Eurasian Journal of Educational Research 90 (2020) 121-138

Eurasian Journal of Educational Research
www.ejer.com.tr

A Comparison of Two Standard-Setting Methods for Tests Consisting of Constructed-Response Items*

Hatun Betul OZARKAN[1], Celal Deha DOGAN[2]

ARTICLE INFO

Article History:
Received: 8 Oct. 2019
Received in revised form: 3 Sept. 2020
Accepted: 17 Sept. 2020
DOI: 10.14689/ejer.2020.90.7

Keywords
standard setting, cut score, constructed response item, extended Angoff method, contrasting groups method

ABSTRACT

Purpose: This study aimed to compare the cut scores obtained by the Extended Angoff and Contrasting Groups methods for an achievement test consisting of constructed-response items.

Research Methods: This study was based on survey research design. In the collection of data, the study group of the research consisted of eight mathematics teachers for the Extended Angoff method, 75 eighth grade students, and a mathematics teacher for the Contrasting Groups method. Data were collected through math achievement test consisting of constructed-response items, scoring rubrics, expert opinion form for Extended Angoff method and student classification form for Contrasting Groups method.

Findings: Cut score was determined to be 13,38 by the Extended Angoff method and 12,50 by the Contrasting Groups method. It has been determined that the two standard-setting methods do not make significant difference between the master accepted student ratios. It has been found that there is a high level of harmony between the methods in terms of classifying the students as qualified and unqualified.

Implications for Research and Practice: This study was restricted to test consisting of constructed-response items and two standard-setting methods. It is suggested that future research tests should include different types of items and compare different methods. In the Covid-19 pandemic process, judge panel discussions can be made online in the Extended Angoff method. Online meetings can allow the creation of a heterogeneous judge group. Therefore, both methods can be used in the Covid-19 process.

* This article was derived from the first author's a master's dissertation conducted under the supervision of the second author.

[1] Corresponding Author: Ministry of National Education, TURKEY, e-mail: hbozarkan@yahoo.com, ORCID: https://orcid.org/0000-0001-5860-5991

[2] Ankara University, Educational Science Faculty, TURKEY, e-mail: dehadogan@gmail.com, ORCID: https://orcid.org/0000-0003-0683-1334

Introduction

Education systems consist of input, process, output and control components. Evaluation is the control mechanism of the interrelated parts of this system. Evaluation not only controls outputs or processes, but also enables the system to be repaired and improved, and the results of the evaluation can be an input into the system becoming an element that prevents the system from entering a vicious circle and ensuring that it is dynamic.

Evaluation is the comparison of the results of the measurement with a criterion or a set of criteria, and then making a decision (Baykul, 2000). The decision, which is a result of evaluation, depends on the measurement results, and in the sense of accuracy, reflects the real value of the variable (measured property). It also depends on whether the criterion is appropriate for the purpose of evaluation, and the accuracy of the procedures during the comparison (Baykul, 1992). The minimum error of measurement is important in terms of reflecting the actual value of the measured property. Although the accuracy of a decision is directly related to the validity and reliability of the measurement results, it does not necessarily guarantee it. The eligibility of the criterion is also related to the accuracy of the decision to be taken. Criteria constitute the decision-making framework used to reach a decision as a result of the measurement, and they play an important role in standardizing the decision that is made.

In an education system, two types of criteria are used. The first is an absolute criterion, which is the same for all individuals involved in the measurement process. It does not change according to a person or group and it corresponds to one or more cut points. The second is a relative criterion based on the results obtained from the group subject to measurement. The value corresponding to the cut point(s) in a relative criterion can vary from group to group (Tekin, 2017; Turgut & Baykul, 2014).

The cut points determined by relative or absolute criteria correspond to scores on a test scale. The cut score divides the scale of the test score into two or more categories or classifications of the examinees taking the test (Cizek & Bunch, 2007). The appropriate adoption of a prescribed, rational system of rules or procedural system, which results in the assignment of a number to differentiate between two or more states or degrees of performance, is called standard-setting (Cizek, 1993). This process is not arbitrary; it is clearly defined and systematic. A standard is the conceptual version of the desired level of competence, and the cut score (or passing score) is the operational version (Kane, 1994). It refers to the minimum level of knowledge and skills required for relevant performance categories. Therefore, a standard is the answer to the question, “How much is enough?”, and in this respect, it can be considered as a criterion.

Jaeger (1989) classified standard-setting methods as “test-centered” and “examinee-centered”. In test-centered methods, judges make decisions about test items, while in examinee-centered methods, they make decisions about the examinees. In this study, Extended Angoff method, which is test-centered, and the Contrasting Groups method, which is examinee-centered, were compared.

Extended Angoff Standard-Setting Method

The Angoff method is a test-centered standard-setting method that is suited for tests consisting of multiple-choice items. In the Angoff method, judges estimate the probability that the borderline examinees would respond to each multiple-choice item correctly in the test. The Extended Angoff method is a test-centered standard-setting method and an adaptation of the Angoff method for constructed-response items (Hambleton and Plake, 1995). In the Extended Angoff method, judges estimate the number of scale points that they believe the borderline examinees will obtain on each constructed-response item in the test (Cizek & Bunch, 2007). Extended Angoff method can be used in combination with the traditional Angoff method in mixed-format tests that contain both multiple-choice and constructed-response items. Here, the borderline refers to performance that is between an acceptable and unacceptable level, and a person who just barely passes.

Contrasting Groups Standard-Setting Method

In the Contrasting Groups method, the cut score is determined based on the test scores of the examinees. Therefore, this method is an examinee-centered standard-setting method. This method is based on the idea that examinees can be divided into two contrasting groups (qualified-unqualified or pass-fail) on the basis of the judgments of their knowledge and skills (Livingston & Zieky, 1982). It is important that the judge, who will decide which examinees will be in two contrast groups, knows the examinees taking the test in terms of their knowledge and skills. The judges place the examinees into two groups without knowing their scores from the test. The category judgments about the examinees are used to form distributions of total test scores for each of the two groups (Cizek & Bunch, 2007). Then, two distributions are plotted and analyzed to arrive at the cut point that separates the groups.

In standard-setting research in Turkey, mostly test-centered standard-setting methods have been compared (Cetin & Gelbal 2010; Demir & Kose, 2014; Gundeger & Dogan, 2014; Korkmaz, 2015; Omur & Selvi, 2010; Tasdelen, Kelecioglu & Guler 2010). These studies were conducted only on tests involving multiple-choice items. No standard-setting research has been found for tests involving different item formats. However, constructed-response items have also been used in recent years in international research (Programme for International Student Assessment, Trends in International Mathematics and Science Study, etc.) along with monitoring and evaluation research (Monitoring and Evaluation of Academic Skills, Student Achievement Monitoring Research etc.) conducted in Turkey. In addition, attempts have been made regarding the inclusion of multiple-choice items as well as constructed-response items in the national exams held in Turkey. Discussions continue on this issue. However, in Turkey, no standard-setting research for a test involving constructed-response items has been found which increases the need for standard-setting studies on this type of test.

In the literature, there is limited research comparing examinee-centered and test-centered methods (Tulubas, 2009) and a similarly lack of research comparing the

Extended Angoff method applied to test-centered methods and the Contrasting Groups method concerning examinee-centered methods. (Konge et al., 2012). Based on these reasons, a comparison of the cut scores obtained from the standard-setting methods of Extended Angoff and Contrasting Groups for an achievement test involving constructed-response items constitutes the main problem of this study.

Aim of the Research

The aim of this study was to compare the cut scores obtained by the Extended Angoff and Contrasting Groups methods that separate the qualified-unqualified levels for an achievement test consisting constructed-response items and designed for the algebra topic included in the eighth-grade mathematics course. According to the purpose of the study, answers to the following questions were sought:

For this type of achievement test,

1. what is the cut score determined by the Extended Angoff method?
2. what is the cut score determined by the Contrasting Groups method?
3. do the classifications (qualified-unqualified) made for examinees differ according to the cut scores determined by the Extended Angoff and Contrasting Groups methods?
    a. is there a significant difference between the percentage of examinees that were considered to be qualified due to scoring above the cut scores obtained from the two different standard-setting methods?
    b. is there consistency between the two methods in classifying examinees as qualified or unqualified?

Method

Research Design

In this study, Extended Angoff and Contrasting Groups standard-setting methods were compared. Accordingly, the current study was based on survey research design since it aimed to describe a situation as it was (Buyukozturk, Kilic Cakmak, Akgun, Karadeniz & Demirel, 2013).

Study Group

For the Extended Angoff method, a judges group consisting of teachers was selected by criterion sampling, one of the purposeful sampling methods. The criterion was accepted as the teachers having been teaching mathematics for at least two years in the eighth grade. Participation was voluntary. A total of eight teachers were included in the group considering the possible number of math teachers that could be found in a school. For the Contrasting Groups method, deciding on the number of students generally requires a balance between costs and benefits (Livingston & Zieky, 1982). It was considered that it would not be appropriate for a teacher to decide on students that they did not actually teach in the school; therefore, 75 students took

part in the research. The judge who would classify the students as contrasting groups was the teacher of mathematics who had taught for at least one year.

Research Instruments and Procedures

Mathematics is one of the main areas of national research as well as large-scale research such as PISA and TIMSS. In addition, because of its nature, mathematics is suitable to be examined with constructed-response items, as it includes high-order thinking processes such as reasoning, problem solving, and organizing problems. That's why in this study, the mathematics achievement test, developed by the researcher, was used. The test prepared for the algebra topic included in the eighth-grade mathematics course contained eight constructed-response items. As a result of the application of this test to 75 students, the reliability coefficient Cronbach α value of internal consistency was calculated as 0.91. The item difficulty index had values between 0.22 and 0.43. The item discrimination index values were above 0.40. According to item difficulty index and item discrimination index values, items are difficult and highly discriminate examinees in terms of the traits measured in the achievement test. Based on these findings, it can be stated that the reliability of the test was high.

A holistic scoring rubric was created for each item in the developed mathematics achievement test. Expert opinion was consulted in the process of developing the scoring rubric, after which the test was finalized. Critical steps of the performance of the examinees for the solution of each item were accepted as scoring criteria. Therefore, each item has been scored in different degrees. The scoring rubrics for the items in the mathematics achievement test were as follows:

- Items 4 and 6: 0-3 points (0, 1, 2, 3)
- Items 1, 2, 3, 7, and 8: 0-4 points (0, 1, 2, 3, 4)
- Item 5: 0-5 points (0, 1, 2, 3, 4, 5)

Kappa statistics were used to ensure the consistency between the scores given by the judges. The kappa values calculated for the items varied between 0.925 and 1, indicating statistically significant ($p < .001$). The consistency between scoring was very high. This finding proved the validity of the scoring rubrics.

In the Extended Angoff method, the judges decided on how many points an examinee on the border of qualified-unqualified would be given, taking into account the criterion in the scoring rubrics for each item in the test. In this process, the judges used the expert opinion form developed by the researcher.

The examinees in the Contrasting Groups method were classified into two groups by the mathematics teachers. For this purpose, a classification form including the names, surnames and grade information of the examinees was used.

The data were collected from the judges who implemented the methods of Extended Angoff and Contrasting Groups and from the students who took the mathematics achievement test. For the Extended Angoff method, the data were collected by considering the steps listed by Hambleton and Plake (1995), while for the Contrasting Groups method, the data were collected by following the steps listed by Brandon (2002).

Data Analysis

In the Extended Angoff method, the judges estimated how many points the students that were on the border of qualified-unqualified for each item could score and recorded their estimation. The arithmetic mean of the points given by the judges was calculated for each item. The sum of these arithmetic means gave the cut score.

In the Contrasting Groups method, the cut score was obtained by calculating the midpoint of the median of the distribution of the test scores obtained from the students who were classified as qualified or unqualified according to the expert judgment. In addition, the cut score was calculated by the logistic regression method. Logistic regression was used to determine the raw score point where the probability of qualified-unqualified category membership was 0.50.

$$y^{*}=a+b(x) \tag{1}$$

where a is a constant, b is the slope of the regression function, and x is the raw score midway between two possible classifications (qualified-unqualified), and y is the predicted value which shows the category (qualified-unqualified) in which an examinee will be located. Since qualified and unqualified were coded as 0 and 1, respectively, .50 was used as the value that best differentiated between these two categories (Cizek & Bunch, 2007).

The difference between the rates of students considered as qualified due to scoring above the cut scores according to the standard-setting methods of Extended Angoff and Contrasting Groups was determined by the difference test between the two dependent ratios. The significance of the difference was determined by the z statistic (Akhun, 1982).

$$z=\frac{b-c}{\sqrt{b+c}} \tag{2}$$

b: Number of students who succeeded in method 1 but failed in method 2

c: Number of students who succeeded in method 2 but failed in method 1

Cohen's Kappa coefficient was used to investigate the consistency of the Extended Angoff and Contrasting Groups standard-seting methods in terms of the classification of the students as qualified-unqualified.

Results

Results Related to the Cut Score Obtained by the Extended Angoff Method

Eight judges were consulted to determine the cut score with the Extended Angoff method in the eight-item mathematics achievement test. The cut score was determined through the scoring rubrics. While determining the cut score, the judges scored two rounds considering the borderline examinee. These scores are shown in Table 1.

Table 1

Judges' Scoring Obtained by the Extended Angoff Method

Judge		Item Number									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	$\bar{X}$	**sd**
1	Round 1	2	1	3	0	1	1	0	0	1.00	1.07
	Round 2	2	2	3	1	1	1	0	0	1.25	1.04
2	Round 1	2	2	3	0	3	3	1	0	1.75	1.28
	Round 2	2	2	4	0	3	3	0	0	1.75	1.58
3	Round 1	2	2	4	0	3	3	0	4	2.25	1.58
	Round 2	2	2	4	0	3	3	0	0	1.75	1.58
4	Round 1	2	2	4	0	3	3	0	4	2.25	1.58
	Round 2	2	2	4	0	3	3	0	0	1.75	1.58
5	Round 1	2	2	2	1	3	3	1	0	1.75	1.04
	Round 2	2	2	2	1	3	3	1	0	1.75	1.04
6	Round 1	4	2	4	1	0	0	0	1	1.50	1.69
	Round 2	4	2	4	0	0	3	0	1	1.75	1.75
7	Round 1	2	2	2	0	3	1	1	1	1.50	0.93
	Round 2	2	2	2	0	3	1	1	1	1.50	0.93
8	Round 1	2	2	4	1	3	3	0	1	2.00	1.31
	Round 2	2	2	4	1	3	3	0	0	1.88	1.46
	Round 1	2.25	1.88	3.25	0.38	2.38	2.13	0.38	1.38	1.75	0.42
$\bar{X}$		0.71	0.35	0.89	0.52	1.19	1.25	0.52	1.69		
sd	Round 2	2.25	2.00	3.38	0.38	2.38	2.50	0.25	0.25	1.67	0.20
		0.71	0.00	0.92	0.52	1.19	0.93	0.46	0.46		

When the cut scores for each item at the end of the second round were analyzed, the lowest cut score was determined for the seventh and eighth items (0.25), and the highest cut score for the third item (3.38) (Table 1). In addition, the judges determined that the students that were on the qualified and unqualified border should have had an average score of 1.67 from each item. The standard deviation value indicated the variability between the scoring of the experts. The variability between expert opinions was less in the second round (.20) than in the first round (.42).

At the end of the second round, the arithmetic mean of the scores determined by each judge for each item was taken. The cut score was calculated by summing all these values.

$$Cut\ score = 2.25 + 2.0 + 3.38 + 0.38 + 2.38 + 2.50 + 0.25 + 0.25$$
$$Cut\ score = 13.39$$

According to the results obtained by the Extended Angoff method, an eighth grader was expected to obtain a minimum score of 13.39 out of 31 in order to be considered 'qualified' in terms of the traits measured in the achievement test.

Results Related to the Cut Scores Obtained by the Contrasting Groups Method

A judge, who taught mathematics, was consulted for the cut scores obtained with the Contrasting Groups method. This judge classified 75 students into two groups as being qualified and unqualified according to their respective definitions (before the test was administered) in the Extended Angoff method.

Table 2 presents the descriptive statistics related to the scores of the students who were classified as qualified-unqualified based on the results of the achievement test administered to 75 students, and Figure 1 shows the related distribution.

Table 2

Descriptive Statistics Related to the Scores of the Students Who Were Classified as Qualified-Unqualified Based on the Judge's Evaluation

Student Group	N	$\bar{X}$	sd	Median	Min score	Max score
Qualified	44	3.09	3.30	3	0	13
Unqualified	31	20.81	7.42	22	3	31
Both Groups	75	10.41	10.28	6	0	31

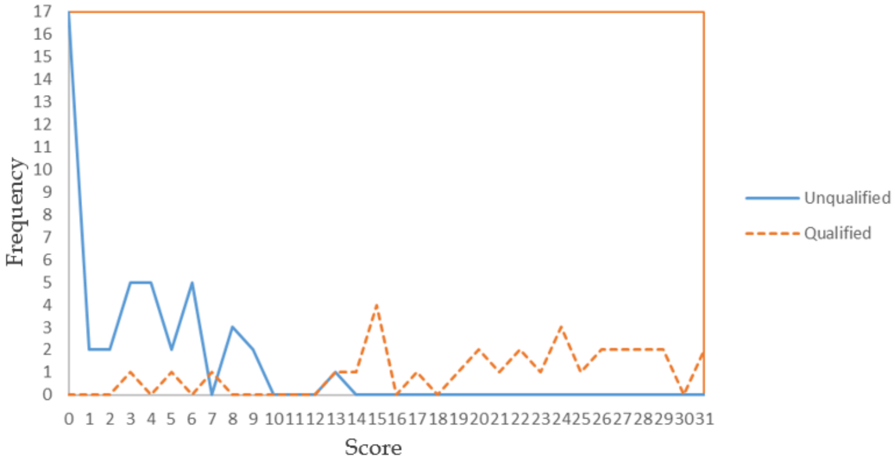

Figure 1. *Score Distribution of the Students Classified as Qualified-Unqualified*

Table 2 shows that the mean of the student group ($\bar{X}$) was 10.41 and the standard deviation was 10.28. In this case, the coefficient of variation (V) for student distribution was calculated as 99%. Since the coefficient of variation (V = 99%) was greater than 50%, which was accepted as the criterion, it can be interpreted that the student group was heterogeneous in terms of the measured traits (Saracbası, Karaagaoglu & Saka, 1986). It is seen that the range of students' scores was 31. As is evident in Figure 1, the cut point of student score distributions was 13. However, it is not always easy to find the intersection of the two distributions on the graph (Cizek & Bunch, 2007). For these reasons, a statistical approach was followed in determining the cut score by the Contrasting Groups method, and the median values were utilized. The cut score was found by calculating the midpoint between the two medians of the groups. As shown in Table 2, the median of the unqualified group was 3, and the median of the qualified group was 22. Accordingly, the cut score obtained by the Contrasting Groups method was determined as 12.50.

According to the results obtained by calculating the midpoint of the median values, an eighth grader was expected to obtain a minimum score of 12.50 out of 31 in order to be considered 'qualified' in terms of the traits measured in the achievement test developed.

One of the ways to calculate the cut score in the Contrasting Groups method is logistic regression. The results of the logistic regression analysis with the data obtained from the Contrasting Groups method are given in Table 3.

Table 3

Logistic Regression Analysis Results for the Contrasting Groups Method

	B	S.E	Wald	sd	p	Exp(B)
Raw score	.440	.110	15.92	1	.000	1.55
Constant	-4.59	1.07	18.43	1	.000	.010

The regression equation according to the results obtained from Table 3 is as follows:

$$0.50 = -4.59 + 0.440.(x)$$

The raw cut score, which is the x value obtained from the equation for the value of y = 0.50, was 11.57. This value was lower than the cut score (12.50), which was obtained from the midpoint of the median values. Considering the size of the sample and the distribution of the scores, it is not particularly unusual that different cut scores results were obtained via different methods (Cizek & Bunch, 2007). There was no significant difference between the cut scores obtained from both methods in terms of classifying students as qualified and unqualified. However, since the use of logistic regression in small samples often yields large standard errors for model parameters and small values for R-squared, it is preferable in many situations to use the midpoint between the median values or the midpoint between the mean values (Cizek & Bunch, 2007).

According to the results obtained by logistic regression analysis, an eighth grader was expected to achieve a minimum score of 11.57 out of 31 to be considered 'qualified' in terms of the traits measured in the achievement test developed.

Results related to the classification of examinees based on the cut scores

In this section, the results and interpretations related to the significance of the difference in terms of the percentage of examinees that were considered qualified as well as those related to the consistency between the classifications according to the two methods are given.

Table 4 shows the examinee frequencies classified as qualified and unqualified according to the cut scores obtained by the Extended Angoff and Contrasting Groups methods.

Table 4

The Examinee Frequencies Classified as Qualified and Unqualified According to the Cut Scores Obtained

	Cut Score	**Number of *Qualified* Examinees**	**Number of *Unqualified* Examinees**
Extended Angoff	13.38	27	48
Contrasting Groups (median)	12.50	29	46
Contrasting Groups (logistic regression)	11.57	29	46

Table 4 reveals that the highest cut score (13.38) was obtained by the Extended Angoff method and the lowest cut score (11.57) was obtained by the Contrasting Group method using logistic regression. Twenty-seven students were accepted as qualified when the students were classified according to the cut score (13.38) obtained by the Extended Angoff method. In this case, the percentage of qualified examinees was 0.36. In the Contrasting Groups method, the number of examinees that were considered qualified according to the cut score calculated in both methods was equal, and it was 29. In this case, the percentage of qualified examinees was 0.39. Although the number of qualified examinees and their proportions was close in both methods, there was a difference between them. The significance of the difference between these two percentages was examined based on the difference between the two dependent percentages. Both cut score calculation techniques for the Contrasting Groups method yielded the same percentage for the examinees that were considered as qualified; therefore, only one of the calculations was used. Table 5 shows the cross-table of the examinees classified as qualified and unqualified according to the cut scores obtained.

Table 5

The Cross Table of Examinees Classified as Qualified and Unqualified According to the Cut Scores Obtained

		Contrasting Groups Method		
		Qualified	**Unqualified**	**Total**
	Qualified	27	0	27
Extended Angoff Method	**Unqualified**	2	46	48
	Total	29	46	75

As shown in Table 5, the number of examinees that were classified as qualified by the Extended Angoff method and unqualified by the Contrasting Groups method was 0. On the other hand, the number of examinees classified as qualified by the Contrasting Groups method but unqualified using the Extended Angoff method was 2. Examining the significance of the difference between the percentages of examinees that were considered to be qualified according to the two methods, the value of z was calculated as follows:

$$z = \frac{|0-2|}{\sqrt{(0+2)}}$$

$$z = 1$$

The z value was not significant at a level of .01 (z < 2.58) indicating that there was no significant difference between the percentages of examinees classified as qualified according to the cut scores obtained by the Extended Angoff and Contrasting Groups methods.

Cohen's Kappa (K) statistic was used when examining the consistency between the Extended Angoff and Contrasting Groups methods. The examinees that scored below the cut score were coded as 0 (unqualified) and those scoring above the cut score were coded as 1 (qualified). Thus, the student scores were categorized. As a result, K was calculated as 0.943 (p < .001). Accordingly, it can be stated that the level of consistency between the Extended Angoff and Contrasting Groups Methods was high in terms of classifying the students as qualified and unqualified.

Discussion, Conclusion and Recommendations

In this study, the cut scores obtained from the standard-setting methods of Extended Angoff and Contrasting Groups were compared for an achievement test consisting of constructed-response items. The conclusions that can be derived from the results are summarized below.

There was no significant difference between the percentages of examinees that were considered to be qualified according to the cut scores obtained by the Extended Angoff method (13.38) and the Contrasting Groups method (12.50). In addition, there was a high level of consistency (K = 0.943) between the two methods in classifying the examinees as qualified and unqualified according to their cut scores. From this perspective, it was concluded that the standard-setting technique employed by the two methods did not differ in terms of the cut scores obtained. However, the circumstances arising from the differences in the processes of standard-setting methods should not be overlooked since they may be the result of the structure of the examinee and judgments groups.

In the context of differences arising from the structure of the examinee group, if the group of examinees shows a heterogeneous distribution of the measured property, both methods can be used. However, for groups with homogeneous distribution, the Contrasting Groups method may have limitations for two reasons: the difficulty of differentiating the definitions of examinees from each other in a group of examinees with homogeneous distribution and the difficulty of separating students into two contrasting groups. In this case, the differences observed between the examinee groups classified as qualified and unqualified may not be obvious. According to Hambleton et al. (2000), cut-scores obtained by contrasting group methods are dependent on the representativeness of the sampled examinees. If the examinees have high performance, there is a potential risk that the cut-score would be too high. A representative sample of groups at different performance levels would give more trustworthy cut-scores (Näsström & Nyström, 2008). Therefore, if the student group shows distorted distribution in terms of the measured property, it would be more appropriate to choose the Extended Angoff method instead of the Contrasting Groups method.

In the context of differences arising from the structure of the judge group, group discussion to define performance categories and determine the cut scores in the Extended Angoff method allow interaction between the judges. In this regard, the participation of the judges with a similar experience in standard-setting studies will contribute to the process. The differences between the definition of performance categories made by the teachers of high and low performing groups may present some potential problems (Näsström & Nyström, 2008). According to the results obtained by Livingston and Zieky (1989), teachers of high ability students tend to set higher standards. Therefore, the difference in the performance levels of the students of the judges group will contribute to the validity of the results obtained. Since the cut score is obtained by statistical means in the Contrasting Groups method, the judges to apply this method should have the necessary statistical background.

The following suggestions are presented to researchers who plan to work on similar issues in the future:

A similar study can be carried out in different ways for an achievement test consisting of constructed-response items. A standard-setting study using the same methods can be carried out for typical performance tests consisting of polytomously scored items and the results can be compared. The level of consistency between the two methods can be studied in a more homogeneous group of examinees. The cut scores obtained using the same methods from a different judge group can be compared. In the Extended Angoff method, the judge group can come together with an online meeting. Considering the difficulties of getting together physically in terms of factors such as time, cost and Covid-19 pandemic, an online meeting can be advantageous. Thus, the diversity of the judge group can be increased. Based on this, researchers who want to conduct a standard-setting study in the Covid 19 Pandemic process can use both methods.

References

Akhun, I. (1982). Iki yuzde arasindaki farkin manidarliginin test edilmesi [Testing the significance of the difference between two percentages]. *Ankara University Journal of Educational Sciences, 15*(1), 240-259. doi: 10.1501/Egifak_0000000817

Baykul, Y. (1992). Egitim sisteminde degerlendirme [Evaluation in education system]. *Hacettepe University Journal of Education* (7), 85-94.

Baykul, Y. (2000). *Egitimde ve psikolojide olcme: klasik test teorisi ve uygulamasi* [Measuring in education and psychology: classical test theory and practice.]. Ankara: OSYM Yayinlari.

Brandon, P. R. (2002). Two versions of the contrasting-groups standard-setting method: A review. *Measurement and Evaluation in Counseling and Development, 35*(3), 167-181. doi: 10.1080/07481756.2002.12069061

Buyukozturk, S., Kilic Cakmak, E., Akgun, O. E., Karadeniz, S., & Demirel, F. (2013). *Bilimsel arastirma yontemleri* [Scientific research methods] (14 b.). Ankara: Pegem Akademi.

Cizek, G. J. (1993). Reconsidering Standards and Criteria. *Journal of Educational Measurement, 30*(2), 93-106. doi: 10.1111/j.1745-3984.1993.tb01068.x

Cizek, G. J., & Bunch, M. B. (2007). *Standart Setting: A Guide to Establishing and Evaluating Performance Standards on Tests.* Sage Publications.

Cetin, S., & Gelbal, S. (2010). Farkli standart belirleme yontemlerinin gecme puanlari uzerine etkisi [Impact of standard setting methodologies over passing scores]. *Ankara University Journal of Educational Sciences, 43*(1), 79-95. doi: 10.1501/Egifak_0000001191

Demir, O., & Kose, İ. A. (2014). Angoff, Nedelsky ve Ebel standart belirleme yontemleri ile belirlenen kesme puanlarinin karsilastirilmasi [A comparison of cutting points determined by Angoff, Nedelsky and Ebel standard setting methods]. *Mersin University Journal of the Faculty of Education, 10*(2), 14-27. doi: 10.17860/efd.18119

Gundeger, C., & Dogan, N. (2014). Angoff, Yes/No ve Ebel standart belirleme yontemlerinin karsilastirilmasi [A comparison of Angoff, Yes/No and Ebel standard setting methods]. *Journal of Measurement and Evaluation in Education and Psychology, 5*(1), 53-60. doi: 10.21031/epod.47091

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8*(1), 41-55. doi: 10.1207/s15324818ame0801_4

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied psychological measurement*, 24(4), 355-366. doi: 10.1177/01466210022031804

Jaeger, R. M. (1989). Certification of student competence. R. L. Linn (Ed). *Educational measurement* (s. 485-514). New York: Macmillan.

Kane , M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*(3), 425-651. doi: 10.3102/00346543064003425

Konge, L., Clementsen, P., Larsen, K. R., Arendrup, H., Bunchwald, C., & Ringsted, C. (2012). Establishing pass/fail criteria for Bronchoscopy performance. *Respiration, 83*, 140-146. doi: 10.1159/000332333

Korkmaz, S. (2015). ***Evet/Hayir, Ebel ve Isaretleme standart belirleme yontemlerinin karsilastirilmasi*** [Comparing Yes/No, Ebel and Bookmark standard setting methods]. Retrieved from http://hdl.handle.net/11655/1849

Livingston, S. A., & Zieky, M. J. (1982). *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests.* Princeton, NJ: Educational Testing Service. Retrieved from https://files.eric.ed.gov/fulltext/ED227113.pdf

Livingston, S. A., & Zieky, M. J. (1989). A comparative study of standard-setting methods. *Applied Measurement in Education*, 2(2), 121-141. doi: 10.1207/s15324818ame0202_3

Näsström, G., & Nyström, P. (2008). A comparison of two different methods for setting performance standards for a test with constructed-response items. *Practical Assessment, Research, and Evaluation*, *13*(1), 9. doi: 10.7275/bhb9-8t88

Omur, S., & Selvi, H. (2010). Angoff, Ebel ve Nedelsky yontemleriyle belirlenen kesme puanlarinin siniflama tutarliliklarinin karsilastirilmasi [Comparison of classification consistency of cutting scores obtained from Angoff, Ebel and Nedelsky methods]. *Journal of Measurement and Evaluation in Education and Psychology, 1*(2), 109-113. Retrieved from https://dergipark.org.tr/tr/pub/epod/issue/5807/77246

Saracbasi, O., Karaagaoglu, E., & Saka, O. (1986). *Basic programlama ve istatistiksel yontemler* [Basic programming and statistical methods]. Ankara: Unalan Ofset.

Tasdelen, G., Kelecioglu, H., & Guler, N. (2010). Nedelsky ve Angoff standart belirleme yontemleri ile elde edilen kesme puanlarinin genellenebilirlik kurami ile karsilastirilmasi [A comparison of cut points obtained from Nedelsky and Angoff standard setting methods with Generalizability Theory]. *Journal of Measurement and Evaluation in Education and Psychology, 1*(1), 22-28. Retrieved from https://dergipark.org.tr/tr/pub/epod/issue/5808/77250

Tekin, H. (2017). *Egitimde Olcme ve Degerlendirme* [Measurement and evaluation in education] (Yirmi besinci b.). Yargi Yayinevi.

Turgut, M. F., & Baykul, Y. (2014). *Egitimde Olcme ve Degerlendirme* [Measurement and evaluation in education] (Altinci b.). Ankara: Pegem Akademi.

Tulubas, G. (2009). *Psikolojik testlerde Angoff ve sinir grup yontemleri ile kesme puanlarinin belirlenmesi* [To determine cut off points with Angoff and Borderline Group Tecniques in psychological testing] (Unpublished master's thesis). Hacettepe University, Ankara.

Açık Uçlu Maddelerden Oluşan Testler İçin İki Standart Belirleme Yönteminin Karşılaştırılması

Atıf:

Ozarkan, H. B., & Dogan, C. D. (2020). A comparison of two standard-setting methods for tests consisting of constructed-response items. *Eurasian Journal of Educational Research 90, 121-138,* DOI: 10.14689/ejer.2020.90.7

Özet

Problem Durumu: Testlerden elde edilen puanlar çeşitli ölçütlerle kıyaslanarak bir karara varılmaktadır. Bu ölçütler bağıl veya mutlak olabilmektedir. Bağıl veya mutlak ölçütle belirlenen kesme noktası test puanları ölçeğini iki veya daha fazla bölgeye ayırıp testi alanları sınıflandırmaktadır. Standart belirleme olarak adlandırılan bu sistematik süreç test geliştirme sürecinin bir parçasıdır. Standart belirleme yöntemlerinden elde edilen kesme puanı test maddelerine (test merkezli) ya da testten elde edilen puanlara (öğrenci merkezli) dayalıdır.

Günümüzde farklı madde formatlarının kullanıldığı testler ön plana çıkmaktadır. Uluslararası Öğrenci Değerlendirme Programı (PISA), Uluslararası Matematik ve Fen Eğilimleri Araştırması (TIMSS) gibi uluslararası araştırmaların yanı sıra Türkiye'de yürütülen Akademik Becerilerin İzlenmesi ve Değerlendirilmesi (ABİDE) izleme araştırmalarında çoktan seçmeli maddelerin yanı sıra açık uçlu maddeler de kullanılmaktadır. Ancak Türkiye'de yapılan standart belirleme araştırmaları çoğunlukla çoktan seçmeli maddelerden oluşan testler üzerinden yürütülmüştür. Ek olarak açık uçlu maddeleri içen bir test için standart belirleme çalışmasına rastlanmamıştır. Alanyazında öğrenci merkezli ve test merkezli yöntemlerin karşılaştırıldığı çalışmalar sınırlı sayıdadır. Ayrıca alan yazın incelendiğinde standart belirleme yöntemleri özelinde test merkezli yöntemlerden Genişletilmiş Angoff ve öğrenci merkezli yöntemlerden Karşıt Gruplar yöntemlerinin karşılaştırıldığı araştırmalar da sınırlı sayıdadır.

Araştırmanın Amacı: Bu çalışmanın amacı ortaokul 8. Sınıf matematik dersi cebir öğrenme alanına ilişkin hazırlanan açık uçlu maddelerden oluşan bir başarı testi için Genişletilmiş Angoff ve Karşıt Gruplar standart belirleme yöntemlerinden elde edilen kesme puanlarını karşılaştırmaktır.

Araştırmanın Yöntemi: Tarama türünde yürütülen bu araştırmanın çalışma grubunu Genişletilmiş Angoff yöntemi için 8 öğretmen, Karşıt Gruplar yöntemi için ise 1 öğretmen ve 75 öğrenci oluşturmaktadır. 8 açık uçlu maddeden oluşan matematik başarı testi ve her bir maddeye ilişkin dereceli puanlama anahtarı (DPA) araştırmacı tarafından geliştirilmiştir. Matematik başarı testinin iç tutarlık anlamındaki güvenirlik katsayısı Cronbach α değeri 0,91 olarak hesaplanmıştır. Bu testin güvenirliğine kanıt oluşturmaktadır. 75 öğrenciye uygulanan matematik başarı testi iki puanlayıcı

tarafından puanlanmıştır. Puanlayıcılar arasındaki uyumu araştırmak amacıyla her bir madde için hesaplanan kappa değerleri istatistiksel açıdan manidar bulunmuştur (p<.001). Elde edilen kappa değerleri 0,925 ile 1 arasındadır. Puanlayıcılar arasındaki uyum çok yüksek düzeydedir. Bu DPA'nın geçerliğine kanıt oluşturmaktadır. Genişletilmiş Angoff yönteminde uzmanlar testte bulunan her madde için DPA'yı dikkate alarak yeterli-yetersiz sınırındaki öğrencinin kaç puan alacağına ilişkin karar vermiştir. Bu süreçte araştırmacı tarafından geliştirilen uzman görüş formunu kullanılmıştır. Karşıt Gruplar yönteminde testi alan öğrenciler uzman tarafından iki grup halinde sınıflandırılmıştır. Bunun için sınıflandırma formu kullanılmıştır.

Genişletilmiş Angoff yönteminde uzmanların yeterli ve yetersiz sınırındaki öğrencilere ilişkin verdikleri puanların her bir madde için aritmetik ortalaması hesaplanmıştır. Bu aritmetik ortalamaların toplamı ise kesme puanını vermiştir. Karşıt Gruplar yönteminde ise uzman yargılarına göre yeterli-yetersiz olarak iki grup halinde sınıflandıran öğrencilerden elde edilen test puanlarının dağılımlarının ortancalarının orta noktası hesaplanarak kesme puanı elde edilmiştir. Ayrıca lojistik regresyon yöntemiyle kesme puanı hesaplanmıştır. Yöntemlerden elde edilen kesme puanlarının üzerinde puan alarak yeterli kabul edilen öğrenci oranları arasındaki fark, bağımlı iki oran arasındaki fark testiyle, farkın manidarlığı ise z istatistiği ile belirlenmiştir. Öğrencilerin yeterli-yetersiz olarak sınıflandırılması bakımından yöntemler arasındaki uyumun araştırılmasında Cohen'in Kappa katsayısı kullanılmıştır.

Araştırmanın Bulguları: Genişletilmiş Angoff yönteminden elde edilen kesme puanı 13,38, Karşıt gruplar yönteminden elde edilen kesme puanı ise 12,50 dir. İki standart belirleme yönteminin yeterli kabul edilen öğrenci oranları arasında manidar fark göstermediği belirlenmiştir. Öğrencilerin yeterli ve yetersiz olarak sınıflandırılması bakımından yöntemler arasındaki uyumun yüksek düzeyde olduğu bulgusuna ulaşılmıştır.

Araştırmanın Sonuçları ve Öneriler: Genişletilmiş Angoff ve Karşıt Gruplar standart belirleme yöntemlerinin elde edilen kesme puanları açısından farklılık göstermediği sonucuna varılmıştır. Ancak standart belirleme yöntemlerinin süreçlerindeki farklılıklardan kaynaklanabilecek durumların gözden kaçırılmaması gerekmektedir. Bu farklılıklar öğrenci veya uzman grubunun yapısı kaynaklı olabilir.

Öğrenci grubunun ölçülen özellik bakımından heterojen bir dağılım göstermesi durumunda her iki yöntem de kullanılabilir ancak homojen dağılım gösteren gruplar için Karşıt Gruplar yöntemi sınırlılık doğurabilir. Bu sınırlılık performans kategorilerindeki öğrenci tanımlarını birbirinden ayırmanın güçlüğü ve dolayısıyla öğrencileri karşıt iki gruba ayırmanın güçlüğünden kaynaklanabilir. Bu durumda yeterli ve yetersiz olarak sınıflandırılan öğrenci grupları arasında gözlenen farklar da belirgin olmayabilir. Bu nedenle öğrenci grubu ölçülen özellik bakımından çarpık bir dağılım gösteriyorsa Karşıt Gruplar yöntemi yerine Genişletilmiş Angoff yönteminin tercih edilmesi daha uygun olacaktır.

Genişletilmiş Angoff yönteminde performans kategorilerinin tanımlanması aşamasındaki grup çalışması ve kesme puanlarının belirlenmesi aşamasındaki grup

tartışması, uzmanlar arasında etkileşime imkân vermektedir. Bu bakımdan standart belirleme çalışmalarına katılacak uzmanların daha önce benzer bir çalışmaya katılmış olması sürece katkı sağlayacaktır.

Karşıt Gruplar yönteminde kesme puanı istatistiksel yollarla elde edildiğinden bu yöntemi uygulayacak uzmanların gerekli istatistiksel alt yapıya sahip olması gerekmektedir.

Açık uçlu maddelerden oluşan bir başarı testi için benzer çalışmanın farklı yöntemlerle yürütmesi önerilebilir. Çoklu puanlanan tipik performans testleri için aynı yöntemlerle standart belirleme çalışması yürütülüp elde edilen sonuçların karşılaştırılması önerilebilir. İki yöntem arasındaki uyum düzeyinin daha homojen bir öğrenci grubunda yeniden araştırılması önerilebilir. Aynı yöntemler kullanılarak farklı bir uzman grubundan elde edilen kesme puanlarının karşılaştırılması önerilebilir. Covid-19 pandemi sürecinde Genişletilmiş Angoff yönteminde uzman paneli görüşmeleri online yapılabilir. Online toplantılar heterojen bir uzman grubu oluşturmaya imkan verebilir. Bu nedenle Covid-19 sürecinde her iki yöntem de kullanılabilir.

Anahtar Sözcükler: standart belirleme, kesme puanı, açık uçlu maddeler, extended Angoff yöntemi, karşıt gruplar yöntemi