diagonal drawn from (0, 0) point to (1, 1) point at an angle of 45° is called the “chance line” of the test. An ideal ROC curve combines the points (0, 0), (0, 1) and (1, 1). Therefore, the closer the ROC curve is to the point (0, 1), the greater the discriminative power of the test is. The curve A shown in Figure 1 is superior to B and C because A is closer to the point where both sensitivity and specificity are 100%. When the tests are compared, it can be said that the test on which the curve A is obtained is capable of performing better classification than the other tests.

The studies on methods of determining cut-off scores in the field of educational sciences reveals examples comparing several methods in success tests finding out pros and cons of the methods (Çetin, 2011; Gündeğer, 2012; Impara and Plake, 1998; Ömür and Selvi, 2010; Stone, Koskey and Sondergeld, 2011; Tanrıverdi, 2006; Taşdelen, 2009; Taşdemir, 2013; Vivo and France, 2008). Besides, there are studies in which methods of determining cut-off scores and examining the psychometric properties in multiple-scoring psychological tests (Konkan, Şenormancı, Güçlü, Aydın and Sungur, 2013; Konuk, Kıran, Tamam, Kararahmet, Aydın and Atık, 2007; Koyuncu, 2015; Tülübaş, 2009). The review of literature shows that none of the methods for determining the cut-off score is the “best method, the method that can make the most accurate classification” does not appear. Irwin (2007) states that the cut-off scores vary when different cut-off score determination methods are applied to the same exam by the same judges. Cizek (2001) recommends using multiple cut-off score determination methods together and emphasizes that similarity of results will enhance the score obtained. In the research, Angoff and Border Group out of test-centred and student-centred standard setting methods, respectively, were preferred as they are easily understood by judges and easier to apply. As for the ROC analysis, it was chosen because sensitivity and specificity values contribute to validity of the test while determining the cut-off score and it offers values related to correctness of classification of individuals. The present study aims at presenting a different view to test developers and those who are involved in adaptation studies by contributing to interpretation of test scores and validity of the test by means of comparing different cut-off determination methods.

It has been determined that there are no cut-off scores determination studies in the developmental or adaptation phase of the multi-scored psychological tests used in the field of education. Therefore, in the interpretation of the scale scores, it is considered that the accuracy of the decisions taken by using the standard determination methods in the psychological tests other than the average mean by establishing the criterion scores. The research problem of this study arises from the need for using methods of determining cut-off scores and examining the psychometric properties in multiple-scoring psychological tests.

In this study, it was aimed to determine the cut-off score of a scale used to find out the anxiety levels of university students by using Angoff, ROC and Border Group methods. In line with this aim, answer was sought to the following questions.

1. For evaluation of the scores of university students in the Trait Anxiety Scale;
 - a. What is the cut-off score determined with the Angoff method?
 - b. What is the cut-off score determined with the Border Group method?
 - c. What is the cut-off score determined with ROC analysis?

2. What is the accuracy level of the area value under the curve obtained from the Trait Anxiety Scale through ROC analysis in distinguishing students with low/high level of anxiety?
3. What is the validity level of the Trait Anxiety Scale with ROC analysis?

This study is believed to have significance as it contributes to validity of the test as well as fresh information for the field and researchers by examining creation of the criterion point of a multi-scored test with different criteria and examining classification of individuals according to these criteria scores.

Method

This study was carried out with basic research method since it intended to determine the cut-off score of the Trait Anxiety Scale by using the Angoff, ROC and Border Group methods. According to Karasar (2009), the purpose of basic research is to add new knowledge to the existing knowledge. At the same time, our research is aimed at generating knowledge in that we investigate variation of the cut-off score of a multi-scoring scale against different methods in order to decide for the most appropriate score.

Participants

The study was carried out with two groups of participants divided into judges and students. The study group consists of 290 students attending the last grade of different undergraduate programs under Faculty of Education of Recep Tayyip Erdoğan University in Rize province during the 2015-2016 Academic Year Spring Semester. A sufficient number of students with specific qualifications was selected by using criteria sampling as a purposeful sampling method (Büyüköztürk, Çakmak, Akgün, Karadeniz and Demirel, 2011). The distribution of the participants by the areas they are majoring in attending is presented in Table 2.

Table 2
The Distribution of the Participants by the Undergraduate Programs

Department	N	%
Science Teaching	41	14
Elementary Mathematics Teaching	89	31
Classroom Teaching	71	24
Social Studies Teaching	32	11
Turkish Language Teaching	57	20
Total	290	100

The participants included students were attending the fourth grade and they were going to take the 2016 Public Personnel Selection Examination.

The judges who were asked to make evaluation in order to determine the cut-off score were selected from instructors who are well-informed about the characteristics of students with low and high levels of anxiety and who are engaged in guidance and psychological counseling at least five years. The literature provides various views on the optimum number of judges. For example, Jeager (1989) refers to advantages and disadvantages of having a large number of experts in his research, and states that a group of 10 to 15 judges is sufficient. In another study on determining the optimum number of judges for the Angoff method, Gündeğer (2012) found

out that reliability does not change if a larger number of judges are employed. As a result, 10 judges were employed in this study. Table 3 displays the list of the universities where the judges work.

Table 3
Universities of the Judges in the Study Group

University	N	%
Başkent University	2	20
Karadeniz Technical University	4	40
Recep Tayyip Erdoğan University	4	40
Total	10	100

As can be seen in Table 3, of the psychological counseling and guidance instructors engaged in the study; two are employed in Başkent University, four are in Karadeniz Technical University, another four are employed in Recep Tayyip Erdoğan University. The judges, who were responsible for making an evaluation on anxiety state, are the instructors who have already completed or are about to complete to their Ph.D.

Procedure

Study data were collected from the students and the judges during the 2015-2016 Academic Year of Spring Semester. Students' data consist of the scores obtained from the Trait Anxiety and Test Anxiety Scales. As for the other part of the data, they include decisions taken by the judges about the items in the Test Anxiety Scale. Since both scales are related to the same concept and are short, they were applied to the students at once. The decisions given by the judges for determining the cut-off score were obtained by means of a form prepared by the researcher. Before the data were collected, the judges were informed about the methods in detail and the meaning of cut-off score was explained by the researcher so that the judges could develop a common understanding of the minimum level of high anxiety and reach consistency between the scorings. The judges were asked to respond to each scale item according to each of the Angoff and Border Group methods. During the first application, the researcher informed the judges about the student-centred Border Group method. The judges responded to the scales taking into account the responses of a student with "the minimum level of high anxiety". In the next application, the judges were given detailed information about the Angoff method, which is a test-centred approach. So they made decision for each item as to what degree of response distinguishes students at "the minimum level of high anxiety". In summary, the judges were told to consider students in the first application, while they considered the scale items in the second application. In both cases, the judges were required to think and respond in reference to "students at the minimum level of high anxiety." The scale was applied to the same judges at an interval of ten days. The scale scores obtained from the judges' responses in accordance with student and test-centred methods consist of the data collected from judges.

Instruments

In the study, cut-off score was determined with different methods on the Trait Anxiety Scale as a multi-scoring test. Based on the two-factor anxiety theory, [Spielberger et al., \(1970\)](#) developed the State and Trait Anxiety Inventory in order to measure two different states of anxiety. It was adapted to Turkish language by Öner and Le Compte in 1983. The inventory is used to measure anxiety levels of high school students and

normal adults and consists of a total of 40 items. The inventory is composed of two separate four-point Likert type scales with 20 items in each (Özgülven, 2007). The Trait Anxiety Scale was selected from this inventory as it determines the anxiety level of individuals independently on the conditions, they are frequently used in the field of guidance and psychological counseling, and it is short enough for participant students to complete before getting bored.

Trait anxiety scale

The scale was developed to identify how an individual usually feels independent of the surrounding conditions. Individuals respond to scale items according to the emotions they generally feel. Respondents are expected to choose the option which best describes their feelings by following the instructions for each item: (1) rarely, (2) sometimes, (3) often, (4) almost always. It takes about 10 minutes to answer the scale. There are positive and negative items on the scale. In scoring of the scale, two separate keys are prepared for each of the positive and negative item groups. The total of reversely scored weighted score of positive items is subtracted from the total weighted score of negative items. Then, a pre-determined and an unchanging value (35 points) is added to the score. The value obtained lastly is the respondent's trait anxiety score. The total score obtained from the scale ranges from 20 to 80. Higher scores refer to higher level of anxiety.

Validity and reliability studies of the Trait Anxiety Scale were conducted by Öner in 1977. The Cronbach alpha internal consistency coefficient of the scale was found to be .85. Construct validity of the scale was examined experimentally on normal and patient groups within the framework of "two-factor anxiety theory". In significant and stressful cases, the lack of significant changes in the former and latter anxiety scores was thought to support the hypotheses derived from the theory and indicate the construct validity. The results of the replications conducted by Öner and Le Compte (1983) in periods ranging from 10 days to 1 year revealed that individuals do not record meaningful change in trait anxiety scores in different conditions. It was concluded that the scale can differentiate neurotic and normal individuals, is valid in Turkish society, and can be used for research on counseling, psychiatric, and mental health centres Öner (1997). In this study, Cronbach alpha internal consistency coefficient of the scale was found to be .82.

When determining the cut-off score with ROC analysis, another scale should be used to identify students with high/low level of anxiety. In this study, the Test Anxiety Scale was used as criterion scale for ROC analysis for two reasons. It was assumed that the study group was experiencing a high level of anxiety due to the upcoming public personnel selection examination. Also, the scale was short enough for the participants to complete in ten minutes. In addition, Öner (1997) calculated the relationship between the Trait Anxiety Scale scores and the subscale scores of the Test Anxiety Scale. Positive and moderate level of relationship ($r=0.45-0.60$) between affective and delusional subscale scores and Trait Anxiety Scale scores were among reasons for using the scales in this study.

Test anxiety scale

The scale developed by Spielberger (1970) was adapted to Turkish by Öner and Le Compte in 1983. This scale, which determines negative feelings and thoughts about exam and testing, can be applied to all

individuals starting from the primary 4th grade. The four-point Likert-type scale consisting of 20 items is a two-factor scale. Factor one and two were called "affective" and "delusional" subscales, respectively. Individuals respond to the scale items according to the frequency of the emotions they generally feel. It takes about 10 minutes to respond according to degree of anxiety as (1) never, (2) sometimes, (3) often, (4) always. The first item of the scale is scored reverse. The scale provides three types of score as delusional, affective, and total score. Total test score ranges from 20 to 80.

The scores obtained from the scale constitute the raw scores. The standard score equivalents of these scores are determined by looking at the norm table. Individuals with a T-score less than 50 according to the norm table are regarded to have low anxiety. In order to perform ROC analysis in this research, the variable considered as a criterion should be at the level of two-category classification scales. Since the study group is composed of university students, the total score of the Test Anxiety Scale equal to or less than 44 refers to low level of anxiety, the score above 44 refers to high anxiety level according to the norm table.

Öner (1997) found the Cronbach alpha internal consistency coefficient of the four-point Likert-type Test Anxiety Scale as .93, which is regarded considerably high. Factor analysis was performed to determine factor structure of the Test Anxiety Scale and also to check whether it resembles the English version. Analysis of basic components was applied as a factorization method in order to reveal the factor design of the scale, and maximum variability (varimax) method was used as a rotation method. The Turkish scale has two dimensions as it does in English form. As in the original version, the same twelve items were collected under factor one and the remaining eight items under the other factor in the Turkish version. The Cronbach's alpha coefficients for the subscales were calculated as .90 and .70 for factor 1 and 2, respectively. It can be inferred that as the number of items in the subscales becomes smaller, the reliability coefficient becomes smaller, too. In this study, Cronbach alpha internal consistency coefficient of the scale was found to be .89.

Data analysis

In relation to subscale one, the cut-off score of the Trait Anxiety Scale was determined from the judges' scores in the former and latter application by using the Angoff and Border Group methods. During the meetings held by the researcher before the practice, a common point of view was drawn about the individual with "minimum level of anxiety" among the judges. However, it is thought that the judges do not take this into consideration during the scale answers. In the meantime, the information about the methods of determining the cutting points of the judges was formed with this research and the answers were forced to be separated as expected. The judges' scale score averages and medians were calculated to decide the cut-off score value in accordance with the student and test-centred methods above. In order to determine the cut-off score of the scale obtained through ROC analysis, students' scores in the Test Anxiety Scale were used, which is taken as criterion. The score of each student in the study group was classified according to the norm table in the scale manual. The scores below the values in the table were coded as "low anxiety", while those above the values were referred to as "high anxiety", as 1-0. In ROC analysis, the cut-off score of the scale was identified at the best point where sensitivity is high and 1-specificity value is low is determined. The value corresponding to that point was taken as the cut-off score that best distinguishes the anxiety level.

Under another sub-scale, in order to find out the real status of the students who were found to have high or low level of anxiety in the Trait Anxiety Scale, ROC analysis was performed as a border value determination method on the basis of their scores in the Test Anxiety Scale.

The third sub-scale aimed at identifying the criteria-related validity of the Trait Anxiety Scale by using ROC analysis. Positive and negative predictive values were calculated and interpreted as evidence of validity. For these values, the students were first classified against the 2x2 table according to their scores in the Trait Anxiety and Test Anxiety Scales. Test Anxiety scores equal to or below 44 referring to low level of anxiety were coded as “negative”; while values above 44 referring to high levels were coded as “positive”. The other classification was based on the cut-off score of the Trait Anxiety Scale determined with ROC analysis. The 2x2 table contains information on number and percentage of students with high and low levels of anxiety according to the cut-off scores. Positive and negative predictive values as evidence of validity were calculated by means of equations found from the table. Predictive values indicate the probability that the scale with determined cut-off score identify students' anxiety states correctly according to the criterion scale. Positive predictive value refers to the probability that the Trait Anxiety Scale correctly identifies the students with high anxiety level in reality. Negative predictive value determines the probability that the Trait Anxiety Scale correctly identifies the students with low anxiety level in reality.

Results and Discussion

The findings are presented in an order following the research questions.

Determining the cut-off score with angoff, roc and border group methods.

As a response to the first research question, the cut-off score was determined with different methods by looking at the judges' decisions for test-centred and student-centred methods. For the Angoff and Border Group methods, 10 judges were asked to make decisions about items of the Trait Anxiety Scale by following the instructions. The judges were supposed to mark the option that best describes their agreement for each item in accordance with the applicable manual. In the four-point Likert-type scale consisting of twenty items, the minimum and maximum scores are 20 and 80, respectively. The first application was made in order to determine the cut-off score with Border Group method. The judges responded to the scale items considering a student who is "at the minimum level of high anxiety", i.e. assuming their reactions. Knowledgeable about students the judges' anxiety levels are defined as "1 (low)", "2 (limit)", and "3 (high)". The scores judged at the level 2 were accepted as "border group" data. In this group the median of the scores is taken as the cut-off score. In the second application, the same scale was reapplied to the judges with different instructions. They made a decision for each item thinking which response degree would distinguish the students at "at the minimum level of high anxiety". While the judges acted with focus on students during the former application, they focused on response degrees of the scale items during the latter application. Scale scores were prepared from decisions of the judges about student and test-centred methods. The median and average of the judges' scale scores were calculated to determine the cut-off score according to the Angoff and Border Group

methods. The scale scores obtained from the two applications and the cut-off scores determined through these methods are presented in Table 4.

Table 4 shows that the total score of the judges according to the Border Group method ranges from 36 to 50. In the Angoff method, the total score varies between 45 and 51. When the twenty decisions about the minimum level of high anxiety were examined, the lowest score was found in the Border Group method while the highest score was in the Angoff method. In the Angoff method, the cut-off score was determined by taking the average of the judges' scale scores. The evaluation of the Trait Anxiety Scale according to the Angoff method resulted in a final score of 48. In the case of the Border Group Method, the cut-off score was determined by calculating the median of the scale scores of the judges. The cut-off score after the evaluation was determined as 44.5 in this case. The differences in the judges' decisions about student and test-centered methods also led to different scores of minimum levels of high anxiety on the scale.

Table 4
Trait Anxiety Scale Cut-Off Scores Determined with Angoff and Border Group Methods

Judge	Angoff	Border Group
Y-1	45	38
Y-2	47	36
Y-3	48	38
Y-4	50	44
Y-5	45	43
Y-6	49	45
Y-7	49	49
Y-8	51	50
Y-9	47	49
Y-10	49	49
Cut-off score	$\bar{X}=48$	Median=44.50

For determining the cut-off score with ROC analysis, criterion scale data were needed for evaluation of the scores obtained by university students from the Trait Anxiety Scale. The criterion scale scores were taken as a reference for real state of the students falling into low and high anxiety groups while determining the cut-off score of the Trait Anxiety Scale with ROC analysis. The scores of the students on the Test Anxiety Scale indicated their real anxiety situations. In order to be able to perform the ROC analysis, the criterion score must be in the two-category classification scale. Therefore, the scores of the students in the Test Anxiety Scale were converted to the classification scale in reference to the score indicated in the scale norm table. Thus, the scale scores equal to or below 44 in the norm table were coded as "SK low = 0" indicating low level of test anxiety; but the scores above 44 were coded as "SK high = 1" indicating high level of anxiety. The classification and distribution of the scores of the study group in the Test Anxiety Scale according to the groups of anxiety are given in Table 5.

Table 5
Classification of the Study Group Anxiety Levels by Test Anxiety Scale

Study Group	N	%
Positive Group (SK high=1)	91	31.40
Negative Group (SK low=0)	199	68.60
Total	290	100.00

As seen in Table 5, the study group consists of 290 students with positive and negative levels of anxiety. Of these students, 31.4% (N = 91) had a score higher than 44 in the Test Anxiety Scale falling into “high anxiety” group (SK high=1), and the rest of 68.6% (N=199) had a score equal to or below 44 classified with “low anxiety” group (SK low=0).

The Receiver Operating Characteristics Curve-ROC was established to determine the cut-off score of the Trait Anxiety Scale based on individual anxiety scores. Sensitivity and specificity values are used for generating the ROC curve. For this scale, each sensitivity and specificity values yielded by the analysis were examined and the most suitable point was selected before deciding the value to be taken as cut-off score. As a result, the point with high sensitivity but low specificity where the difference between the two values reaches the highest point was selected. The coordinates for the cut-off score determined with ROC analysis are presented in Table 6.

Table 6
Cut-Off Score in Trait Anxiety Scale Determined with ROC Analysis

Cut-Off Score	Sensitivity	Specificity	Cut-Off Score	Sensitivity	Specificity
34.00	1.00	1.00	49.50	.593	.246
35.50	1.00	.995	50.50	.473	.176
37.00	1.00	.990	51.50	.407	.121
38.50	1.00	.925	52.50	.341	.090
39.50	.989	.925	53.50	.242	.060
40.50	.978	.859	54.50	.187	.045
41.50	.956	.819	55.50	.143	.035
42.50	.923	.769	56.50	.099	.015
43.50	.912	.744	58.00	.088	.010
44.50	.890	.658	59.50	.055	.005
45.50	.857	.573	60.50	.033	.005
46.50	.813	.472	62.00	.011	.000
47.50	.725	.402	64.00	.000	.000
48.50	.716	.366			

In Table 6, when the cut-off score is selected as 48.5, it is seen that the corresponding sensitivity and specificity values are 72% and 37%, respectively. This cut-off score corresponds to the point where the difference between the values of sensitivity and specificity (716-366 = 350) is the highest. The reason for determining this value as the cut-off score is that both sensitivity and specificity for the scale refer to the best point at the same time.

In order to evaluate university students' Trait Anxiety scores, the cut-off value was determined as 48 through Angoff method, 44.5 with Border Group method, and 48.5 through ROC. The comparison of the cut-off scores demonstrates that the highest value is obtained with ROC analysis as a border value detection method, while the lowest value is obtained with Border Group as a standard setting method. The different cut-off scores seem to be in parallel with findings of the study by Wang, Pan and Austin (2003) that even the methods in the same class differ considerably within the class. The different cut-off scores determined for the Trait Anxiety Scale can be attributed to different rationales of the decisions given in the three methods. The close values of cut-off scores in Angoff and ROC analysis can be explained with similar focal points of the two methods. In their criticism on standard setting methods, Livingston and Zeiky (1989) asserted that an individual at the lowest level of high anxiety will be imagined differently by each judge. Sensitivity and

specificity values of ROC analysis that do not apply to other methods offer more information about the scale for researchers.

Accuracy level of the area value under the curve obtained from the trait anxiety scale through roc analysis in distinguishing students with low/high level of anxiety.

Receiver Operating Characteristics Curve-ROC was established to determine the cut-off score of the Trait Anxiety Scale based on individual anxiety scores. The ROC curve generated using sensitivity and specificity values is shown in Figure 2.

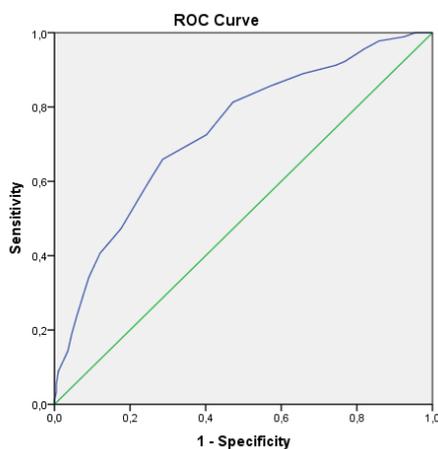


Figure 2. The area under the ROC curve regarding the Trait Anxiety Scale

As an attempt to find out the cut-off value for this scale, each of the sensitivity and specificity values given in Figure 2 were examined and the optimum point (sensitivity = .716, specificity= .366) was selected. The area located under the curve takes values between 0 and 1. The more the curve approaches to the point (0, 1), the bigger the area value is. In his study, Eröz (2010) investigated the variance of the area under the curve in cases of continuous or discontinuous data in the criterion test. It was found out that in case of discontinuous data, the cut-off score of the test led to the highest value of the area remaining under the ROC curve. Therefore, we converted the scores in the Test Anxiety Scale, criterion scale, into discontinuous. The individuals were divided into two with high and low level of anxiety according to the ROC analysis in reference to the area remaining under the curve shown in Figure 2. Table 7 shows the values of that area.

Table 7
The Value of the Area under the ROC Curve

AUC	s	p	95% confidence interval lower limit	95% confidence interval upper limit
.734	.032	.000	.672	.796

It is seen in Table 7 that ROC analysis resulted in a statistically significant operative characteristics curve with the value of (AUC)=0.734 (p<0.05). Hosmer, Lemeshow and Sturdivant (2013), carried out a

classification study on levels of the AUC value and found out that the value closer to 1 would indicate a more accurate classification relying on the cut-off score. In present study, the value under the curve was calculated as 0.734. Thinking the classification based on levels of the AUC value, since the value 0.734 falls into the range of $0.7 \leq AUC < 0.8$, it was concluded that the cut-off score could help classify at an “acceptable” level. This value reveals that based on the cut-off score determined with ROC analysis, the Trait Anxiety Scale is capable of classifying individuals at high and low level of anxiety with 73% accuracy (moderate level). In the ROC analysis performed by *Koyuncu (2015)* for another multi-scoring anxiety scale, the AUC value was found at acceptable level, which seems to support our findings. In concurrence with determination of the cut-off score through ROC analysis, interpreting the area under the curve also provides evidence for accurate classification.

Validity level of the trait anxiety scale through roc analysis.

Assuming that validity determination is an evidence-gathering process, measurement instruments must be re-examined considering the changing conditions. According to *Korkmaz (2007)*, it is essential whether a previously developed measurement instrument measures the psychological traits intended to measure in different cultures, so psychometric qualities of such scale must be examined. In order to answer the third research question, the data collected from the students by means of the criterion scale, supposed to reflect the real situation, and the Trait Anxiety Scale. In order to provide evidence of validity with ROC analysis, the students were classified according to their anxiety state by taking into consideration the Trait Anxiety and Test Anxiety Scales. In order to find out the validity level of the Trait Anxiety Scale through ROC analysis, the data collected by means of the two anxiety scales were analyzed for negative and positive predictive values, sensitivity and specificity values and interpreted accordingly.

The students with scale score equal to or below 44 in the Test Anxiety Scale norm table were coded as "negative" since they show low test anxiety. On the other hand, they were coded as "positive" because values above 44 refer to high anxiety. In the case of Trait Anxiety Scale with ROC analysis, scale scores less than 48.5 were coded as "negative" meaning individuals with low level of anxiety, while scores more than 48.5 were coded as "positive" referring to high level of anxiety. The number and percentage of students with high and low levels of anxiety according to the cut-off scores are given in Table 8.

Table 8
Number and Percentage of Students with Low/High Level of Anxiety in the Scales

Test Anxiety Scale	Trait Anxiety Scale			Total
		Low Anxiety (Negative)	High Anxiety (Positive)	
Low Anxiety (Negative)	N	142	57	199
High Anxiety (Positive)	N	31	60	91
	N	173	117	290

As evidence of validity, positive and negative predictive values were calculated from the values in Table 8 with formulas and interpreted accordingly. Positive predictive value refers to the probability that the Trait Anxiety Scale correctly identifies high anxiety level. Negative predictive value determines the probability that the Trait Anxiety Scale correctly identifies low anxiety level. The cut-off score was calculated as 48.5 based

on the ROC analysis on the Trait Anxiety Scale. Table 8 shows that 117 students in the study group are in the positive group with continuous anxiety because they scored above 48.5, the cut-off score, in the Trait Anxiety Scale. Another 173 students appear in the negative group, which indicates low level of continuous anxiety, as they had scores lower than 48.5 cut-off scores. Of the 117 students in the positive group, 60 are in true positive group, while 57 are in false positive group. As for those in the negative group, 142 fall into true negative, and the remaining 31 in false negative group. Uluç (2007) states that specificity and sensitivity values are not sufficient to determine the actual state of the measured trait, so the information obtained with predictive values must be supported. Predictive values vary depending on the frequency of the measured trait. The frequency of the measured trait was calculated based on the number of individuals with high and low level of anxiety in the study group. Here the positive predictive value was found to be 0.52 ($PYD=60/117$) from ratio of the number of individuals with high anxiety level on both of the scales in Table 8 to the number of individuals with high anxiety level in the Trait Anxiety Scale. It can be understood that the Trait Anxiety Scale predicts high test anxiety with 0.52 accuracy. To put in another way, if the individual's continuous anxiety level is determined as high, the test anxiety is high in reality with 0.52 probability. On the other hand, the negative predictive value was found to be 0.82 ($NYD = 142/173$) from individuals with a lower level of anxiety on both of the scales in Tablo 8 to the number of individuals with low anxiety level in the Trait Anxiety Scale. It indicates that the Trait Anxiety Scale predicts low test anxiety with 0.82 accuracy. In other words, if the individual's continuous anxiety level is determined low, it is actually low with 0.82 probability.

In the study, the negative predictive value of the Trait Anxiety Scale was found to be higher than the positive predictive value, which seems to be in parallel with findings by Konuk *et al.*, (2007) and Konkan *et al.*, (2013), which both reported higher negative predictive values than positive predictive values in scale adaptation studies. Regarding those studies, it seems worth noting that the number of individuals having the measured trait or a high level of the measured trait is smaller than that of individuals who do not have the trait or show a low level of that trait. Similarly, we found a higher number of individuals with low level of anxiety. Hence, unequal numbers of the groups within the study group could explain the high negative predictive values. Uluç (2007) suggests that in cases employing a measurement instrument with higher positive or negative predictive values, the values constitute basis for interpretation vary because the frequency of the trait measured in the study group. It was found out that the Trait Anxiety Scale is more likely to correctly predict low level of test anxiety than high level. Since our study resulted in higher number of individuals with low level of anxiety in both scales, the psychometric characteristics of the scale were examined according to the negative predictive value. Classification of individuals against the 48.5 cut-off score of the scale calculated with ROC analysis yielded the sensitivity value of 72% and specificity value of 37%. Thus, the predictive value of the scale can be said to be high considering this cut-off score. This scale can help identify students at high or low anxiety level according to the cut-off score calculated with ROC analysis. Due to the fact that the scale is used for description rather than diagnosis, the results can be regarded as evidence of validity of the ROC analysis. Taking this cut-off score into consideration, it can be said that guidance can be made easily to identify the factors potentially related to anxiety and handle factors that raise or exacerbate anxiety among university students.

Conclusion

In this study, the cut-off score was determined on a scale measuring anxiety level of university students by using the Angoff, ROC and Border Group methods. Contribution was brought to validity studies of the scale by examining the high/low anxiety classification compliances of these methods.

- i. In line with the study aim; the cut-off values evaluating the scores of university students in the Trait Anxiety Scale were found to be 48 with Angoff method, 44.5 with Border group method, and 48.5 with ROC analysis. Although the methods are item-oriented, student-oriented and criterion-oriented, Angoff and ROC analysis yielded cut-off scores which are quite close. It should be noted that the scale is used for descriptive purposes rather than diagnostic. It was concluded that individuals with scores equal to and above those cut-off scores can be called to have high level of anxiety, implying the need for more detailed interviews with such individuals.
- ii. Investigation was carried out on accuracy of the cut-off score calculated with ROC in classifying anxiety levels of students. The Receiver Operating Characteristics Curve-ROC was developed in order to identify the cut-off score of the Trait Anxiety Scale against the criterion of students' scores in the Test Anxiety Scale. It was concluded that according to the AUC (area under the curve) obtained in the study, the Trait Anxiety Inventory is capable of distinguishing individuals by level of anxiety. In other words, it can identify individuals with high and low level of anxiety with 73% accuracy (intermediate level).
- iii. Validity of the Trait Anxiety Scale was checked with ROC analysis. Positive and negative predictive values were calculated from the number of students with low and high level of anxiety in relation to test anxiety and trait anxiety. Since our study resulted in higher number of individuals with low level of anxiety in both scales, the psychometric characteristics of the scale were examined according to the negative predictive value. The scale yielded higher levels of sensitivity, specificity and negative predictive value calculated with ROC analysis. Due to the fact that the scale is used for description rather than diagnosis, it is quite likely to identify anxiety levels.

As a result, different cut-off scores were calculated for a multi-scoring psychological measuring instrument in reference to standard setting and border value detection methods. When ROC analysis was used to determine the cut-off score; availability of values such as sensitivity, specificity and positive/negative predictive values that are not available in the other methods provided more detailed and objective information. In the case of the cut-off score calculated with ROC analysis, the accuracy rate of the decisions can also be calculated. It was demonstrated that psychometric characteristics of scale can be determined from the values calculated with ROC analysis.

Suggestions

- i. In this study, the cut-off score was determined in multiple-scoring psychological tests by using three different methods as Angoff, Border Group and ROC analysis. The Angoff and Border

Group methods were preferred in the study as they are easily understandable for judges and allow much convenience in practice. The ROC analysis was preferred because sensitivity and specificity values contribute to validity of the test and provide values related to accurate classification of individuals during determination of the cut-off score. Similar research can be done by using other methods to determine the cut-off score.

- ii. In this study, the scales were applied to the student group consecutively. For judges' decisions, the scales were applied at an interval of one week in accordance with standard setting methods. In future studies, this time difference can be set as a variable in determining the cut-off score and different methods can be utilized to find out whether there is difference between the scores.
- iii. Providing training on methods for judges who will contribute to the cut-off score determination studies may enable the process to be carried out more controlled if the research is started with sufficient ones in the success tests to be made later. Thus, the differences that can arise from the variation of the reaction rates of the judges can disappear.
- iv. In the context of education, validity studies can be done by using ROC analysis in scale development and adaptation studies as well as in medicine and health fields.

References

- Akçay, A., and Demirel, A. (2011). Determination of Optimal Positivity Threshold of Some Blood Parameters in Dogs with Pyrometer by ROC (Receiver Operating Characteristic) Curve Method. *Erciyes University Journal of Veterinary Faculty*, 8(3), 153-163.
- Anastasi, A. (1988). *Psychological testing*. New York: MacMillan Pub. Co. Inc.
- Büyüköztürk, Ş., Çakmak, E., Akgün, Ö. E., Karadeniz, Ş., ve Demirel, F. (2011). *Bilimsel araştırma yöntemleri* [Scientific research methods]. Ankara, Turkey: Pegem Akademi Yayınları.
- Cizek, G. J. and Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage Publications.
- Cizek, G. J. (2001). *Setting performance standards: Concepts, methods and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Crocker L., and Algina J. (1986). *Introduction to classical and modern test theory*. Orlando, Florida: Holt, Rinehart and Winston Inc.
- Cronbach, L. J. (1984). *Essentials of psychological testing*. New York: Harper Collins Publishers.
- Çetin, S. (2011). *İşaretleme ve Angoff standart belirleme yöntemlerinin karşılaştırılması* [A Comparison of Bookmark and Angoff Standard Setting Methods]. (Doctoral dissertation, Hacettepe University, Ankara, Turkey). Retrieved from <https://tez.yok.gov.tr/UlusalTezMerkezi/>
- Dirican, A. (2001). Tam testi performansının değerlendirilmesi ve kıyaslanması. [Evaluation of the diagnostic test's performance and their comparisons]. *Cerrahpaşa Tıp Dergisi*, 32(1), 25-30.

- Erkuş, A. (2007). Ölçek geliştirme ve uyarlama çalışmalarında karşılaşılan sorunlar [Scale development and adaptation problems]. *Türk Psikoloji Bülteni*, 13(40), 17-25.
- Eröz, B. (2010). *Veri yapısına bağlı olarak ROC eğrisi altında kalan alana ilişkin istatistiksel yöntemlerin karşılaştırılması* [Comparison of the statistical methods for the area under the ROC curve depending on the data structure]. (Master's thesis, Hacettepe University, Ankara, Turkey). Retrieved from <https://tez.yok.gov.tr/UlusalTezMerkezi/>
- Guilford, J. P. (1954). *Psychometrics methods*. New York: McGraw-Hill Book Co.
- Gündeğer, C. (2012). *Angoff, Yes/No ve Ebel standart belirleme yöntemlerinin karşılaştırılması* [A Comparison of Angoff, Yes/No and Ebel Standard Setting Methods]. (Master's thesis, Hacettepe University, Ankara, Turkey). Retrieved from <https://tez.yok.gov.tr/UlusalTezMerkezi/>
- Hosmer, D. W., Lemeshow, S., and X. Sturdivant, R. (2013). *Applied logistic regression*. New York: John Wiley and Sons.
- Impara, J. C. and Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, 35(1), 69-81. <http://dx.doi.org/10.1111/j.1745-3984.1998.tb00528.x>
- Irwin, P. (2007). An alternative examinee-centered standard setting strategy. University of Nebraska, USA.
- Jeager, R. M. (1989). *Certification of student competence*: In R.L. Linn (Ed) Educational Measurement.(pp. 485-514). New York: Macmillan.
- Karasar, N. (2009). *Bilimsel araştırma yöntemi*. [Scientific research methods] Ankara, Turkey: Nobel Yayınları.
- Konkan, R., Senormanci, O., Guclu, O., Aydın, E., and Sungur, M. Z. (2013). Anxiety Disorder-7 (GAD-7) scale / Common Anxiety Disorder-7 (YAB-7) test. *Archives of Neuropsychiatry*, 50 (1), 53-59.
- Konuk, N., Kiran, S., Tamam, L., Karaahmet, E., Aydın, H., ve Atık, L. (2007). Duygu durum bozuklukları ölçeğinin Türkçe uyarlamasının bipolar bozukluk taramasında geçerliği [Validate the Turkish version of the Mood Disorder Questionnaire (MDQ) as a screening tool]. *Türk Psikiyatri Dergisi*, 18(2), 147-154.
- Korkmaz, M. (2007). Psikolojik ölçmenin yeni kuralları ve Türkiye'deki durumu. [New rules of psychological measurement and the situation in Turkey]. *Türk Psikoloji Bülteni*, 13(40), 8-14.
- Koyuncu, S. (2015) *Psikolojik ölçeklerde ROC analizi yöntemiyle standart belirleme* [Standard determination in psychological scales using roc analysis] (Master's thesis, Gazi University, Ankara, Turkey). Retrieved from <https://tez.yok.gov.tr/UlusalTezMerkezi/>
- Livingston, S. A., and Zieky M. J. (1989). A comparative study of standard setting methods. *Applied Measurement in Education*, 2(2),121-141. http://dx.doi.org/10.1207/s15324818ame0202_3
- Nunnally, J. C. (1970). *Introduction to psychological measurement*. New York: McGraw-Hill Book Co.
- Obuchowski, N. A. (2005). ROC analysis. *American Journal of Roentgenology*, 184(2), 364- 372. <http://dx.doi.org/10.2214/ajr.184.2.01840364>
- Ömür, S., ve Selvi, H. (2010). Angoff, Ebel ve Nedelsky yöntemleriyle belirlenen kesme puanlarının sınıflama tutarlılıklarının karşılaştırılması [Comparison of classification consistency of cut points determined by Angoff, Ebel and Nedelsky methods]. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 1(2), 109-113.

- Öner, N. ve A. Le Compte. (1983). *Durumluk-sürekli kaygı envanteri el kitabı*. [State-trait anxiety inventory handbook] İstanbul,Turkey: Boğaziçi Üniversitesi Yayınları.
- Özgülven, E. İ. (2007). *Psikolojik testler* [Psychological tests]. Ankara, Turkey: PDREM Yayınları.
- Sasse, E. A. (2002). Objective evaluation of data in screening for disease. *Clinica Chimica Acta*, 315(1), 17-30. [http://dx.doi.org/10.1016/S0009-8981\(01\)00710-0](http://dx.doi.org/10.1016/S0009-8981(01)00710-0)
- Spielberger, C. D., Gorsuch, R. C., and R. E. Luschene. (1970). *Manual for the state-trait anxiety inventory*. California: Consulting Psychologists Press.
- Stone, G. E., Koskey, K. L. K., and Sondergeld, T. A. (2011). Comparing construct definition in the Angoff and objective standard setting models: playing in a house of cards without a full deck. *Educational and Psychological Measurement*, 71(6), 942-962. <http://dx.doi.org/10.1177/0013164410394338>
- Tanrıverdi, S. (2006). *Standart belirleme yöntemlerinin geçme puanları üzerindeki etkisi* [The effect of standard setting methods on passing scores]. (Master's thesis, Hacettepe University, Ankara, Turkey). Retrieved from <https://tez.yok.gov.tr/UlusalTezMerkezi/>
- Taşdelen, G. (2009). *Nedelsky ve Angoff standart belirleme yöntemlerinin Genelenebilirlik kuramı ile karşılaştırılmasına ilişkin bir araştırma* [A comparison of Angoff and Nedelsky cutting score procedures using generalizability theory]. (Master's thesis, Hacettepe University, Ankara, Turkey). Retrieved from <https://tez.yok.gov.tr/UlusalTezMerkezi/>
- Taşdemir, F. (2013). *Angoff (1-0), Nedelsky ve sınır değerleri saptama yöntemleri ile bir testin sınıflama doğruluklarının incelenmesi Angoff (1-0)*, [Nedelsky and examination of classification accuracies of a test by determination methods of limit values]. (Doctoral dissertation, Ankara University, Ankara, Turkey). Retrieved from <https://tez.yok.gov.tr/UlusalTezMerkezi/>
- Tülübaş, G. (2009). *Psikolojik testlerde Angoff ve Sınır grup yöntemleri ile kesme puanlarının belirlenmesi Psikolojik testlerde Angoff ve Sınır Grup Yöntemleri ile kesme puanlarının belirlenmesi* [To determine cut off points with Angoff and Borderline Group Techniques in psychological testing]. (Master's thesis, Hacettepe University, Ankara, Turkey). Retrieved from <https://tez.yok.gov.tr/UlusalTezMerkezi/>
- Uluç, S. (2007). ROC Analizi I: Özgüllük, duyarlılık, pozitif yordayıcı değer ve negatif yordayıcı değer hesaplamaları [ROC analysis I: specificity, sensitivity, positive predictive value and negative predictive value]. *Türk Psikoloji Bülteni*, 13(40), 26-31.
- Vivo, J. M., and Franco, M. (2008). How does one assess the accuracy academic success predictors? ROC analysis applied to university entrance factors. *International Journal of Mathematical Education in Science and Technology*, 39(3), 325-340. <http://dx.doi.org/10.1080/00207390701691566>
- Wang, L., Pan, W., and Austin, J. T. (2003). Standards-setting procedures in accountability research: Impacts of conceptual frameworks and mapping procedures on passing rates. *Paper presented at the annual meeting of the American Educational Research Association*, Chicago.