International Journal of Assessment Tools in Education
2021, Vol. 8, No. 1, 9–20
https://doi.org/10.21449/ijate.747105
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

An Evaluation of Pass/Fail Decisions through Norm- and Criterion-Referenced Assessments

Ismail Cuhadar[1,*], **Selahattin Gelbal**[2]

[1]Ministry of National Education, General Directorate of Measurement, Evaluation and Examination Services, Ankara, Turkey

[2]Hacettepe University, Faculty of Education, Department of Educational Sciences, Ankara, Turkey

ARTICLE HISTORY

Received: June 02, 2020

Revised: Dec. 06, 2020

Accepted: Jan. 04, 2021

KEYWORDS

Norm-referenced Assessment,

Criterion-referenced Assessment,

Standard-setting,

Angoff Method,

Nedelsky Method.

Abstract: The institutions in education use various assessment methods to decide on the proficiency levels of students in a particular construct. This study investigated whether the decisions differed based on the type of assessment: norm- and criterion-referenced assessment. An achievement test with 20 multiple-choice items was administered to 107 students in guidance and psychological counseling department to assess their mastery in the course of measurement and evaluation. First, the raw scores were transformed into T-scores for the decisions from norm-referenced assessments. Two decisions were made to classify students as passed/failed comparing each student's T-score with two common cutoffs in education: 50 and 60. Second, two standard-setting methods (i.e., Angoff and Nedelsky) were conducted to get two cut scores for the criterion-referenced assessment with the help of experts in measurement and evaluation. Two more decisions were made on the classification of students into pass/fail group by comparing the raw scores and the cut scores from two standard-setting methods. The proportions of students in pass/fail categories were found to be statistically different across each pair of four decisions from norm- and criterion-referenced assessments. Cohen's Kappa showed that the decisions based on Nedelsky method indicated a moderate agreement with the pass/fail decisions from the students' semester scores in measurement and evaluation while the other three decisions showed a lower agreement. Therefore, the criterion-referenced assessment with Nedelsky method might be considered in making pass/fail decisions in education due to its criterion validity from the agreement with the semester scores.

1. INTRODUCTION

Educational institutions make high-stakes decisions about students to determine who has mastered the objectives of a course, who will be promoted to the upper grades or who will be selected to a particular school. Because a false decision can cause some problems in reaching the next level objectives in education, educational institutions should be careful in making these decisions. For example, students' learning should be assessed carefully in secondary school due to the fact that it may have impact on the dropout rates in higher education (Paura & Arhipova, 2014; Vossensteyn et al., 2015). Therefore, it is important to build appropriate decision

CONTACT: Ismail CUHADAR ismail.cuhadar@gmail.com Ministry of National Education, General Directorate of Measurement, Evaluation and Examination Services, Ankara, Turkey

ISSN-e: 2148-7456 /© IJATE 2021

mechanism to minimize these false decisions. A valid assessment method is one required tool for the appropriateness of the decision-making mechanism in education (Aiken, 2000; Crocker & Algina, 2008).

Assessment is a decision-making process that involves the comparison of the measurement results with a criterion (Baykul, 2010; Turgut & Baykul, 2010). In other words, assessment is composed of three items: measurement results, criterion, and decision. Accordingly, a decision is made on the measured construct by comparing the measurement results with a criterion. For this reason, the criterion is required to be defined in order to evaluate the measurement results. Based on the type of the criterion in the assessment process, assessment is grouped in two categories: norm-referenced assessment and criterion-referenced assessment (Murphy & Davidshofer, 2005).

Norm-referenced assessment is a formal evaluation process where the performance of a student is compared with the performance of a scientifically selected group of test takers (Crocker & Algina, 2008; Kline, 2000). In the norm-referenced assessment, the performance of the selected group of test takers is the criterion for making the decision on the measurement results of students. On the other hand, a student's score is compared with a predetermined level of performance in the criterion-referenced assessment, and this particular level of performance can be determined using a specific ability set or knowledge area in an educational assessment (Cizek & Bunch, 2007; Murphy & Davidshofer, 2005). In the criterion-referenced assessment, one approach for establishing the criterion is to determine a cut score through the standard setting methods based on the subject matter experts' judgements (Cizek & Bunch, 2007; Crocker & Algina, 2008; Murphy & Davidshofer, 2005; Urbina, 2004).

Standard setting is a process to identify a number that separates different performance levels (Cizek, 1993). There are several standard setting methods in the literature. Jaeger (1989) categorized the standard setting methods in two groups: test-centered and examinee-centered methods. Nedelsky method (Nedelsky, 1954), Angoff method (Angoff, 1971), Ebel method (Ebel, 1972), and Yes/No method (Impara & Plake, 1997) are some examples of the test-centered standard setting methods while borderline group method (Zieky & Livingston, 1977) and contrasting groups method (Berk, 1976) are some examples of examinee-centered standard setting methods. Due to the convenience for the multiple-choice test items, the easiness to administer to the subject matter experts, and the popularity in literature and practice, Angoff and Nedelsky methods were used in the current study to identify the cut scores for classifying student into the performance levels (Cizek, 1993; Murphy & Davidshofer, 2005). These two methods were the basis of the criterion-referenced assessment, and briefly introduced in the next sections.

1.1. Angoff Method

Because Angoff method (1971) is a convenient procedure for the tests with the multiple-choice items, it is commonly used in the practice, including license and certificate programs (Cizek & Bunch, 2007). The first step in the application of Angoff method is to define the minimum qualification level for being categorized in the particular performance level with respect to the test purpose (Livingston & Zieky, 1989). Then, the subject matter experts determine the probability that each item can be answered correctly by examinees with this minimum qualification level. An average probability is obtained across all subject matter experts for each item in the test. The sum of these probabilities from each item corresponds to the cut score for the test. This process can be expressed as a formula using Equation 1 to obtain the cut score of a test via Angoff method.

$$Cut\ Score_{Angoff} = \frac{\sum_{j=1}^{R}\sum_{i=1}^{K}(p_{ij})}{R} \quad (1)$$

In Equation 1; *R* indicates the number of subject matter experts, *K* indicates the number of items in a test, and p_{ij} is the probability determined by expert *j* to item *i*.

1.2. Nedelsky Method

Because it is easy to apply the method proposed by Nedelsky in 1954, Nedelsky method is still in practice, and it was one of the methods that accelerated the transition from the norm-based performance level decisions to the assessment type showing examinees' true performance levels (Cizek & Bunch, 2007). The number of distractors in each item is important in determining the cut score using this method (Arrasmith, 1986). The subject matter experts determine how many distractors the minimum qualified examinees can eliminate in each test item taking the measurement construct into consideration. Accordingly, the number of options that the minimum qualified examinees cannot eliminate is determined for each test item. A probability of correct response is obtained considering the number of remaining options. The probabilities across all items are summed to get a cut score for each subject matter expert. The average of these cut scores across all experts indicates the cut score of the test by Nedelsky method. This process can be expressed as a formula using Equation 2 to obtain the cut score of a test via Nedelsky method.

$$Cut\ Score_{Nedelsky} = \frac{\sum_{j=1}^{R}\sum_{i=1}^{K}(d_i - e_{ij})^{-1}}{R} \quad (2)$$

In Equation 2; *R* indicates the number of subject matter experts, *K* indicates the number of items in a test, d_i is the number of options in item *i*, and e_{ij} is the number of distractors eliminated in item *i* by expert *j*.

1.3. Which Assessment Type?

The type of assessment depends on how measurement results are intended to be used. When the measurement results are used for the selection and placement purpose, the norm-referenced assessment is advantageous over the criterion-referenced assessment (McCauley & Swisher, 1984). However, the decisions from the norm-referenced assessment do not correspond to the true ability level in the target construct of the measurement tool (Johnson & Martin, 1980). For this reason, the norm-referenced assessment is open to misuse in evaluating examinees' performance levels and the effectiveness of a program (McCauley & Swisher, 1984). On the other hand, the criterion-referenced assessment is very useful in determining the examinees' performance levels and replanning curriculum based on the identified needs of the examinees from the criterion-referenced assessment (Freeman & Miller, 2001). Accordingly, the type of assessment that needs to be used in making decisions depends on the purpose of a measurement.

The goal of education is to provide intentional and sustainable changes in students' behavior through a curriculum and based on the objectives of that educational institution (Ertürk, 1998; Tyler, 2013). When the assessment types are reviewed in education and practice, it is seen that different approaches are taken for the similar educational goals. For example, the criterion-referenced assessment is used in the primary education and secondary education while the assessment type differs across the universities in the higher education, although the aforementioned goal of the education is the similar across all levels in the education system (e.g., the minimum scores for being evaluated as successful are 45 and 50 out of 100 in the primary and secondary education in Turkey, respectively; Milli Eğitim Bakanlığı, 2014, 2016). The assessment type is not consistent within the university among the departments, and either the norm- or criterion-referenced assessment can be chosen for evaluating student achievement in some universities (e.g., Akdeniz University, 2017; Ankara University, 2018; Erciyes University, 2015; Sakarya University, 2019). Furthermore, the passing grade is not consistent across the universities (e.g., 50 in Sakarya University, 60 in Ankara University). Thus, a score of 55 is considered insufficient to pass a course in some universities, but the same score means

a sufficient score in the others. In other words, the same score can result in pass or fail decision based on the assessment procedure in the educational institutions. Accordingly, it is important to determine which assessment procedure provides more valid decisions for which situations. Otherwise, the pass/fail decisions can be incorrect or inappropriate, and the incorrect decisions can cause problems in reaching the next level objectives of the curriculum (e.g., Paura & Arhipova, 2014; Vossensteyn et al., 2015).

There are studies in the literature for comparing the norm- and criterion-referenced assessments over different tests (e.g., Mohr, 2006; Oescher, Kirby, & Paradise, 1992; Pester, 2003; Visintainer, 2002). In addition, a few studies investigated the differences and the similarities in the decisions from these two assessment types (e.g., Jacobson, 2008; Nartgün, 2007; Toprakçı, Baydemir, Koçak, & Akkuş, 2007). However, the standard setting methods used in the criterion-referenced assessments were not compared with the norm-referenced assessments in these studies. Furthermore, two assessment types have not been investigated using the same test for decision making. Accordingly, this study purports to compare the decisions on the same group of examinees from the same test with two different assessment procedures: the norm-referenced assessment and the standard setting-based criterion-referenced assessment.

It is not only important to test the differences in the pass/fail decisions from the norm- and criterion referenced assessments, but also to investigate which assessment type produces more valid decisions under which conditions. The criterion validity might be used to investigate the validity of decisions from the norm- and criterion-referenced assessments (see Aiken, 2000; Baykul, 2010; Kline, 2000; Montgomery & Connolly, 1987; Murphy & Davidshofer, 2005; Turgut & Baykul, 2010 for more information about validity). Despite several studies comparing the two assessment types (e.g., Jacobson, 2008; Nartgün, 2007), the criterion validity of the decisions based on two assessment types has not yet been investigated. Therefore, another purpose of this study is to investigate the criterion validity of the decisions from the norm- and criterion-referenced assessments. Based on two purposes of the study, two research questions were tested: a) “Is there a significant difference between the student-passing rates from the norm- and criterion-referenced assessments?”, and b) “How is the criterion validity of the decisions from the norm- and criterion-referenced assessments?”.

2. METHOD

2.1. Participants

Because the purpose was to compare the assessment types, and the findings were not generalized to a population, there was no sampling procedure in the current study. Accordingly, a purposive study group was chosen that fits the goal of the study. The study group was composed of the second-grade students studying in the guidance and psychological counseling department of Kayseri Erciyes University in Turkey. The fact that these students took a measurement and evaluation course, and the achievement test was designed to measure this content area were the reasons for the selection of them in the current study. In addition, some experts participated in the study for the application of the standard setting methods. These experts had at least a master’s degree in the measurement and evaluation field. In total, there were 107 students from the guidance and psychological counseling department, and there were 11 and 10 experts for the application of Angoff and Nedelsky methods, respectively.

2.2. Procedure and Instrument

In the study, the data were collected in three steps. First, a test was administered to the guidance and psychological counseling students to measure their achievements in the measurement and evaluation course. For this reason, an achievement test with the multiple-choice items was constructed considering the content of the measurement and evaluation course. After the items

were reviewed by two experts in the measurement and evaluation field, a test form with 36 items was obtained. A pilot study was conducted to investigate the statistical characteristics of the items. Then, a final test form composed of 20 items with five options in each was obtained considering the test content, item difficulties, and item discriminations from the pilot study. The data from the final test administration showed that the item discrimination indices ranged between 0.33 and 0.80, and the item difficulty indices ranged between 0.11 and 0.85. In addition, the internal consistency reliability based on Kuder-Richardson formula 20 (KR-20) was equal to 0.71. The second step of the data collection process involved using the experts' opinions to calculate the cut scores based on Angoff and Nedelsky methods. Angoff method was the first application for obtaining the cut scores, and Nedelsky method was administered one week after the application of Angoff method. At the last step of the data collection process, the pass/fail decisions for the students from the measurement and evaluation course at the end of the semester was gathered so that these decisions can be used as a criterion to examine the validity of the norm- and criterion-referenced assessment procedures in the current study.

2.3. Data Analysis

All data analyses were conducted using SPSS 18 (SPSS Inc., 2009) and Microsoft Excel (2013). First, Kendall's coefficient of concordance (i.e., Kendall's W; Kendall & Smith, 1939) was used to examine the agreement among the experts in the standard setting methods. Then, two decisions for each student on their achievements were made as "pass" or "fail" comparing their raw scores with the cut scores from Angoff and Nedelsky methods. In this way, two decisions based on the criterion-referenced assessment were obtained: one from Angoff method and one from Nedelsky method. For the norm-referenced assessment, the raw scores (i.e., the number of correct responses) were first transformed into T-scores (see Sönmez & Alacapınar, 2011). Two more decisions were made on the classification of students into pass and fail categories comparing the T-scores with two passing scores: 50 and 60. These two passing scores were chosen since they are commonly used in the assessment of the students' achievements (e.g., Ankara University, 2018; Sakarya University, 2019). At the end of whole process, there were four decisions for each student on their classifications into passing/failing groups: two decisions from the norm-referenced assessments (i.e., when 50 and 60 were the passing scores in T-score scale) and two decisions from the criterion-referenced assessments (i.e., when two cut scores from Angoff and Nedelsky methods were applied in the raw-score scale).

For the first research question, z-test was used to test whether the decisions based on the four methods in the study statistically differ. Z-test is used to analyze the statistical difference between two proportions from the same group of examinees (Calmorin & Calmorin, 2007). Z-statistic is calculated through dividing the observed proportion difference between the variables by its standard error, as seen in Equation 3 (Jekel, 2007).

Z-test Proportions

		Method II		
		Pass	Fail	
Method I	Pass	*a*	*b*	p_1
	Fail	*c*	*d*	q_1
		p_2	q_2	1.00

$$z = \frac{p_1 - p_2}{\sqrt{\frac{b+c}{N}}} \tag{3}$$

In Equation 3, *N* indicates the number of examinees; *a* indicates the proportion of examinees who pass from both methods; *b* indicates the proportion of examinees who pass from Method I, but fail from Method II; *c* indicates the proportion of examinees who pass from Method II, but fail from Method I; *d* indicates the proportion of examinees who fail from both methods;

$$p_1 = a + b;\ q_1 = 1 - p_1;\ p_2 = a + c;\ \text{and}\ q_2 = 1 - p_2.$$

For the second research question, Cohen's Kappa (1960) was used to investigate the criterion validity of the decisions based on the norm- and criterion-referenced assessments by determining the agreement between the pass/fail decisions from the four methods and the pass/fail decisions from the students' semester scores. Cohen's Kappa statistic is used to determine the level of agreement between two categorical variables correcting the agreement rates by chance (Clark-Carter, 2005). The level of agreement based on Cohen's Kappa can be considered as poor for the values < 0.2; fair between .2 and 0.4; moderate between 0.4 and 0.6; good between 0.6 and 0.8; and perfect between 0.8 and 1 (Cohen, 1960; Landis & Koch, 1977; McHugh, 2012). Cohen's Kappa can be calculated using Equation 4 (Cohen, 1960).

$$\kappa = \frac{Sum\ f_o - Sum\ f_e}{N - Sum\ f_e} \quad (4)$$

In Equation 4; κ is Cohen's Kappa coefficient, $Sum\ f_o$ indicates the sum of observed frequencies in agreement between the methods, $Sum\ f_e$ indicates the sum of expected frequencies in agreement between the methods, and *N* is the number of examinees.

3. RESULT / FINDINGS

Before determining the cut scores from Angoff and Nedelsky methods, the agreement among the experts was examined. Kendall's W indicated a statistically significant agreement among 11 experts in the application of Angoff method ($W = 0.45$, $p < 0.01$). Similarly, a statistically significant agreement among 10 experts in the application of Nedelsky method was found ($W = 0.44$, $p < 0.01$). The cut scores across the experts ranged between 9.90 and 17.15 with an average of 13.20 in Angoff method, and between 5.28 and 15.58 with an average of 8.52 in Nedelsky method. Accordingly, the final cut scores of the achievement test was 13.20 and 8.52 based on Angoff and Nedelsky methods through the criterion-referenced assessments, respectively.

The results of the four methods for determining the passing and failing students from the achievement test in the current study was presented in Table 1. When the norm-referenced assessment was used with a cut score of 50 and 60 in T-score scale, 47% and 16% of the students passed the achievement test, respectively. For the criterion-referenced assessments, 23% of the students passed the test from Angoff method while it was 78% when the Nedelsky method was used to determine the cut score of the test. Accordingly, the minimum percent of passing students was from the norm-referenced assessment with a cut score of 60 in T-score scale, and the maximum percent of passing students was from the criterion-referenced assessment with Nedelsky method being the standard setting method. These two methods produced 62% gap with respect to the students classified as pass from the achievement test.

Table 1. *The Cut Scores, The Number of Passing Students, and The Proportion of Passing Students across The Four Methods (n = 107)*

Method	Assessment	Cut Score	Number of Passing	Proportion of Passing (%)
Angoff	Criterion-referenced	13.20	25	23
Nedelsky	Criterion-referenced	8.52	83	78
T-score	Norm-referenced	50.00	50	47
T-score	Norm-referenced	60.00	17	16

Z-test indicated that the proportion of passing students differed statistically among each pair of the four methods in the study at α = 0.01 (i.e., z = 2.83 for the proportion difference between Angoff method and T-score of 60; z = -5.00 for the proportion difference between Angoff method and T-score of 50; z = 8.12 for the proportion difference between Nedelsky method and T-score of 60; z = 5.74 for the proportion difference between Nedelsky method and T-score of 50; z = -5.74 for the proportion difference between T-score of 50 and T-score of 60; z = 7.62 for the proportion difference between Angoff and Nedelsky methods). Accordingly, the passing rates depends on the chosen method, and a student can be classified into pass or fail category based on which method is applied in the assessment procedure. Therefore, it is important to determine which method produces more valid decisions among the four methods in the study.

The agreement between the four assessment procedures and the students' semester scores in classifying the students into pass/fail categories was presented in Table 2. When the norm-referenced assessment was used with a rule of 60 to pass the test, 36% of the pass/fail decisions was in agreement with the decision from the students' semester scores. When the passing score was 50 rather than 60 in the norm-referenced assessment, the level of agreement with the semester decisions went up to 61%. The former rule produced a poor agreement (κ = 0.09), and the agreement was fair from the later rule (κ = 0.24) in the norm-referenced assessments when the agreement was corrected by chance. For the criterion-referenced assessments, the percent agreement between the Angoff method and the external criterion was equal to 43% with a poor agreement based on Kappa value of 0.14. Among the four methods in the study, Nedelsky method produced the decisions with the highest agreement with the pass/fail categories from the students' semester scores. Nedelsky method and the semester scores resulted in classifying 81% of the students into the same category. In addition, Cohen's Kappa indicated a moderate agreement (κ = 0.41) between these two ways to categorize students into passing and failing groups. As a result, it was found that Nedelsky method, which is the procedure under the criterion-referenced assessment, provided the best decisions in classifying students into pass/fail categories with respect to the criterion validity.

Table 2. *The Agreement between the Pass/Fail Decisions from the Four Assessment Procedures and the Semester Scores (n=107)*

Method	Assessment	Cut Score	Frequency of Agreement	Percent of Agreement	Kappa
Angoff	Criterion-referenced	13.20	46	43	0.14
Nedelsky	Criterion-referenced	8.52	86	81	0.41
T-score	Norm-referenced	50.00	65	61	0.24
T-score	Norm-referenced	60.00	38	36	0.09

4. DISCUSSION and CONCLUSION

Norm- and criterion-referenced assessments are two major procedures in the assessment of student skills in education. Although which assessment type needs to be used depends on their advantages and disadvantages for different measurement situations, the two assessment types are sometimes used for the same measurement goals. Accordingly, the current study investigated the differences in the pass/fail decisions from the norm- and criterion-referenced assessments, and the criterion validity of the two assessment procedures. Under the norm-referenced assessments, two decisions were made on the classification of students into passing/failing groups using two cut scores: 50 and 60 in the T-score scale. Angoff and Nedelsky methods were used to determine two more cut scores in the raw score metric for categorizing students into passing/failing groups by the criterion-referenced assessment.

The findings indicated that all four methods produced statistically different rates of passing students from the achievement test. Accordingly, a different percent of students might pass a test depending on the type of assessment: norm- or criterion-referenced assessment. This difference between the norm- and criterion-referenced assessments is in line with the findings from Nartgün (2007), but inconsistent with the results in Oescher et al. (1992). The difference might have resulted from using a different subject area with two different tests (one test based on norm-referenced assessment and another test based on criterion-referenced assessment) in the study by Oescher et al. (1992). Nedelsky method provided the highest passing rate (i.e., 78%) among the four methods in the study, which is in line with the findings in Çetin and Gelbal (2010). The lowest percent of passing (16%) resulted from the norm-referenced assessment with a cut score of 60 in T-score scale. This result is understandable since the cut score is one standard deviation above the mean in T-score scale. Therefore, approximately 84% of students are expected to have a lower score than the cutoff in this norm-referenced procedure.

The analyses for investigating the criterion validity of the four methods in the norm- and criterion-referenced assessments indicated different agreement rates between the four methods and the external criterion (i.e., the pass/fail decisions from the students' semester scores). Nedelsky method provided the most valid decisions on the students' achievement groups with respect to the external criterion considering the agreement rates and Cohen's Kappa values. The reason for the consistency between Nedelsky method and the semester scores might be the high success of the students from the exams and projects in the measurement and evaluation course, and the relatively low cut score from Nedelsky method for the achievement test in the current study. However, unlike Nedelsky method, other three methods (i.e., Angoff method, T-score of 50, and T-score of 60) used a harder cut score to pass from the test, and so more students failed from these three assessment procedures causing poor to fair agreement rates with the decisions from the semester scores. This finding is not in line with the results in Jacobson (2008), where both norm- and criterion-referenced assessments were good at classifying examinees into two performance levels. Jacobson (2008) investigated the two assessment procedures in a different subject area and used one test per assessment. The difference in the findings might be attributed to the number of tests and the content of the tests in the studies.

Because the percentage of students classified in the passing performance level depended on the type of assessment in the current study, it is recommended that the educational institutions determine the assessment procedure based on their assessment purpose. When the purpose of the assessment is to determine the performance level of examinees or the proficiency in a construct, the criterion-referenced assessment is recommended. Accordingly, Nedelsky method can be used in determining how much is enough to pass from a course or curriculum considering the criterion validity of the method in the current study. However, the limitations of the current study need to be taken into consideration before generalizing the results into the other settings. For example, the course of measurement and evaluation was the subject area for the

achievement test in the current study. It is also possible to study the same research questions in other subject areas (e.g., comparing if the results differ across the subject areas requiring verbal skills or numerical skills). In addition, Angoff and Nedelsky methods were chosen for the criterion-referenced assessment in the current study, but some other standard-setting methods (e.g., borderline group method, contrasting groups method, etc.) can be considered in a future study. Furthermore, the number of performance levels was two in the current study. More than two performance levels can be studied in a future work (e.g., the number of letter grades in universities).

Acknowledgements

This study is composed of some parts of the master's thesis entitled as "*A Study upon Comparison of Norm- and Criterion-referenced Assessments*".

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 author(s).

Authorship contribution statement

Ismail Cuhadar: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing - original draft. **Selahattin Gelbal:** Methodology, Supervision and Validation.

ORCID

Ismail CUHADAR https://orcid.org/0000-0002-5262-5892
Selahattin GELBAL https://orcid.org/0000-0001-5181-7262

5. REFERENCES

Aiken, L. R. (2000). *Psychological testing and assessment*. Boston: Allyn and Bacon.

Akdeniz University. (2017, September 17). *Akdeniz Üniversitesi Ön Lisans ve Lisans Eğitim-Öğretim ve Sınav Yönetmeliği [Akdeniz University Regulations for Associate and Undergraduate Degree Education and Examinations]*. Retrieved May 27, 2020, from http://oidb.akdeniz.edu.tr/wp-content/uploads/2017/02/Akdeniz-Üniversitesi-Ön-Lisans-ve-Lisans-Eğitim-Öğretim-ve-Sınav-Yönetmeliği-17.09.2017.pdf

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement*. Washington, DC: American Council on Education.

Ankara University. (2018, September 4). *Ankara Üniversitesi Ön Lisans ve Lisans Eğitim-Öğretim Yönetmeliği [Ankara University Regulations for Associate and Undergraduate Degree Education and Examinations]*. Retrieved May 27, 2020, from http://oidb.ankara.edu.tr/files/2018/04/ÖN-LİSANS-VE-LİSANS-EĞİTİM-ÖĞRETİM-YÖNETMELİĞİ.pdf

Arrasmith, D. G. (1986). *Investigation of judges' errors in Angoff and contrasting groups cut of score methods* [Doctoral dissertation, University of Massachusetts]. ProQuest Dissertations and Theses.

Baykul, Y. (2010). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması [Measurement in Education and Psychology: Classical test theory and applications]*. Ankara: Pegem Yayıncılık.

Berk, R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education, 45*(2), 4-9.

Calmorin, L. P., & Calmorin, M. A. (2007). *Research methods and thesis writing*. Manila: Rex Book Store.

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement, 30*(2), 93-106.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousands Oaks, CA: Sage Publications.

Clark-Carter, D. (2005). *Quantitative psychological research: a student handbook.* New York, NY: Psychology Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory.* Ohio: Cengage Learning.

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

Erciyes University. (2015, December 27). *Erciyes Üniversitesi Ön Lisans ve Lisans Eğitim-Öğretim Yönetmeliği [Erciyes University Regulations for Associate and Undergraduate Degree Education and Examinations]*. Retrieved May 27, 2020, from https://www.erciyes.edu.tr/kategori/ERU-E-BELGE/Yonetmelikler/131/136

Ertürk, S. (1998). *Eğitimde program geliştirme [Program development in education]*. Ankara: Meteksan.

Freeman, L., & Miller, A. (2001). Norm-referenced, criterion-referenced, and dynamic assessment: What exactly is the point? *Educational Psychology in Practice, 17*(1), 3-16.

Çetin, S., & Gelbal, S. (2010). Impact of standard setting methodologies over passing scores. *Ankara University, Journal of Faculty of Educational Sciences, 43*(1), 79–95.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*(4), 353-366.

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 and National Council on Measurement in Education.

Jacobson, R. Y. (2008). *Examination of the potential of selected norm-referenced tests and selected locally developed criterion-referenced tests to classify students into performance categories* [Doctoral dissertation, University of Nebraska]. ProQuest Dissertations and Theses.

Jekel, J. F. (2007). *Epidemiology, biostatistics and preventive medicine.* Philadelphia: Saunders/Elsevier.

Johnson, D. L., & Martin, S. (1980). Criterion-referenced testing: New wine in old bottles. *Academic Therapy, 16*(2), 167 - 173.

Kendall, M. G., & Smith, B. B. (1939). The problem of m rankings. *The Annals of Mathematical Statistics, 10*(3), 275-287.

Kline, P. (2000). *Handbook of psychological testing.* London and New York: Routledge.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

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

McCauley, R. J., & Swisher, L. (1984). Use and misuse of norm-referenced tests in clinical assessment: A hypothetical case. *Journal of Speech and Hearing Disorders, 49*(4), 338-348.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276-282.

Microsoft Corporation. (2013). *Microsoft Excel*. Retrieved from https://office.microsoft.com/excel

Milli Eğitim Bakanlığı. (2014, July 26). *Milli Eğitim Bakanlığı okul öncesi eğitim ve ilköğretim kurumları yönetmeliği [Ministry of National Education regulations for preschool and primary school education institutions]*. Retrieved May 27, 2020, from http://mevzuat.meb.gov.tr/dosyalar/1703.pdf

Milli Eğitim Bakanlığı. (2016, October 28). *Milli Eğitim Bakanlığı ortaöğretim kurumları yönetmeliği [Ministry of National Education regulations for secondary school education institutions]*. Retrieved May 27, 2020, from https://ogm.meb.gov.tr/meb_iys_dosyalar/2016_11/03111224_ooky.pdf

Mohr, A. K. (2006). *Criterion referenced and norm referenced predictors of student achievement: Teacher perceptions of, and correlations between, Iowa test of basic skills and the palmetto achievement challenge test* [Doctoral dissertation, University of South Carolina]. ProQuest Dissertations and Theses.

Montgomery, P. C., & Connolly, B. H. (1987). Norm-referenced and criterion referenced tests: Use in pediatrics and application to task analysis of motor skill. *Physical Therapy, 67*(12), 1873-1876.

Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological testing: Principles and applications.* New Jersey: Pearson.

Nartgün, Z. (2007). Aynı puanlar üzerinden yapılan mutlak ve bağıl değerlendirme uygulamalarının notlarda farklılık oluşturup oluşturmadığına ilişkin bir inceleme [An investigation on whether the applications of the norm- and criterion-referenced assessments over the same scores make a difference in grading]. *Ege Eğitim Dergisi, 8*(1), 19-40.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14*(1), 3–19.

Oescher, J., Kirby, P. C., & Paradise, L. V. (1992). Validating state-mandating criterion-referenced achievement tests with norm-referenced test results for elementary and secondary students. *Journal of Experimental Education, 60*(2), 141-150.

Paura, L., & Arhipova, I. (2014). Cause analysis of students' dropout rate in higher education study program. *Procedia-Social and Behavioral Sciences, 109*, 1282-1286.

Pester, A. M. (2003). *Language intervention effects of norm-referenced and criterion referenced test scores* [Master's thesis, Miami University]. https://etd.ohiolink.edu/!etd.send_file?accession=miami1050351250&disposition=inline

Sakarya University. (2019, April 18). *Sakarya Üniversitesi Ön Lisans ve Lisans Eğitim-Öğretim ve Sınav Yönetmeliği [Sakarya University Regulations for Associate and Undergraduate Degree Education and Examinations]*. Retrieved May 27, 2020, from https://www.sakarya.edu.tr/yeni-lisans-ve-onlisans-egitim-ogretim-ve-sinav-yonetmeligi-d330.html

SPSS, Inc. (2009). PASW Statistics for Windows (Version 18.0) [Computer Program]. Chicago: SPSS Inc.

Sönmez, V., & Alacapınar, F. G. (2011). *Örneklendirilmiş bilimsel araştırma yöntemleri [Scientific research methods with examples]*. Ankara: Anı Yayıncılık.

Toprakçı, E., Baydemir, G., Koçak, A., & Akkuş, Ö. (2007, September). *Eğitim fakültelerinin eğitim-öğretim ve sınav yönetmeliklerinin karşılaştırılması [A comparison of regulations for education and examinations in faculty of education]*. Paper presented at the meeting of 16. Ulusal Eğitim Bilimleri Kongresi, Tokat, Turkey.

Turgut, M. F., & Baykul, Y. (2010). *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education]*. Ankara: Pegem Yayıncılık.

Tyler, R. W. (2013). *Basic principles of curriculum and instruction.* Chicago: The University of Chicago Press.

Urbina, S. (2004). *Essentials of psychological testing.* New York: Wiley

Visintainer, C. (2002). *The relationship between two state-mandated, standardized tests using norm-referenced Terranova and the criteria-referenced, performance assessment developed for the Maryland school performance assessment program* [Doctoral dissertation, Wilmington College]. ProQuest Dissertations and Theses.

Vossensteyn, J. J., Kottmann, A., Jongbloed, B. W., Kaiser, F., Cremonini, L., Stensaker, B., Hovdhaugen, E., & Wollscheid, S. (2015). *Dropout and completion in higher education in Europe: Main report*.

Yıldırım, C. (2011). *Bilim felsefesi [Philosophy of science]*. İstanbul: Remzi Kitabevi.

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