International Journal of Assessment Tools in Education
2025, Vol. 12, No. 1, 78–92
https://doi.org/10.21449/ijate.1371162
journal homepage: https://dergipark.org.tr/en/pub/ijate
Research Article

Examining the cut-off score of the English B1 progression exam according to different standard setting methods

Rümeysa Kaya[1*], **Bayram Çetin**[2]

[1]Gaziantep University, Gaziantep, Türkiye
[2]Gaziantep University, Faculty of Education, Department of Educational Sciences, Gaziantep, Türkiye

ARTICLE HISTORY
Received: Oct. 5, 2023
Accepted: Apr. 16, 2024

Keywords:
Angoff,
Angoff Y/N,
Nedelsky,
Ebel.

Abstract: In this study, the cut-off scores obtained from the Angoff, Angoff Y/N, Nedelsky and Ebel standard methods were compared with the 50 T score and the current cut-off score in various aspects. Data were collected from 448 students who took Module B1+ English Exit Exam IV and 14 experts. It was seen that while the Nedelsky method gave the lowest cut-off score, Angoff Y/N method gave the highest cut-off score. The z test was used to determine the difference between the percentages of students who were considered successful according to the methods, and all *z* values were found to be significant. The classification of students according to their achievement status was examined with the Cohen's Kappa test. Spearman Brown Rank Differences Correlation coefficient was calculated to examine the relationship between the MPSs of the experts according to the methods, and the highest correlation was found between the Angoff-Ebel methods. Wilcoxon test was used to examine the significance of the difference between the MPS of the methods. Because of the test, the difference between Angoff-Nedelsky, Angoff-Ebel, Angoff Y/N-Nedelsky and Nedelsky-Ebel methods was found to be significant. Among the expert decisions, it was seen that there was a moderate level of agreement in the Angoff, and a high level of agreement in the Ebel and Nedelsky methods. A significant difference was found between the current cut-off score, the 50 T score, and the percentages of students considered successful according to the methods.

1. INTRODUCTION

Measurement tools are used when determining the impact of educational activities on individuals. The measurement tool can be written or oral. Evaluation is made when the measurement result obtained from the measurement tool is compared with a criterion, and a decision is made about the individual's success. Having common goals and criteria in the assessment - evaluation process will ensure standardization in education. This standardization will develop a common language even at the international level. For example, for the English language level, an individual at the B1 level is expected to be able to talk about experiences in daily life, daily events, and topics of interest.

*CONTACT: Rümeysa KAYA rumeysakayarusen@hotmail.com Gaziantep University, Gaziantep, Türkiye

e-ISSN: 2148-7456

The cut-off score is used to determine the level of language skills an individual possesses according to his/her performance. Before determining the cut-off score, it would be more appropriate to determine and define the performance levels. The cut-off score and performance levels do not have to be determined by the same experts.

The steps and methods used in the cut-off point determination require a certain process called the standard-setting process. There are many methods that can be used in the standard-setting process. The method of application may differ in terms of analysis and interpretation of the obtained data. Jeager (1989) divided these methods into two groups: test-centered methods and student-centered methods. In test-centered methods, experts form the minimum passing score based on their judgments about the test items, while in student-centered methods, they create a cut-off point based on the knowledge and skills of the individuals who answered the test. The test-centered methods that are commonly used are Angoff, Angoff Y/N, Nedelsky, Ebel, and Marking methods, while the student-centered methods mostly utilized are the Boundary Group method and Opposite Groups methods. One of the advantages of these methods is that the cut-off score from the test-centered methods can be obtained without applying the test to the students and that the experts are not affected by the characteristics of the student groups while determining the cut-off score. The test-centered methods used in this study are briefly mentioned below.

1.1. Angoff Method

In this method, developed by William H. Angoff in 1971, experts are asked to predict how many of the 100 students on the pass-fail limit will be able to answer the item correctly for each item in the test. The minimum passing score of that expert is obtained by adding the probability values given by the expert for the items, dividing by the number of items in the test, and multiplying the result with the evaluation score of the test (the highest score that can be obtained from the test). The mean score of the test is obtained by taking the average of the MPS (minimum passing score) found in this way.

1.2. Angoff Y/N Method

In this method developed by Impara and Plake in 1997, experts are asked to give one point for each item in the test if they think an individual on the pass-fail limit will answer that item correctly, and zero points if they think they will answer incorrectly. After adding the points given by the expert for the items and dividing by the number of items in the test, the expert's MPS is obtained by multiplying the result with the evaluation score of the test. The cut-off score of the test is found by taking the mean of the MPS.

1.3. Nedelsky Method

In this method developed by Leo Nedelsky in 1954, experts are asked to estimate the number of options that a pass-fail student can eliminate when reaching the correct answer for each item in the test. The probability of answering the item correctly is found with the formula '1/number of remaining options'. This method can only be applied in tests containing multiple-choice items. The expert's MPS is by adding these probability values calculated based on expert judgments, dividing by the number of items in the test, and multiplying the result by the evaluation score of the test. The cut-off score of the test is obtained by averaging the MPSs.

1.4. Ebel Method

In this method, developed by Ebel in 1972, experts are asked to evaluate each item in the test in two stages. In the first stage, the experts examine the items in two dimensions, namely convenience and difficulty, and place them in a 3x4 table. There are four subgroups in the dimension of relevance: necessary, important, acceptable, and debatable. In the difficulty dimension, there are three subgroups as easy, medium, and difficult. In the

second stage, they predict how many of the 100 students on the pass-fail limit will be able to answer the items in each cell correctly. A score is obtained for the cell by multiplying the number of items in the cell with the percentage determined for that cell. The result obtained by adding the cell scores and dividing by the number of items in the test is multiplied by the evaluation score of the test, and the expert's MPS is found. The cut-off score of the test is obtained by averaging the MPSs.

The standard-setting method to be used should be understandable by experts, and the results should be interpretable. Working with a large group of experts will provide a more accurate cut-off score. The expert group should be informed about the method of application, the purpose, and the characteristics of the test.

Studies comparing different standard-setting methods are avaliable in the literature (Berk, 1986; Boduroğlu, 2017; Buckendahl *et al*., 2002; Livingston & Zieky, 1983; Norcini *et al*., 1987; Ömür & Selvi, 2010). In this study, it was aimed to examine how the cut-off points changed according to the four test-centered standard-setting methods, how the obtained cut-off scores affected the percentage of students who were considered successful, how the decisions of the experts about the items changed according to the methods, and the consistency between the expert decisions. In addition, the cut-off score obtained from the standard-setting methods and the 50 T score as a norm-based assessment method, were compared in various aspects. In this study, answers were sought for the following problem statements:

1. What are the cut-off scores for Module B1+ Exit Exam IV using Angoff, Angoff Y/N, Nedelsky, and Ebel standard-setting methods?
2. Is there a significant difference between the percentages of successful students according to the cut-off points obtained from the standard-setting methods used?
3. Is there a consistency between the standard-setting methods used to classify students as successful or unsuccessful according to the methods?
4. Is there a consistency between the standard-setting methods used regarding minimum passing scores among experts?
5. What are the relationships between the actual difficulty values of the items, the estimated item response probabilities given by the experts using the Angoff method, and the estimated item response probabilities given by the experts using the Ebel method?
6. What is the level of agreement between the experts' decisions on the items according to the standard-setting methods used?
7. Do the percentages of students who score above the current cut-off score of Module B1+ Exit Exam IV and the cut-off scores obtained by the standard-setting methods used in the research differ?
8. What is the cut-off score obtained according to the 50 T score, the number of students accepted as successful according to this score, and the percentage of students, and is there a significant difference between the 50 T score and the percentage of students who are considered successful according to the cut-off scores obtained from the standard-setting methods used in this study?
9. Is there harmony in classifying students as successful or unsuccessful according to the standard-setting methods used in this study with a T score of 50?

2. METHOD

This study aimed to obtain cut-off points from different standard-setting methods and examine the obtained cut-off scores in different centers. In this context, it is a descriptive and relational study. Excel, JASP 0.16.1.0, and SPSS Statistics v23 x64 programs were used during the tests and analyses. The significance value was accepted as .05 in all analyses in the study.

2.1. Study Group

In this study, data were collected from two different groups. The 1st group consisted of 448 students who answered the Module B1+ Exit Exam IV. The second group was 14 lecturers working at the School of Foreign Languages and filling out the standard-setting methods forms. While determining the number of experts, previous studies on this subject were taken into account (Hurtz & Hertz, 1999).

2.2. Data Collection Tools

This study used Module B1+Exit Exam IV, which was held at the end of the 2021-2022 academic year of the School of Foreign Languages of a state university, was used. There are 62 items in the exam, which consists of four sub-sections: Listening, use of English, vocabulary and reading. Student scores were calculated in accordance with the exam guidelines. As a result of the analyses made on these scores, it was seen that the difficulty and distinctiveness of the test were moderate (KR20=0.69, test difficulty (*P*)=0.51). Student responses showed a normal distribution (kurtosis=0.02, skewness=0.20).

While obtaining data from the experts, expert evaluation forms were given to the experts along with the exam questions. Experts filled out the forms following the instructions. In this study, pass-fail students were identified as individuals with B1-level characteristics made by the Common European Framework of Reference for Languages.

B1 Level (Intermediate-Independent User):

- He/she can convey the events and experiences he/she has lived; can talk about their dreams, hopes, and wishes, and briefly explain their views and plans with their reasons.
- Can handle most situations encountered when traveling, where the language is spoken. Can understand the main lines of written expressions based on familiar topics in daily life.
- Can express himself/herself in line with his/her interests or on the subjects he/she knows through simple texts with links between ideas.

2.3. Analysis of Data

For the first sub-problem of the study, expert evaluation forms prepared in accordance with the application of the methods used in the study were given to the experts. While 14 expert forms were used for Angoff, Angoff Y/N, and Nedelsky methods, the forms belonging to 4 experts were deemed invalid in the Ebel method and 10 expert forms were used.

In the solution of the second sub-problem of the study, the student scores were classified as successful or unsuccessful according to the cut-off points obtained from the methods. The number and percentage of successful students were determined and the significance of the difference between these percentages was examined with the *z*-test. The *z*-test is used to check the significance of the difference between two dependent percentages in sample numbers larger than 30.

Cohen's Kappa test was used to determine the compatibility between the classification of students' achievement status according to the methods in the solution of the third sub-problem of the study. In order to make the scores suitable for the test, the cut-off score of the method and above were converted to 1 and other scores to 0. The fit rating scale suggested by Landis and Koch (1977) was used to interpret the results. This scale is as follows:

0.00 - 0.20 = slight
0.21 - 0.40 = fair
0.41 - 0.60 = moderate
0.61 - 0.80 = substantial
0.81 - 1.00 = almost perfect

In the solution of the 4th sub-problem of the study, the relationship between the expert MPS was examined by calculating the Spearman-Brown Rank Differences Correlation Coefficient. The Spearman-Brown Rank Correlation Coefficient is a statistical method used to examine the relationship between variables when the data is less than 30. The following rating scale was used to interpret this correlation coefficient (İlhan, 2022).

$r < 0.20$ = no relationship
$0.20 < r < 0.39$ = weak relationship
$0.40 < r < 0.59$ = moderate correlation
$0.60 < r < 0.79$ = high level of association
$0.80 < r < 1.00$ = very high correlation

In the continuation of the solution, the Friedman chi-square test was performed to examine the significance of the difference between the mean of the MPSs obtained from the methods. Friedman chi-square test is a non-parametric test used to check whether the mean scores of two or more groups differ significantly Wilcoxon Signed Ranks test was used to see the difference between the mean of MPS and which methods were significant.

In the solution of the fifth sub-problem of the research, the average of the percentage estimates of the experts for answering the items based on the Angoff and Ebel method (considering the percentages obtained in the Ebel method on an item basis). With these averages, descriptive statistics based on students' exam results were found. Pearson Product Moments Correlation Coefficient was calculated since the data showed normal distribution.

In the solution of the sixth sub-problem of the study, the expert evaluation forms were transferred to Excel according to the methods filled by the experts. Kendall's W fit coefficient was calculated by considering the agreement between the expert decisions, Kendall's W fit coefficient in Angoff method, Cochran Q test in Angoff Y/N method, Intraclass Correlation Coefficient for Nedelsky method and the percentage values given by the experts about cells in Ebel method on an item basis. Kendall's W concordance coefficient is used when the number of raters is more than two and a single cohesion coefficient is desired to be obtained from the data. The scale used in the interpretation of this coefficient is given below (Rovai *et al*., 2014):

0.00 – 0.20 = very weak effect
0.21 – 0.40 = weak effect
0.41 – 0.60 = medium effect
0.61 – 0.80 = strong effect
0.81 – 1.00 = very strong effect

Since the Cochran Q test examines the agreement between expert evaluations in two categories, such as 1-0 or positive-negative, this test was preferred in the Angoff Y/N method.

For the solution of the seventh sub-problem of the study, the passing grade of the B1 level of the School of Foreign Languages, where the study was carried out, was 60, and it was assumed in this study that the passing grade was created only according to Module Exit Exam IV. The number and percentages of students who got the current cut-off score and above of the methods and the exam were found. Then, the significance of the difference between these percentages was examined with the formula of the *z*-test.

In the solution of the eighth sub-problem of the study, the scores obtained by the students from the exam were converted into T scores. In this study, 50 T score was determined as a criterion as a norm-based assessment. The number and percentage of students considered successful according to the 50 T score were found. The significance of the difference between the percentages of students who were considered successful according to the methods and those who were considered successful according to the 50 T score was examined by performing the *z*-test.

In the solution of the ninth sub-problem of the study, the scores of the students who were considered successful according to the 50 T score and the cut-off point of the methods were converted to 1 and the other scores to 0. Then, Cohen's Kappa test was performed on these data.

3. RESULTS

In the solution of the first sub-problem of the study, MPSs of the methods were calculated based on the standard-setting methods forms filled by the experts. Since four expert forms were deemed invalid in the Ebel method, the MPS of four experts could not be calculated for this method. In Table 1, the MPSs of the experts according to the methods are given:

Table 1. *MPS of experts by methods*.

Experts	Minimum Passing Score (MGP) for Angoff Method	Minimum Passing Score (MGP) for Angoff Y/N Method	Minimum Passing Score (MGP) for Nedelsky Method	Minimum Passing Score (MGP) for Ebel Method
Expert 1	73.71	72.58	64.06	73.15
Expert 2	94.48	77.42	33.00	83.63
Expert 3	49.76	45.16	40.94	48.65
Expert 4	56.05	64.52	51.18	52.10
Expert 5	63.23	82.26	65.11	-
Expert 6	72.10	62.90	53.23	70.56
Expert 7	72.02	64.52	64.19	70.48
Expert 8	67.82	46.77	39.19	52.58
Expert 9	58.06	72.58	39.02	41.53
Expert 10	58.39	62.90	34.66	50.48
Expert 11	57.34	74.19	39.29	41.53
Expert 12	39.81	58.06	42.03	-
Expert 13	56.69	56.45	37.66	-
Expert 14	53.95	61.29	57.65	-

As can be seen in Table 1, since the MPPs of the Angoff method contain extreme values, the cut-off scores of the methods were obtained by taking the mean of the corrected (pruned) mean in this method and the MPS of the other methods, since the MPS of the other methods did not contain extreme values. The cut-off points calculated according to the MPSs obtained from the experts are given in Table 2.

Table 2. *Cut-off scores of Angoff, Angoff Y/N, Nedelsky, and Ebel methods.*

Methods	Angoff	Angoff Y/N	Nedelsky	Ebel
Cut-off Score by Method	61.59	64.40	47.23	58.47

When Table 2 is examined, the highest cut-off score in this study was obtained from the Angoff Y/N (64.40) method, while the lowest cut-off score was obtained with the Nedelsky method (47.23). It was observed that there was a difference of 14.36 points between the highest cut-off score and the lowest cut-off score. This may be due to the way the methods are applied. It is possible that the Nedelsky method, which involves focusing on all options together with the item root, may have been overlooked in this instance. This may have resulted in the clues provided by the correct option being misinterpreted, leading experts to consider the items in question to be more challenging than they actually were. In the Angoff Y/N method, on the other hand, it may be due to the decrease in the judgment options related to the items by evaluating the items according to only two value judgments (1-0). The cutoff scores of the Angoff and Ebel methods are close to each other because both methods contain an estimate of the percentage of students at the minimum proficiency level. The fact that the lowest cut-off score belongs to the Nedelsky

method also coincides with the results of the studies conducted by Tanriverdi (2006), Taşdemir (2009), and Yildirim Kan (2019).

For the second sub-problem of the study, the cut-off points obtained from the methods and the number and percentages of students who scored above were calculated. Then, a z-test was performed to test the significance of the difference between these percentages. Table 3 gives the percentage of students who are considered successful according to the methods and the results of the z-test.

Table 3. *The number of students deemed successful according to the methods, their percentage, and z-test results.*

Methods	*N*	%	*z*
Angoff	79	17.63	5.1*
Angoff Y/N	53	11.83	
Angoff	79	17.63	13.68*
Nedelsky	26	59.38	
Angoff	79	17.63	4.58*
Ebel	100	22.32	
Angoff Y/N	53	11.83	14.60*
Nedesky	266	59.38	
Angoff Y/N	53	11.83	6.86*
Ebel	100	22.32	
Nedelsky	266	59.38	12.89*
Ebel	100	22.32	

*$p<.05$

The value required for a significant difference at the .05 level in the *z*-test is 1.96. All *z*-values found as a result of comparing the methods' percentages in pairs were greater than 1.96. It was seen that the difference between the percentages of students who were considered successful according to the methods was significant. This result was obtained because the difference in cut-off scores affects the percentage of students who are considered successful according to the methods.

In the solution of the third sub-problem of the study, Cohen's Kappa test was performed to determine the fit in terms of classifying the students according to their success status according to the methods and the degree of this fit, if any, and the values found were interpreted. The results of the Cohen's Kappa test are given in Table 4.

Table 4. *Cohen's Kappa test results.*

Methods	Kappa coefficient (*k*)	Compliance Level
Angoff - Angoff Y/N	0.77	substantial fit
Angoff – Nedelsky	0.26	fair fit
Angoff Y/N- Nedelsky	0.17	slight fit
Angoff – Ebel	0.85	Almost perfect fit
Angoff Y/N – Ebel	0.64	Substantial fit
Nedelsky – Ebel	0.33	fair fit

As seen in Table 4, all *k* values are positive, which indicates that the methods were correctly understood by the experts and that the expert's decisions about the item were consistent. Considering the level of fit, the best fit was between Angoff and Ebel methods (Kappa=0.85,

Kappa>0.75, almost perfect fit), and the lowest fit between Angoff Y/N and Nedelsky methods (Kappa=0.17, Kappa<0.20, slight fit). As the cut-off points of the methods get closer to each other, the fit value between them also increases. The results found between Angoff and Ebel also coincide with the results of previous studies. (Demir, 2014; Gündeğer, 2012).

In the solution of the fourth sub-problem of the study, the Spearman-Brown Rank Correlation Coefficient was calculated to examine the relationship between MPSs obtained from experts according to the methods. The Friedman Chi-Square test was used to check the existence of agreement between all methods in terms of the mean of MPSs. The Spearman-Brown Rank Differences Correlation Coefficient results are given in Table 5.

Table 5. *Spearman Brown rank differences correlation coefficients between MPSs.*

		Angoff	Angoff Y/N	Nedelsky	Ebel
Angoff	N	-			
	R	-			
	P	-			
Angoff Y/N	N	14	-		
	R	0.51	-		
	P	0.06	-		
Nedelsky	N	14	14	-	
	R	0.03	0.17	-	
	P	0.92	0.55	-	
Ebel	N	10	10	10	-
	R	0.86*	0.16	0.24	-
	P	0.00	0.67	0.51	-

A statistically significant relationship was found only between the experts' MPSs for the Angoff and Ebel methods. ($p<.05$). In addition, the correlation value between these two methods was positive and very high ($r>.80$, $p<.05$). As a result of the Friedman Chi-Square Test, it was observed that at least one of the MGP averages differed significantly from the others ($\chi^2=13.29$, $p<.05$). Wilcoxon Signed Ranks Test was used to check which mean of MGP of the methods was significant. The results of the Wilcoxon Signed Ranks Test are given in Table 6.

Table 6. *Wilcoxon signed-row test results.*

Methods	*N*	*Z*	*p*
Angoff Angoff Y/N	14	0.41	.68
Angoff Nedelsky	14	2.92*	.004
Angoff Ebel	10	2.81*	.005
Angoff Y/N Nedelsky	14	3.30*	.001
Angoff Y/N Ebel	10	0.66	.507
Nedelsky Ebel	10	2.80*	.005

*$p<.05$

As can be seen in Table 6, the methods with a significant difference in terms of MPS averages are Angoff - Nedelsky, Angoff - Ebel, Angoff Y/N - Nedelsky and Nedelsky - Ebel methods. While there is a very high correlation between the MPSs of the Angoff and Ebel methods, the

significant difference between the MPS averages indicates that the MPSs of the experts according to the two methods are in the same direction, but the MPS averages of one of the methods differ due to the lower MPSs of the other methods. While there is no relationship between the MPSs of Angoff Y/N – Ebel and Nedelsky - Ebel methods, the lack of a significant difference between the MPS averages shows that the experts' perception of ease-difficulty regarding the whole test for the two methods has changed. However, when the averages of these MPSs are averaged, the results are close to each other.

In the solution of the fifth sub-problem of the study, the difficulty levels of the items were calculated based on the answers of the students who participated in the exam. Then, the average of the item answer probability estimates made by the experts using the Angoff and Ebel methods were taken. Thus, the average response percentage of each item was found according to both methods. In Table 7, descriptive statistics based on real item difficulty with Angoff and Ebel methods are given:

Table 7. *Descriptive statistics for item difficulty and actual item difficulty based on Angoff and Ebel methods.*

	Estimated Item Difficulty Based on Angoff Method	Estimated Item Difficulty Based on Ebel Method	Real Item Difficulties
N	62	62	62
Minimum	0.54	0.51	0.13
Maksimum	0.72	0.89	0.89
Average	0.62	0.58	0.51
Standard deviation	0.04	0.04	0.20
Distortion	0.14	0.09	0.11
Kurtosis	0.33	0.56	0.72

When Table 7 is examined, it is seen that the difficulty levels estimated according to the Ebel and Angoff judgment method are easier than they actually are. Since the data showed a normal distribution, the relationship between the item difficulties according to the three conditions was examined by calculating the Pearson Product Moments Correlation Coefficient. The results are given in Table 8.

Table 8. *Correlation between Angoff and Ebel methods estimated item difficulties and actual item difficulties.*

		Real Item Difficulty	Angoff-Based Item Difficulty	Item Difficulty Based on Ebel
Real Item Difficulty	*r*	-		
	p	-		
Angoff-Based Item Difficulty	*r*	0.52*	-	
	p	<.001	-	
Item Difficulty Based on Ebel	*r*	0.36*	0.67*	-
	p	0.004	<0.001	-

*$p<.05$

It was observed that there was a positive and moderately significant correlation between the experts' average of the estimated item difficulties based on the Angoff method and the actual item difficulties ($r=0.52$, $p<.05$, $N=62$). This result coincides with the result of Çetin (2011)'s study. It was observed that there was a positive and weakly significant correlation between the experts' estimated item difficulties based on the Ebel method and the actual item difficulties ($r=0.36$, $p<.05$, $N=62$). It was observed that there was a positive and highly significant correlation between the experts' mean estimated item difficulties based on the Angoff and Ebel methods ($r=0.67$, $p<.05$, $N=62$).

The significant relationship between the average of the estimates made by the experts about the item difficulties according to the Angoff and Ebel method and the actual item difficulties indicate that the predictions made by the experts using the methods are valid. The weak correlation between the estimated item difficulty averages based on the Ebel method and the actual item difficulties may be because the percentage values given for cells in the Ebel method are considered on an item basis.

In the solution of the sixth sub-problem of the study, the harmony between the expert decisions was examined. Kendall's W coefficient of agreement was found to be .561 for the agreement between the estimates of 14 experts for 62 items in the Angoff method ($\chi^2=451.943$, $sd=13$, $p<.05$). This value shows that the expert decisions are moderately compatible in the Angoff method. This harmony also coincides with the results of Kiliç (2013) study.

Cochran Q coefficient of agreement was checked for the consistency between the decisions made by 14 experts for 62 items in the Angoff Y/N method, and it was seen that the expert decisions were compatible ($Q=43.356$, $p<.05$). In the Nedelsky method, it is seen that the In-Class (Cluster) correlation coefficient of agreement between the decisions made by 14 experts for 62 items is 0.70. This value shows that the expert decisions are highly compatible with the Nedelsky method.

The Kendall W agreement coefficient for the agreement between the estimates of 10 experts for 62 items in the Ebel method was found to be .691 ($\chi^2=385.220$, $sd=9$, $p<.05$). This value shows that the expert decisions are strongly compatible in the Ebel method. The increase in the number of experts and the number of items in the test makes it difficult to achieve high agreement among experts.

In the solution of the seventh sub-problem of the study, 21.21% (95 students) of the students who took the exam according to the current cut-off score were successful. The significance of the difference between the current cut-off score and the percentages of students who were considered successful according to the cut-off scores obtained from the methods was examined with the *z*-test. The *z* test results are given in Table 9.

Table 9. *z-test results for the percentage of successful students according to the methods and current cut-off score.*

	N	%	*z*
Angoff Method	79	17.63	4*
Current Passing Score	95	21.21	
Angoff Y/N Method	53	11.83	6.48*
Current Passing Score	95	21.21	
Nedelsky Method	266	59.38	13.08*
Current Passing Score	95	21.21	
Ebel Method	100	22.32	2.23*
Current Passing Score	95	21.21	

*$p<.05$

When the current cut-off points and the methods were compared one by one in terms of the percentage of students who were considered successful, it was seen that all *z* values were significant. This shows that the current cut-off score and the cut-off score of the methods differ significantly from each other.

In the solution of the eighth sub-problem of the study, student scores were converted to T scores. In this evaluation, 50 T points were taken as a criterion. According to the 50 T score, 47.32% of the students (212 students) were successful. The significance of the difference between the 50 T score in terms of the percentage of students considered successful and those considered

successful according to the cut-off points obtained from the methods was examined with the *z*-test. The *z*-test results are given in Table 10.

Table 10. *z-test results for the percentage of students deemed successful according to methods and 50 T-scores.*

	N	%	*z*
Angoff Method	79	17.63	11.53*
50 T Points	212	47.32	
Angoff Y/N Method	53	11.83	12.61*
50 T Points	212	47.32	
Nedelsky Method	266	59.38	7.35*
50 T Points	212	47.32	
Ebel Method	100	22.32	10.58*
50 T Points	212	47.32	

*$p<.05$

Looking at Table 10, it was seen that all *z* values were significant. This indicates that the cut-off scores of standard-setting methods and the 50 T score, which is an assessment method based on norms, differ significantly. This result is similar to that of the study of Çukadar (2013) and Şahin (2019).

For the solution of the ninth sub-problem of the study, 50 T points and student scores considered successful according to the cut-off point of the methods were converted as 1, and student scores considered unsuccessful were converted to 0. Then, Cohen's Kappa Test was performed on these data. Statistical information about the test result is given in Table 11.

Table 11. *The results of the Cohen's Kappa test were performed with a T score of 50 and the level of agreement between the methods.*

Methods	Kappa coefficient (k)	Compliance Level
Angoff – 50 T	0.39	Fair fit
Angoff Y/N -50 T	0.26	Fair fit
Nedelsky – 50 T	0.76	Substantial fit
Ebel-50 T	0.49	Moderate fit

It was seen that Nedelsky method ($k = 0.76$, substantial fit) gave the best fit with a T score of 50, and Angoff Y/N method ($k = 0.26$, fair fit) gave the lowest fit, in terms of classifying students according to their achievement status. This is because the T score of 50 and the cut-off score of the Nedelsky method are close to each other.

4. DISCUSSION and CONCLUSION

In this study, the cut-off score of Gaziantep University foreign language B1 level exam was calculated using Angoff, Angoff Y/N, Nedelsky and Ebel standard-setting methods. These scores were then compared, in various aspects, with the existing cut-off score and the 50 T score, which is one of the norm-based evaluation methods. The results obtained and discussions based on these results are given below.

As evidenced by the findings, the cut-off scores of the methods in question exhibited notable discrepancies. These discrepancies can be attributed to the fact that the specific areas of focus for experts may vary depending on the method being employed. The result of the lowest cut-off point in this study belongs to the Nedelsky method, which is in line with the results of the previous studies, except for the study of Taşdelen (2009). This may be because the experts perceive the items as more difficult than they are since the Nedelsky method examines all the

options one by one. The result of the Angoff Y/N method, which acts with only two judgments, has a very low cut-off score. This result is consistent with the results of the previous study. The Angoff Y/N method's ability to make values over only two sources from the fact that its results differ significantly from other methods. The cut-off score of the Ebel method is lower than the cut-off scores of the Angoff and Angoff Y/N methods. It has been shown that the more complex the understanding and application of the standard-setting method is, the lower the cut-off score is.

The results indicate that the percentages of students who are considered successful according to the cut-off scores differ significantly for all methods, and this finding showcases that even minor differences between the cut-off scores significantly impact the exam results. It has also been observed that there is an inverse proportion between the cut-off score and the percentage of students considered successful. In cases where the cut-off points of the methods were close to each other, it was seen that the results of the classification of the students according to their success were close to each other. The Nedelsky method gave lower coefficients in terms of compatibility with other methods because the cut-off score was much lower than the other cut-off scores. The perfect harmony between Angoff and Ebel methods stems from the common points in the way the methods are applied. The large difference between the percentages of students who are considered successful according to the standard-setting methods reveals the importance of making decisions by using more than one method in creating cut-off points for the exams.

The moderate relationship between the Angoff method and the Angoff Y/N method in terms of MPSs shows that the experts' perception of the difficulty of the exam is similar according to these two methods. The fact that these two methods do not differ significantly in terms of MPS averages shows that the MPS averages of the methods are close to each other. The fact that there is no relationship between Angoff-Nedelsky, Angoff Y/N- Nedelsky and Nedelsky in terms of MPSs and that there is a significant difference between the MPS averages of these methods shows that experts' ideas about the structure of the exam have changed while working with the Nedelsky method. The very high level of correlation between the MPS of the Angoff method and the MPS of the Ebel method may be because both methods involve estimating over 100 students at the pass-fail limit. Although there was a high level of correlation between the MPSs of these two methods, the differentiation in terms of MPS averages indicates that the experts perceived the items more easily in one of the methods. It was observed that experts made similar decisions using the Angoff method.

Although there is no relationship between the MPSs of the Ebel method and the MPS of Angoff Y/N and Nedelsky methods, the lack of difference between MPS averages indicates that the perceptions of the experts about the difficulty of the items in the test have changed. However, MPS averages of the methods are close to each other. Since there is a high level of agreement between the MPS of the Angoff and Ebel methods, only one of the methods can be used when the aim is to save time in determining the cut-off point.

The weak correlation between estimated item difficulties based on the Ebel method and actual item difficulties indicates that it is not a correct practice to consider the percentage values given by the experts for cells in the Ebel method on an item basis. A different study could examine whether the number of items in the test and the structure of the test have an impact on the relationship between actual item difficulties and experts' method-based item difficulty estimates. Angoff method is more appropriate to implement when estimating item difficulty in the test development process.

In order to see why the agreement between experts was at a medium level in the Angoff method, the expert forms were examined, and it was seen that one of the experts gave all probability values at a very high level. In cases where two judgments are used, such as the Angoff Y/N method, it has been found that it is more appropriate to check whether there is harmony between

expert decisions. In cases where the Nedelsky method is used, the high level of agreement between expert decisions shows that the more detailed the experts examine the items, the greater the agreement between them. The higher agreement between experts in the Ebel method than in the Angoff method may be because fewer experts are employed in the Ebel method. The effect of the number of items on the harmony between experts can be examined by looking at the harmony between the experts' judgments in the first and last half of the test.

The divergence between norm-based assessment and standard-setting methods results is observed due to the fact that test-centered methods are not affected by student characteristics. Student-centered methods and norm-based assessment results are likely to yield similar results. As seen in the study, if a cut-off score is created without using the standard-setting method in exams that aim to recognize and place students, judging students' level of language skills, the results based on this cut-off score do not make accurate decisions about the students. In exams with high student participation, creating a cut-off score using at least one standard-setting method with a broad group of experts will increase the reliability and validity of the exam criteria.

In light of all these findings, it is seen that it is important to use various standard-setting methods together and keep the expert group-wide when determining the cut-off score in exams where absolute evaluation becomes important. In addition, the test items should be reviewed by looking at which items the expert judgments differ significantly on.

Acknowledgments

This study was conducted by Rümeysa KAYA at Gaziantep University, Institute of Educational Sciences, as a Master's Thesis. It is summarized from the master's thesis that was conducted under the supervision of Bayram ÇETİN.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Gaziantep University, Social and Human Sciences Ethics Committee, 12.10.2022-246368.

Contribution of Authors

The authors contributed equally to all the stages of the study.

Orcid

Rümeysa Kaya https://orcid.org/0000-0003-3212-3032
Bayram Çetin https://orcid.org/0000-0001-5321-8028

REFERENCES

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), Educational measurement (2nd ed.). *American Council on Education.*

Berk, R.A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, *56*(1), 137-172. https://doi.org/10.2307/1170289

Boduroğlu, E. (2017). *Yükseköğretime geçiş sınavının sınıflama tutarlılığının farklı yöntemlerle elde edilen kesme puanlarına göre incelenmesi* [*The study of classification consistency of transition to higher education examination according to the cut-off scores obtained from different methods*] [Master's dissertation, Mersin University]. Higher Education Institution National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Buckhendal, W.C., Smith, W.R., Impara, C.J., & Plake, S.B. (2002). A comprasion of Angoff and Bookmark standard setting methods. *Journal of Educational Measurement, 39*(3), 253-263. https://doi.org/10.1111/j.1745-3984.2002.tb01177.x

Çetin, S. (2011*). İşaretleme ve angoff standart belirleme yöntemlerinin karşılaştırılması* [*Comparison of bookmark and angoff standard setting methods*] [Doctoral dissertation, Hacettepe University]. Higher Education Institution National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Çukadar, İ. (2013). *Norm ve ölçüt dayanaklı değerlendirmelerin karşılaştırılmasına ilişkin bir çalışma* [*A study upon comparison of norm and criterion referenced assessment*] [Master's dissertation, Hacettepe University]. Higher Education Institution National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Demir, O. (2014). *Angoff, nedelsky ve ebel standart belirleme yöntemleri ile belirlenen kesme puanlarının karşılaştırılması* [*A comparison of cutting points determined by angoff, nedelsky and ebel standard setting methods*] [Master's dissertation, Abant İzzet Baysal University]. Higher Education Institution National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Ebel, R.L. (1972). *Essentials of educational measurement*. Englewood Cliffs NJ: Prentice-Hall.

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 dissertation, Hacettepe University]. Higher Education Institution National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

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. https://doi.org/10.1177/01466210022031804

Hambleton, R.K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89–116). Lawrence Erlbaum Associates Publishers.

Hurtz, G.M., & Hertz, N.R. (1999). How many raters should be used for establishing cutoff scores with the Angoff method? a Generalizability Theory Study. *Educational and Psychological Measurement, 59* (6), 885-897. https://doi.org/10.1177/00131649921970233

Impara, J.C., & Plake, B.S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*(4), 353-366. https://doi.org/10.1111/j.1745-3984.1997.tb00523.x

İlhan, M. (2022). Korelasyon [*Correlation*]. In Çetin B. (Ed.), *Eğitimde ölçme ve değerlendirme* [*Measurement and evaluation in education*]. (2nd ed., pp. 23-43). Ani Publishing.

Jaeger, R.M. (1989). Certification of student competence. In R.L. Linn (Ed.), *Educational measurement* (3rd ed. pp 485-514). Macmillan Publishing Co, Inc; American Council on Education.

Kiliç, A. (2018). *Angoff, yes/no ve sınır grup yöntemlerine göre kesme puanlarının karşılaştırılması [Comparison of cutting points by Angoff, yes / no and borderline group methods]* [Master's dissertation, Abant İzzet Baysal University]. Higher Education Institution National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Kaya, R. (2023). *İngilizce B1 seviye atlama sınavının kesme puanının farklı standart belirleme yöntemlerine göre incelenmesi* [*Examination of the cutting score of the English B1 leveling exam according to different standard determination methods*] [Master's dissertation, Gaziantep University]. Higher Education Institution National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Landis, J.R., & Koch, G.G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics, 33*(2), 363-374. https://doi.org/10.2307/2529786

Livingston, S.A., & Zieky, M.J. (1983). A comparative study of standard-setting methods. *ETS Research Report Series*, *1983*(2), i-48. https://doi.org/10.1002/j.2330-8516.1983.tb00038.x

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14*(1), 3-19. https://doi.org/10.1177/001316445401400101

Norcini, J.J. (2003). Setting standards on educational tests. *Medical education, 37*(5), 464-469. https://doi.org/10.1046/j.1365-2923.2003.01495.x

Norcini, J.J., Lipner, R.S., Langdon, L.O., & Strecker, C.A. (1987). A Comparison of Three Variations on a Standard-Setting Method. *Journal of Educational Measurement*, *24*(1), 56–64. https://doi.org/10.1111/j.1745-3984.1987.tb00261.x

Ömür, S., & Selvi, H. (2010). Angoff, Ebel ve Nedelsky yöntemleriyle belirlenen kesme puanlarinin siniflama tutarliliklarinin karşilaştirilmasi. *Journal of Measurement and Evaluation in Education and Psychology, 1*(2), 109-113. https://dergipark.org.tr/tr/download/article-file/65991

Rovai, A., Baker, J., & Ponton, M. (2014). *Social science research design and statistics: a practitioner's guide to research methods and ibm spss analysis.* Chesapeake, VA: Watertree Press LLC.

Şahin, T. (2019). *Nedelsky, sınır grup ve karşıt gruplar standart belirleme yöntemlerinin norma dayalı değerlendirmelerle karşılaştırılması* [*Comparison of nedelsky, borderline group and constrasting groups standard setting models with norm referenced assessment*] [Master's dissertation, Hacettepe University]. Higher Education Institution National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Tanriverdi, S. (2006*). Standart belirleme yöntemlerinin geçme puanları üzerine etkisi* [*Impacts of standard setting methods over passing*] [Master's dissertation, Hacettepe University]. Higher Education Institution National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Taşdelen, G. (2009*). Nedelsky ve Angoff standart belirleme yöntemlerinin genellenebilirlik kuramı ile karşılaştırılmasına ilişkin bir araştırma* [*A comparison of Angoff and Nedelsky cutting score procedures using generalizability theory*] [Master's dissertation, Hacettepe University]. Higher Education Institution National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

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]. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Wang, L., Pan, W., & 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.

Yildirim Kan, N. (2019). *İngilizce hazırlık atlama sınavı için kesme puanı belirlenmesinde standart belirleme yöntemlerinin karşılaştırılması* [*Comparing standard setting methods while determining cut point for English proficiency exam*] [Master's dissertation, Hacettepe University]. Higher Education Institution National Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Zieky, M.J., & Livingston, S.A. (1977). *Basic skills assessment: Manual for setting standards on the Basic Skills Assessment tests*. Educational Testing Service.