A Comparative Study of the Nedelsky and Angoff cut score procedures

Djilali Mezainia[1], Hakim Bouamama[2]

[1]Department of Psychology and Education Sciences, University Djilali Bounaama khemis miliana, Algeria.

[2]Educational Governance Department, The National Institute for Research in Education, Algeria,

hakim.bouamama@inre.dz, https://orcid.org/0000-0002-0909-984X

Received: 11/2023, Published: 01/2024

Abstract:

This study aimed at investigating the validity evidence of both the Angoff and Nedelsky methods, as well as comparing the determined a cut score used to separates two different levels (masters and non-masters of the Mathematics related content). In order to gather data a criterion referenced achievement test was developed. Two samples were selected. The first sample consisted of 12 judges (subject matter experts) whose task was to estimate the performance level of the hypothetical group of "minimally competent" examinees for each item. The second sample consisted of 173 fourth grade middle school students who were randomly selected to answer the test questions. The results revealed that the cut score resulted from the application of the Angoff procedures was significantly higher than when using the Nedelsky procedures. The findings also showed that there was a significant correlation of the propability that the "minimally acceptable person" would answer each item correctly estimated by the expert judges using both methods to the actual difficulty of the test items. Finally, it was concluded that although the Angoff method demonstrated somewhat a higher internal consistency, no significant difference existed between the two methods.

Keywords: Angoff; Nedelsky; criterion-referenced; cut score.

Tob Regul Sci.™ 2023;9(2):2320-2332

DOI: doi.org/10.18001/TRS.9.2.150

Introduction

The process of setting standards and establishing appropriate cut scores and decide "how much is enough" is a critical step in any context that uses test scores to make decisions about to be classification of examinees into two or more categories. This process is the most difficult and controversial step within the procedures of criterion referenced tests development. Kane suggest that "By setting a standard that yields a cut score on a test score scale, we can change a subjective evaluation of a person's performance level in some domain into a simple, objective comparison of

a test score to the cut score "(Kane,2017,12). All standard setting methods are subjective to some extent." They all involve judgments about "how much is enough" or "how much is too much". But once the standard is set, the operational subjectivity is eliminated, or at least, enormously reduced" (Kane, 2017, 12).

The critical step in the development and use of some tests is to establish one or more cut points dividing the score range to partition the distribution of scores into categories. . . . Thus, in some situations, the validity of test score interpretations may hinge on the cut scores (AERA et al., 2014, p. 100). And as Hansche (1998) pointed out "the probably the most challenging problem today in educational assessment concerns setting performance standards on the test score scale to separate students into performance categories (e.g., certifiable and not certifiable)" (Hansche, 1998, 87). This challenge is due to the fact that a true cut score does not need to be discovered through research and studies, but it should be determined through judging methods according to certain criteria and across several sessions. For this purpose, a number of methods were proposed.

Most standard setting methods are consensual and use panels of expert judges to establish a cut score that seperates two or more performance categories. Among the judge-based standard setting methods, Angoff and Nedelsky method have been used in a range of educational settings. Commonly, Angoff and Nedelsky are procedures used to estimate performance standards to 'separate the competent from the non-competent candidate'.

In addition, since setting a cut score is a judge-based process it is likely to include a certain degree of error, which may result in some false positive and false negative decisions (The first one occurs when the non-competent candidate is classified as competent, while the second occurs when the competent candidate is classified as non-competent). Thus, more studies should be conducted to find the procedure that minimizes proficiency classification errors. In addition, since the validity evidence of the judgemental method to set a cut score depends widely on the accuracy of the judges ratings, it seems reasonable and necessary to conduct studies in order to establish a framework that enables experts to reach credible, defensible, and reasonable cut scores and valid decisions.

Most of the studies conducted in this context focused on researching the adequacy of judges, their number and level of training, their characteristics, the amount and type of information they need, and the appropriate number of rounds to reach the final cutoff score. However, "the consistency and appropriateness of judgmental standard setting in education has been repeatedly questioned. Different methods tend to give different results, and there has been no obvious way to choose among the conflicting results" (Kane, 2017, 12).

The internal consistency of the judges ratings is considered one of the most important criteria that reflects the internal validity and adequacy of the cut score. It represents one of the few possibilities for objectifying the judgemental procedures by linking the judges estimates to the reality (Cizek & Bunch, 2007). Preference for one method over another is related to the extent to

which the method yields estimates of item difficulty that are more consistent with the actual performance of the targeted respondents (Chang, 1999). The weak internal consistency of the judges estimates is also one of the main obstacles in most of the judgemental methods.

Kane pointed out that in spite of the number of standard-setting methods and variations on these methods has expanded greatly, along with the range of applications of these methods, the issues involved in developing and validating performance requirements have not changed much. It is still difficult to set standards and even more difficult to validate standards (Kane, 2001).

Moreover, and despite the large number of studies concerned with methods of determining cut scores, the debate is still open about the importance of each method, as there are no clear guidelines to date in choosing the appropriate method. Shoukry points out that studies that attempted to study the validity of the cut-off score are scarce, besides most of these studies focused on comparing the scores resulting from the use of different methods in calculating the cut-off scores (Shoukry, 2006, 7).

Kane (2001) believes that there are three types of evidence of validity evidence (procedural, internal and external). The first type is concerned with the extent to which the methodological steps are followed. The second type is related to the internal data generated from standard setting procedures, with a focus on the consistency of the results. The third type is based on comparing results with external sources.

The current study aims to compare the cut scores resulting from the use of each two different methods (Nedelsky and Angoff), as well as to explore the internal validity evidence of each method. The questions addressed in this research are:

(1) Do the different procedures (Angoff and Nedelsky) yield different cut scores?

(2) To what extent do ratings using each method correlate to the actual difficulty coefficients of the test items?

(3) Is there a significant difference in the internal consistency between the Angoff and Nedelsky methods?

Definition of terms

Setting performance

standards refers to the process of deriving level of performance on a test, and the classification of test takers into performance categories is most commonly operationalized by application of a cut score to performance on a test, and standard setting can involve establishing one or more cut scores on an examination (Cizek,2017).

A cut score

is defined as a selected point on the score scale of a test. This point is used to determine whether a particular test score is sufficient for some purpose. It functions to divide the distribution of test performances into two or more performance categories.

Standard setting methods can be classified according to different taxonomies. Historically, the simplest classification has been that of test-centered versus examinee-centered methods. In the former, panelists make judgments about test questions, whereas in the latter the focus is on the examinees (Pitoniak & Cizek, 2016).

Nedelsky Method

This method, suggested by Leo Nedelsky in 1954, can be used only with multiple-choice tests, since it requires a judgment about each possible wrong answer. The judges' task is to look at the question and identify the options that they believe a hypothetical minimally competent examinee would rule out as incorrect. The reciprocal of the remaining number of options becomes each item's "Nedelsky value." That value is interpreted as the probability that the borderline student will answer the item correctly (Cizek & Bunch, 2007).

Angoff Method

This method, suggested by William H. Angoff in 1971, is similar to Nedelsky method, but it can be used with different types of tests. In this process, each judge estimates the proportion of minimally competent examinees who would give a correct answer to each of the items. Those estimates are then summed across items for each judge, with the average of the sums across judges determining the test cut-score (Shulruf, et.al, 2016).

A two-choice version of the Angoff method, where judges are asked whether a person with minimum competence should be able to answer the item correctly. Possible responses are "yes, «no," and "I don't know." Statistically significant interjudge agreement on "yes" responses determines appropriate minimum competence items. The percentage of those items on the test serves as the initial standard. This standard is then adjusted for measurement errors.

Another modification of Angoff's method took place in 1984, where judges are provided the test item difficulties the intent of this modification is to improve interjudge agreement, then they are asked to adjust their ratings accordinly, and the cut score is the average of their modified estimates.

One of the advantages of the Angoff method is that it is easy to use. (Hambleton & Pitoniak, 2006) .It can be used before applying the test (Kane, 1998). Among its drawbacks is the difficulty in estimating the performance on the items for the group of individuals with minimal adequacy, and the tendency to overestimate the easy items (Hambleton & Pitoniak, 2006).

Method

Participants

Raters

In order to establish the appropriate cut scores using both methods (Angoff and Nedelsky) a group of 12 judges were selected. All the panel of judges selected to participate in the standard setting procedures were experienced in teaching Mathematics fourth grade middle school students, and were familiar with the content.

Test Taker

The test scores were obtained from the fourth grade middle class of 173 students who were randomly selected and available take the test in the content area of Mathematics.

Items: The achievement criterion referenced test in Mathematics used in this study was developed by the researcher. It consisted of 20 items four- option multiple -choice format. Alpha reliability estimates was found to be 0,74. For the test evaluation, one mark (1) for the correct answer and zero (0) for the incorrect one, so the total score will range between 0 and 20.

Procedures

Training on the content-based standard setting process, a psychometrician and a subject matter expert in Mathematics explained the objectives of the meeting. They introduced the Angoff procedure and led the discussion. Judges were first asked to reach common understanding of the concept of hypothetical minimally competent examinee to get familiar with it. Bearing that definition in mind, instructions on the Angoff rating procedures were provided. The panelists, then, were asked to consider a group of 100 minimally competent examinees and estimate how many of them would be able to answer the question correctly. All ratings were collected and the mean of each rater's total judgment scores was calculated. This mean score indicates, in the rater's judgement, the score that a minimally competent examinee would obtain. The standard setting process in this research consisted of only a single round.

A week later the panel of judges participated in a single round to determine a cut score on the criterion referenced test in Mathematics using the Nedelsky method. As the judges were familiar with the concept of the minimally competent examinee, Nedelsky rating procedures were explained. Every rater estimated the number of options that the minimally competent examinee would be able to eliminate as incorrect. All ratings were collected and the mean of each rater's total judgment scores was calculated.

Results and Discussion

Procedures

The first research question addressed the difference in cut scores generated by the same group of judges using different methods (Angoff and Nedelsky). The mean across judges ratings was calculated in both procedures. The judges using the Angoff method produced a mean cut score of 70%, the mean cut scores reached by the panel using this method ranged from low of 66% to high 74%.According to the final cut score produced using the Angoff method, it is assumed that the examinee is assumed to have reached the accepted level of performance if at least 70% of his answers are correct.

The judges using the Nedelsky method produced a mean cut score of 55%, the mean cut scores reached by the panel using this method ranged from low of 50% to high 61%. According to the final cut score produced using the Nedelsky method, it is assumed that the examinee is assumed to have reached the accepted level of performance if at least 55% of his answers are correct. The data are presented in table No. (01)

The Wilcoxon test was used to compare the ratings of judges when the Angoff method is used to determine the cut score and the ratings produced when the Nedelsky method is used. This test is considered as a non-parametric statistical method used to find out the difference between two related samples to the data (Dunn, 2001, 547), and it becomes an alternative to the t-ratio of two related samples when the sample size is small (n) ranging between (6-25). In addition, since the two samples are related and the sample size is small (n = 12), the Wilcoxon test was used and it was calculated by using the statistical package for social sciences (Spss). The results are illustrated in table N°. (2).

Table 1. Shows the results of the Wilcoxon test.

Sig (2-tailed)	P value	Sample Size	Mean Rank	Sum of Ranks
0.01	3.06	12	6.50	78

Table 1. Shows that there was a significant difference between the judges rating using different methods (t=3.06, p> 0.01). This result indicates that the Angoff method set a significantly higher standard than did the Nedelsky method.

The low cut score resulting from the Nidelsky method compared to Angoff method is due the fact that the judges ratings in the Nidelsky method on a four-options test are four fixed values (0.25, 0.33, 0.5 and 1) , which do not allow for giving estimates that range between (0.5) and (1), and accordingly, most judges do not tend to set the probability value (1) or the level of proficiency 100% Shepard, 1984), and thus the judges are forced to set the value (0.5), for most

of the items, which in turn leads to a decrease of the resulting cut score when using the Nedelsky method. (p. 72).

Moreover, it is hypothesized that the Nedelsky method produces a lower cut score because the ratings are influenced by the test level of difficulty. Since the judges' task is to identify the wrong answers that a minimally competent examinee would be able to recognize as wrong, or not the best of the answer. Moreover, the Nedelsky method requires the judges to fully examine the alternatives; this procedure makes them tend to judge the item as difficult due to the similarity of the alternatives and thus sets a lower cut score for the item .In this regard, Burton and Cross (1982, 1978) pointed out , the lower judges rating in the Nidelsky method is not only due to the complexity of the concept to be measured, but also to the plausibility of the distractors (Chang, 1996,17-18). The plausibility of the alternatives may influence the item level of difficulty, which is not a characteristic of the Angoff methods. As for the Angoff method, where the judge does not have to examine the alternatives and evaluates the term as soon as he reads the question, which may lead him to rate the item as easy, which results in a high cut score.

The results of many studies agree with the results of the current study, i.e. The Angoff method produces higher cut-scores compared to the Nedelsky method, and some of these studies are presented in table N°. (3).

Table 2 shows higher cut-scores produced by the Angoff compared to the Nedelsky method.

Studies	Number of Items	Angoff		Nedelsky	
		Cut Score %	Cut Score	Cut Score %	Cut Score
Livingston & zieky (1989)	70	51%	36	44%	31
Allem S .E (1991)	60	71.7%	43	66.77%	40
Chang lei (1996)	09	71%	06	57%	05
Hadjadj .G (2007)	87	79%	69	54%	47
The current Study	20	70%	14	55%	11

Table 2. Shows that the cut scores resulting from the Angoff method in all six studies are higher compared to the Nedelsky method. The highest value of the five cut scores resulting from the implementation of the Angoff method was 79% compared to 54% when the Nedelsky method was used for the same test. The lowest value of cut scores in Angoff method was 51% compared to 44% when the Nedelsky method was used. The mean cut score resulting from the Angoff method when comparing the five studies was 68%, and 55% in the Nidelsky method.

The second research question addressed the extent to which the judges ratings of the test items in both methods correlate to the actual items difficulty. Data were gathered and processed through the statistical package for social sciences (SPSS). (See table N° 4).

Table 3. Shows the mean cut score for the test items and their actual difficulty coefficients.

Items	Item Difficulty	Ratings		items	Item Difficulty	Ratings	
		Angoff	Nedelsky			Angoff	Nedelsky
01	0.98	0.88	0.71	11	0.75	0.57	0.43
02	0.72	0.62	0.47	12	0.59	0.69	0.58
03	0.93	0.85	0.71	13	0.39	0.40	0.40
04	0.57	0.65	0.50	14	0.60	0.58	0.58
05	0.77	0.81	0.54	15	0.58	0.65	0.51
06	0.86	0.84	0.71	16	0.44	0.69	0.42
07	0.88	0.78	0.67	17	0.81	0.72	0.67
08	0.94	0.76	0.67	18	0.72	0.74	0.46
09	0.88	0.78	0.54	19	0.66	0.71	0.44
10	0.80	0.68	0.54	20	0.82	0.60	0.43

The value of the correlation coefficient of item ratings using Angoff method and the actual item difficulty was 0.72 for the Angoff method and 0.69 for the Nedelsky method, which indicates a strong and positive correlation between the judges estimates and item difficulty coefficients. Correlation results support the validity of the two methods.

The test of significance proved the existence of significant correlations for both methods .For the Angoff method ($t=5.89$, $P > 0.05$) and ($t=5.59$, $P > 0.05$) for the Nedelsky method.

And among the studies that support the current result is the findings of the study (Al-Shuraim and Sawalma, 2006), which indicated that the Pearson correlation coefficient of the judges ratings using the Angoff method and the actual difficulty coefficients of the test items was 0.57 when information were not available . While, the correlation coefficient reached 0.98 when they were provided with them. As for the Nedelsky method, the correlation coefficient reached 0.77 when judges were provided with item difficulty and 0.52 in the absence of information about the item difficulty. Those coefficients were significant at (P > 0.05).

Based on these results, we can conclude that the results reflect the importance of the two methods in providing estimates that are directly proportional to the actual difficulty coefficients, which supports the validity evidence of the resulting cut score and consequently increases the predictive ability of the two methods.

According to research literature, the justification for this result is that the estimates of the judges using the Angoff method give free estimates that may range from (0) to (1), meaning that the judge can give any percentage that can correspond to the actual difficulty of the item. As for the rating model using the Nedelsky method on a four-alternative test, judges find four fixed probability values to choose from (0.25, 0.33, 0.5, 1), and therefore most judges do not tend to choose the probability value (1) and choose the probability value (0.5).

Unlike the Angoff procedures, the Nedelsky rating procedures do not permit to a judge to to provide an estimate between (0.5 and 1) with the four-option multiple-choice test, which results in providing an estimate a lower or higher than the judge expected rating. This indicates that the rating scale of the Needelsky method forces the judge to be inconsistent to a certain extent.

Through the third research question the difference in the internal consistency of the two methods was researched .In order to test the difference in the internal consistency of the two methods the researcher followed the same procedures suggested by Chang Lei (1996) in calculating the internal consistency, where the variance was estimated to the mean ratings for all the items resulting from the application of the Angoff and Nedelsky methods, and the actual item difficulty coefficients estimated on the basis of the minimally competent examinees (Chang,1996,4)

Through calculating the variance used to estimate the internal consistency of the ratings for each method with the actual test difficulty coefficients, it was noted that the variance value for the Angoff method was 0.028 and 0.048 for the Nedelsky method, and the square root of these variances, reached 0.17 for the Angoff method, and 0.22 for the Nedelsky method, which means that the deviation of the cut scores from the actual difficulty values was 3.4 or 17% of the total items for the Angoff method and 4.4 or 22% of the total items for the Nedelsky method.

Whereas, the lower the value of the standard deviation of the judges estimates, the more this indicates a high internal consistency. It is noted from the results that the value of the standard

deviation of the judges' estimates when using the Angoff method was lower compared to the Nedelsky method, which indicates a decrease in the internal consistency of the Nedelsky method compared to the Angoff method.

In order to test the difference in internal consitency of the two methods (Angoff & Nedelsky) a comparison was made between the two-variance values for the two methods by dividing the larger variance by the smaller variance (Chang, 1996, 12). To obtain the variance ratio. The calculated F value was (1.71).

It is smaller than the tabulated value (2.15) at the significance level of 0.05. Hence, no significant difference exists in the internal consistency of the estimates between the two methods.

Table 4. Shows data about ratings and minimally competent examinees performance.

Items	Item Difficulty	Ratings		Items	Item Difficulty	Ratings	
		Angoff	Nedelsky			Angoff	Nedelsky
01	0.98	1,00	1,00	11	0.75	0,80	0,73
02	0.72	0,75	0,55	12	0.59	0,55	0,27
03	0.93	1,00	0,82	13	0.39	0,30	0,27
04	0.57	0,50	0,18	14	0.60	0,45	0,36
05	0.77	0,75	0,36	15	0.58	0,65	0,55
06	0.86	0 ,80	0,64	16	0.44	0,50	0,10
07	0.88	0,80	0,55	17	0.81	0,75	0,45
08	0.94	0,95	0,91	18	0.72	0,65	0,45
09	0.88	0,85	1,00	19	0.66	0,60	0,36
10	0.80	0,85	0,82	20	0.82	0,80	0,73

It is noted in table no. (4) That the highest value of the ratings with the Nedelsky method is 0.71, while the percentage of the actual difficulty coefficient values for the minimally competent examinee in the Nedelsky method, which exceeds 0.71, was 35 %. Besides, the lowest value of the estimate in the Nedelsky method is 0.25, meaning that the judge is not allowed to give an estimate less than this value, (see table N°5). That the value of the actual difficulty coefficient for item No. (04) Is 0.18, which It increases the degree of the discrepancy between the ratings and

the item difficulty coefficients for the examinee with the minimum level of proficiency which may influence the method's internal validity.

As for the Angoff method, it is noted that the percentage of the judges estimates that exceed 0.50 reached 95 %, and the percentage of actual difficulty coefficients for the examinee with the minimum level of competency reached 90 %, as shown in table N° (5).

The second factor that might have caused a lower internal consistency of the Nedelsky method compared to the Angoff method was the judge experience associated with the test content. A number of studies indicated that the judge's experience, the nature and duration of training he received on the process of determining the cut score are among the important factors affecting the value of internal consistency. In addition, that these factors allow them to predict the performance of the minimally competent examinee.

Conclusion

This study aimed to compare the methods of Angoff and Nedelsky to estimate the cut score on a reference-based test in mathematics for the fourth grade middle level class, and to investigate the underlying validity inferences of each method.

In order to achieve the study objectives, an achievement criterion- referenced test in Mathematics was developed .The test consisted of twenty multiple-choice items, each with four alternatives. The study included two types of samples. The first is a sample of judges, and it consisted of a group of twelve experienced teachers whose role was to estimate the test items in question using two standard setting procedures (Angoff and Nedelsky) in order to establish a cut score for the test. While the second sample consisted of 173 student randomly selected to answer the test questions.

The use of the two methods resulted in two different cut scores, as the cut score resulting from the application of Angoff procedures was (0.70) or (70%), and it was higher than the cut score resulting from the application of the Nedelsky procedures, which was (0.55) or (50%) .

The results of the non-parametric analysis (Wilcoxon test) indicated that there is a statistically significant difference at the significance level of 0.01 between the two cut scores resulting from the two methods.

The results also revealed that the judges ratings in both methods correlate significantly to the actual items difficulty level, which supports the validity of both methods. In addition, to make sure that there is a difference in internal consistency between the Angoff and Needelsky methods. The results indicated that there was no difference in the internal consistency between the two methods, except that the Angoff method was characterized by higher internal consistency, and this is due to many factors mainly the nature and characteristics of each method and the judges experience. However, although Angoff method has greater internal consistency compared to

Nedelsky method, the results show that the two methods have acceptable internal consistency, which is evidence supporting the validity of the two methods. However, this result remains related to the nature of the test and the nature of the subjects.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

References

[1] Allam, S.D, M. (2007). Diagnostic Tests as a Reference Test in the Educational, Psychological and Training Fields, *Cairo, Dar Al-Fikr Al-Arabi*.

[2] Al-Shuraim, A. & Sawalmeh, Y. (2006). Determining the cut-off score for the referenced test in mathematics using the "Angov" and "Nedelsky" models. A comparative study of knowing the difficulty of the paragraphs and not knowing them. *The Jordanian Journal of Educational Sciences*, 2(1), 1 – 10.

[3] American Educational Research Association. (2014). American Psychological Association & National Council on Measurement in Education. Standards for educational and psychological testing. Washington, DC: *American Educational Research Association*.

[4] Berk, R. A. (1986) .A Consumer's Guide to Setting Performance Standards on Criterion Referenced Tests. *Review of Educational Research*, 56, 137 – 172.

[5] Berk, R. A. (1986) . A Consumer's Guide to Setting Performance Standards on Criterion Referenced Tests. *Review of Educational Research,* 56, 137 – 172.

[6] Chang, L. (1999). Judgemental item analysis of the Nedelsky and Angoff standard-setting methods. *Applied Measurement in Education,* 12, 151–165.

[7] Chang, L.A .(1996). Comparison between the Nedelsky and Angoff Standard Setting Methods ? Paper presented At the Annual Meeting of The National Council on Measurement in Education, *New York*.

[8] Cizek, G J. Bunch. B, M .(2007). Standard setting: A guide to establishing and evaluating performance standards on tests." *Sage Publications, Inc*. London.

[9] Demauro, G & Powers, D.(1990). Logical consistency of The Angoff Method of Standard Setting .,*Paper Presented at The Annual Meeting of The National Council on Measurement in Education.* Boston, MA,17-19 April.

[10] Dunn, D.S .(2001). Statistics and Data Analysis for the Behavioral Sciences. *M.C. Graw-Hill*.New York. NY.

[11] Ghanem, H. .(2007). buhuth mueasarat qi alqias alnafsii w eilm alnafs altarbawi, *ta1 alqahirat ealam alkutub.*

[12] Goodwin, L, .(1996). Focus on Quantitative Methods Determining Cut – Off Scores. *Research in Nursing & Health*.19 (1), 857 – 872.

[13] Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards". In R. L. Brennan (Ed.) Educational Measurement (4th Ed.). *Westport:* American Council on Education & Praeger Publishers.

[14] Hansche, L .(1998). Handbook for the Development of Performance Standards: *Meeting the Requirements of Title I.* Washington, US Department of Education and the Council of Chief State School Officers.

[15] Jaeger, R. M. .(1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). Washington, DC: American Council on Education.

[16] Kane, M. (1998). Choosing between examinee-centred and test-centred standard-setting methods. *Educational Assessment,* 5(3), 129 –145.

[17] Kane, M. (1994). Validating the performance standards associated with_passing scores. *Review of Educational Research,* 64(3), 425–461.

[18] Kane, M. T. (2017). Using empirical results to validate standards. In S. Blömeke, & J. Gustafsson (Eds.), Standard setting in education: *The Nordic countries in an international perspective*, (pp. 11–29). Springer.

[19] Livingston, S. A., & Zieky, M. J. (1983). A Comparative Study of Standard-Setting Methods. *Educational Testing Service, Princeton, N.J.*

[20] Pitoniak, M. J., & Cizek, G. J. (2016). Standard setting. In C. S. Wells & M. Faulkner-Bond (Eds.), *Educational measurement*: From foundations to future (pp. 38–61). The Guilford Press.

[21] Shoukry, S. (2006). faeiliat baed turuq taqdir darajat alfasl faa altanabuw bialtahsil allaahiq faa alhandasat lilmarhalat al'iiedadia «risalat majistir ghayr manshurat, kuliyat altarbiat, jamieat alminya

[22] Shulruf, B., Wilkinson, T., Weller, J., Jones, P., & Poole, P. (2016). Insights into the Angoff method: results from a simulation study. BMC Med Educ, 16, 134.

[23] 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, IL.