ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(1);12-29

A Statistical Comparison of Norm-Referenced Assessment Systems Used in Higher Education in Turkey*

Erkan Hasan ATALMIŞ **

Abstract

The purpose of this study is to identify different norm-referenced assessment systems used in Turkish higher education, and to compare them empirically. Norm-referenced assessment regulations of 70 universities in Turkey was primarily analyzed, and universities were divided into four different groups depending on their norm-referenced assessment systems (only applying T-score conversion, the most commonly used method; applying T-score conversion and quantiles together; applying T-score conversion, quantiles and standard deviation together; applying standard deviation based norm-referenced assessment system). After the algorithms of two universities applying T-score conversion and three universities applying other norm-referenced assessment system were selected, they were used to convert end-of-year grade for each course of 19,574 students in a state university into letter grades and 4-point system. To test the differences of the norm-referenced assessment systems used in these universities, the norm-referenced system of a university were compared with the criterion-referenced system of the same university as well as norm-referenced systems of other universities. The paired t-test was used to identify the difference between norm-referenced and criterion-referenced assessment, while the differences between norm-referenced assessment systems were analyzed through one-way analysis of variance. The findings revealed that the letter grades calculated through the norm-referenced assessment were statistically different than the ones calculated with criterion-referenced; besides, a statistically significant difference was identified between the letter grades obtained using the norm-referenced assessment systems of universities. At the end of the study, the findings were discussed in term of students and instructors.

Key Words: Norm-referenced assessment, criterion-referenced assessment, assessment in higher education, grading system.

INTRODUCTION

The aim of education and training is the disclosure of the cognitive, affective and psychomotor skills to the students in a planned and programmed way. To ensure this, the curriculum consisting of four basic components should be primarily determined. These components include (a) determining the behavioral objectives, (b) constructing the content in accordance with these objectives and the readiness of the students, (c) creating learning and teaching activities with the idea that each student learns differently; and (d) performing meaningful assessment and evaluation (Tan, 2015). In particular, the significance of measurement and assessment cannot be underestimated in terms of determining the extent to which the behavioral objectives within the program reflect the readiness of the students and identifying as to what extent learning and teaching activities are appropriate to the objectives and behaviors.

The concepts of measurement and assessment are different and even complementary concepts. Specifically, measurement refers to a variable or an object with numbers or symbols, while assessment provides a meaningful interpretation of the results obtained from the measurement by comparing them through a frame of reference. Previous studies have revealed that a frame of reference will vary across teacher notions, student success distribution in the class, student ability and their achievement scores related to the program (learning difference at the beginning and end of the program) as well as the

* A part of the study was presented at 2018 EDUCCON Education Conference (Ankara University, Ankara, Turkey.
** Assistant professor, Kahramanmaraş Sütçü Imam University, Faculty of Education, Kahramanmaraş-Türkiye, eatalmis@ksu.edu.tr, ORCID ID: orcid.org/0000-001-9610-491X

To cite this article:
Atalmış, E., H. (2019). A statistical comparison of norm-referenced assessment systems used in higher education in Turkey. *Journal of Measurement and Evaluation in Education and Psychology, 10*(1), 12-29. DOI: 10.21031/epod.487335

Received: 25.11.2018
Accepted: 28.01.2019

objectives of the program (Martin & Jolly, 2002; Turgut & Baykul, 2015; Yorke, 2011). Considering these situations, one of the most essential factors of an accurate assessment is the selection of an appropriate reference. The assessment is already categorized depending on the reference in use. Criterion-referenced in which absolute criterion is used is defined as an assessment which is accepted by everyone in the same way without reference to the group and group characteristics. Norm-referenced is a type of assessment which yields for the relative criteria and the assessment made depending on the criteria selected according to the predefined group and especially the success of the group.

Thorndike (2005) argues that the criterion-referenced assessment plays a significant role in directing learning and teaching activities since the use of this type of assessment is more relevant to what extent people achieve the level of targeted knowledge. Sadler (2005) suggests that the criterion-referenced assessment provides students with the grades they deserve due to the fact that the grades based on this assessment are calculated regardless of each student's achievement. In contrast to the criterion-referenced assessment, it is recommended that norm-referenced assessment be used for sorting, placement and in distinguishing the achievement sequence of the students (American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME), 2014). Nartgün (2007) has noted that the use of norm-referenced assessment is particularly relevant in the large-scale and national-scale examinations that have upper levels and that require placement.

Although there does not exist a definite line in terms of the use of criterion-referenced and norm-referenced assessment, the exams that the criterion-referenced method is applied may be prepared in accordance with the exam preparation guidelines (Thorndike, 2005). Haladyna and Rodriguez (2013) have indicated that susceptibility should be displayed for the construction of exam questions based on the objectives, item-writing guidelines and the cognitive taxonomy levels. Otherwise, the scores obtained as a result of the examination will tend to be less homogeneous; the variation of the measured feature will not be explained at the maximum level, and in this case, ranking or placement of the students according to the scores will not be reasonable in terms of measurement and assessment (Hambleton et al., 1978).

In particular, the question related to what kind of assessment type is applied is the subject of hot debate in higher education today. The current relevant studies have put forward that the use of norm-referenced assessment in universities with high competition and success motivation will be much more effective, and that criterion-referenced assessment will be more appropriate in schools with low achievement motivation since norm-referenced assessment will cause grading inflation (Başol, 2013; Selvi, 1998). In one of the most comprehensive studies on this subject, Johnson (2003) has argued that grading inflation is a serious problem in universities and that the method used for grading can vary across universities and faculties. Unfortunately, a limited number of studies have been conducted to compare criterion-referenced and norm-referenced assessment.

Having analyzed the guidelines of two different state universities using criterion-referenced and norm-referenced assessment, Nartgün (2007) has compared the grades with different distributions and suggested that the grades calculated with the norm-referenced assessment provide more than those of the students deserve. In other words, when compared with the criterion-referenced assessment, it is evident that norm-referenced assessment leads to grading inflation. However, criterion-referenced assessment method offers more effective results than the norm-referenced assessment in cases when the raw achievement grades are close to one another, that is, the scores are similar (standard deviation is low). Atılgan, Yurdakul, and Öğretmen (2012) have found different results compared to the findings of Nartgün (2007). 3,120 grades obtained from the students studying at the faculty of education and calculated through use of norm-referenced assessment have been converted into criterion-referenced assessment, and it has been determined that 42% of the grades are free from any change. Twenty-two percent of these grades increase in favor of the norm-referenced assessment, while the rest increases concerning criterion-referenced assessment. In addition, the results of the interviews conducted with the faculty members in the same study have shown that the grading system calculated through the

norm-referenced assessment has a negative impact on the interaction among students. This situation results in negative competition among the students and grouping; moreover, it also decreases the sense of trust towards each other and damages the value education of the students.

Besides, Duman (2011) has emphasized the positive side of the norm-referenced assessment method. In a study conducted with primary school prospective teachers, the norm-referenced assessment has been identified to partially compensate the grading deficiencies emerging due to the faculty members and the exam questions. Similar results have emerged in previous studies. In particular, both classroom assessment questions prepared by teachers and questions in question banks are of low quality (Demir & Atalmış, 2017; Downing, 2005; Masters et al., 2001; Mehrens & Lehmann, 1991; Tarrant et al., 2006). As a result of the analysis on measurement and assessment, both the reliability and validity of the test scores with low-quality questions will decrease and a false assessment mechanism (decision) will be formed depending on the test scores that are calculated incorrectly (Çelen & Aybek, 2013).

Besides, Nartgün (2007) has argued that the norm-referenced assessment would provide an advantage to students after graduation. The use of norm-referenced assessment system in higher education in the countries, especially in Turkey, where competition is common in terms of both university entrance exam and after graduation provides an opportunity to compare the grades of the students graduating from different universities on the same scale. To illustrate, the student who graduated with a score of 60 from a university that requires high score is considered to have near or equal achievement score with the student who has graduated with the score of 80 from a university requiring low achievement score.

In a more explicit way, just as the person who graduated with a lower GPA from University A that admits those with high scores cannot be perceived in the same way as the person who graduated with a higher GPA from University B that accepts those with low scores, the person graduating from University B can be perceived as more successful at first sight. However, the person graduating from University A may have received far more comprehensive, well-equipped and innovative education. Therefore, s/he may have experienced a difficult process and graduated with a low GPA. For this reason, a norm-referenced assessment system may enable to compare two students graduating from these two different universities through using the same scale.

The Assessment Systems Used at Universities in Turkey

Upon reviewing the state universities in Turkey, the norm-referenced assessment system is used as a grading system in the majority of the universities. Considering the universities that use criterion-referenced assessment, they have been identified to hold different pass/fail cutoff score and their letter grades are determined variously.

The letter grade of CC equals to 60 in most of these universities (e.g. Selçuk University, Şırnak University, Uludağ University), one of which is 54 (Amasya University), one is 64 (Abdullah Gül University), while 50 in others (e.g. Recep Tayyip Erdoğan University, Gümüşhane University, and Bayburt University); moreover, the grade of CC corresponds to 65 in some of the universities (e.g. Hacettepe University and Bolu Abant İzzet Baysal University) and 70 in only a small number of them (METU, Boğaziçi University, and Gaziantep University). In particular, the universities that keep CC score range high were observed to have high university entrance grades (e.g. Hacettepe University, Gebze Technical University, Boğaziçi University, and METU) or they have been separated from such universities in the following years. For instance, Gaziantep University may be considered a university separated from METU but still able to protect METU traditions.

Given the instructions of the universities using norm-reference assessment system in Turkey, they use different methods and algorithms depending on the number of students in the class, classroom grade point average, percentiles and standard deviations of grades. While some universities determine the grade of CC, they allow the instructors to intervene in addition to the classroom grade point average (Ankara University and Istanbul Technical University). Besides, criterion-referenced assessment or norm-referenced assessment systems is implemented in many universities according to the grade

interval by calculating T score (Akdeniz University, Aksaray University, Bartın University, Bitlis Eren University, Bursa Technical University, Bülent Ecevit University, Ege University, Fırat University, Hitit University, Karamanoğlu Mehmet Bey University, Kırıkkale University, Kilis 7 Aralık University, Mehmet Akif Ersoy University, Muğla Sıtkı Koçman University, Muş Alparslan University, Tunceli University, Uşak University).

Likewise, the letter grades of the students are given with the percentages of the students in the class in some of the universities that apply T score calculation method (Artvin Çoruh University, Atatürk University, Balıkesir University, Celal Bayar University, Cumhuriyet University, Çankırı Karatekin University, İzmir Katip Çelebi University, Kafkas University, Karadeniz Technical University, Marmara University, Namık Kemal University, Niğde University, Ondokuz Mayıs University, Süleyman Demirel University, Trakya University), while some of the universities determine the letter grades by taking the standard deviation of the class in addition to those mentioned above (Selçuk University, Yalova University).

In some universities that do not use the T score method, a criterion-referenced or norm-referenced assessment system is applied depending on the standard deviation of the grade distribution in the class as well as the number of students and classroom grade point average (İstanbul University, Çukurova University, Harran University, İnönü University, Kahramanmaraş Sütçü İmam University). Similarly, the upper and lower limits of the letter range in these universities are determined by the university administration. These different norm-referenced systems are presented in detail as follows:

Norm-Referenced Assessment System Only Applied According to T Score Conversion

This assessment system initially calculates students' raw success grades (RSG) through using midterm and final exam grades. RSG is calculated by taking 40% of midterm and 60% of the final exam. While determining students who will be included in the norm-referenced assessment system, RSG lower limit and the lower limit of number of students are used. Table 1 displays RSG lower limit and the number of students in universities applying norm-referenced assessment.

Table 1. The Lower Limit of RSG (TLLORSG) and the Number of Students (TLLOTNOS) in Universities Applying Norm-Referenced Assessment

University Name	TLLORSG	TLLOTNOS	University Name	TLLORSG	TLLOTNOS
Adana Sci. and Tech. Uni.	15	11	Iğdır Uni.	20	11
Adıyaman Uni.	20	20	İnönü Uni.	40	11
Ağrı İbrahim Çeçen Uni.	30	10	İskenderun Tech. Uni.	15	30
Akdeniz Uni.	20	15	İstanbul Tech. Uni.	-	-
Aksaray Uni.	35	11	İstanbul Uni.	35	20
Alanya Alaaddin Key. Uni.	20	16	İzmir Katip Çel. Uni.	30	11
Anadolu Uni.	25	30	Kafkas Uni.	40	11
Ankara Uni.	-	30	Kahraman. S. I. Uni.	25	15
Ardahan Uni.	20	10	Karadeniz Tech. Uni.	15	11
Artvin Çoruh Uni.	15	11	Karamanoğlu M. Uni.	20	10
Atatürk Uni.	-	10	Kırıkkale Uni.	15	30
Balıkesir Uni.	15	10	Kırklareli Uni.	20	-
Bandırma Onyedi Ey. Uni.	15	11	Kilis 7 Aralık Uni.	20	11
Bartın Uni.	15	10	Marmara Uni.	20	10
Bilecik Şeyh Edebali Uni.	45	10	Mehmet A. E. Uni.	15	20
Bitlis Eren University	20	20	Mimar Sinan F. A. Uni.	-	30
Bozok Uni.	-	-	Muğla S. K. Uni.	10	30
Bursa Technical Uni.	20	20	Muş Alparslan Uni.	15	10
Bülent Ecevit Uni.	35	25	Namık Kemal Uni.	15	11
Celal Bayar Uni.	20	20	Nevşehir H. B. V. Uni.	-	-
Cumhuriyet Uni.	15	11	Niğde Ö. H. Uni.	10	11

Çankırı Karatekin Uni.	25	10	Ondokuz Mayıs Uni.	20	11
Çukurova Uni.	35	20	Osmaniye K. A. Uni.	20	1
Dokuz Eylül Uni.	-	-	Sakarya Uni.	-	-
Dumlupınar Uni.	15	-	Selçuk Uni.	15	20
Ege Uni.	15	30	Süleyman Demirel Uni.	15	11
Erciyes Uni.	-	-	Trakya Uni.	15	11
Erzincan B. Y. Uni.	-	10	Tunceli Uni.	15	10
Eskişehir Osmangazi Uni.	-	-	Uludağ Uni.	20	-
Fırat University	10	15	Uşak Uni.	20	10
Gazi University	-	-	Yalova Uni.	20	20
Harran University	40	20	Yıldız Tech. Uni.	-	-
Hitit University	30	20	Yüzüncü Yıl Uni.	15	30

As can be seen in Table 1, the lower limit of RSG and the number of students are determined as 15 and 11 in 65 state universities using norm-referenced assessment system in Turkey. Specifically, 11 students with RSG scores greater than 15 are required to use a norm-referenced assessment system. Otherwise, criterion-referenced assessment system is supposed to be used rather than norm-referenced assessment.

Following this stage, students' scores will be converted into T scores by using the following formulas in the norm-referenced assessment system, which requires only "T-score conversion" (Güler, 2017).

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} \quad , \quad \sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}} \quad , \quad T = [(\frac{X_i - \mu}{\sigma}) \times 10] + 50$$

Here, N refers to the number of students participating in the assessment, X_i signifies students' RSG, μ, represents the students' RSG average, σ is the standard deviation of the students' RSG and T is the score converted from students' RSGs. After each student's T score is obtained, the letter grades are given to the students by using the value in Table 2 depending on the RSG average of the class.

Table 2. Calculation of Letter Grades in terms of T Score

Class Level	RSG average of the class	AA (4)	BA (3.5)	BB (3)	CB (2.5)	CC (2)	DC (1.5)	DD (1)	FF (0)
Outstanding	$80<\mu\leq100$	≥ 57	52-56.99	47–51.99	42–46.99	37–41.99	32–36.99	27–31.99	< 27
Excellent	$70<\mu\leq80$	≥ 59	54-58.99	49–53.99	44–48.99	39–43.99	34–38.99	29–33.99	< 29
Very Good	$62.5<\mu\leq70$	≥ 61	56–60.99	51–55.99	46–50.99	41–45.99	36–40.99	31–35.99	< 31
Good	$57.5<\mu\leq62.5$	≥ 63	58–62.99	53–57.99	48–52.99	43–47.99	38–42.99	33–37.99	< 33
Satisfactory	$52.5<\mu\leq57.5$	≥ 65	60–64.99	55–59.99	50–54.99	45–49.99	40–44.99	35–39.99	< 35
Sufficient	$47.5<\mu\leq52.5$	≥ 67	62–66.99	57–61.99	52–56.99	47–51.99	42–46.99	37–41.99	< 37
Poor	$42.5<\mu\leq47.5$	≥ 69	64–68.99	59–63.99	54–58.99	49–53.99	44–48.99	39–43.99	< 39
Fail	$\mu<42.5$	≥ 71	66–70.99	61–65.99	56–60.99	51–55.99	46–50.99	41–45.99	< 41

Table 2 suggests that a class in which the norm-referenced assessment system is applied is in one of eight different levels according to RSG average of the class. To illustrate, the class whose RSG average is between 80 and 100 is considered "outstanding", while the class whose RSG average varies between 57.5 and 62.5 is regarded as "Good". Taking the students' letter grades into account, the letter grade of a student whose T score is 58 and who is in the class with 55 RSG average (Good) is BB; whereas the student with the same T score but in the class with 65 RSG average (Very Good) has BA letter grade.

When the norm-referenced assessment regulations of the universities are examined, most of the universities use the chart as in Table 2 while calculating T score (Celal Bayar University, Kafkas University, Marmara University), while others use criterion-referenced assessment system for upper-level classes. In no uncertain terms, some of the universities with 60 and over (Akdeniz University), 70 and over (Bartın University, İzmir Katip Çelebi University, Muğla Sıtkı Koçman University), 80 and over RSG average (Karadeniz Technical University, Ondokuz Mayıs University, Uşak University) use criterion-referenced assessment system. Table 3 depicts the letter ranges in the criterion-referenced assessment applied in some universities. Besides, some of the universities use norm-referenced assessment system by decreasing the number of class levels (below 8) and enlarging the class level intervals in Table 2 (Bülent Ecevit University, Çankırı Karatekin University, Niğde University, Süleyman Demirel University).

Table 3. Letter Intervals of the Universities Regarding Criterion-Referenced Assessment System

4-point Grading System	Letter Grade	Aksaray Uni. RSG Intervals	Akdeniz Uni. RSG Intervals	Karadeniz Technical Uni. RSG Intervals	Selçuk Uni. RSG Intervals	İstanbul Uni. RSG Intervals
4.00	AA	90 – 100	87.5 – 100	90 – 100	88 – 100	88 – 100
3.50	BA	85 – 89	80.5 – 87.4	80 – 89	80 – 87	80 – 87
3.00	BB	80 – 84	73.5 – 80.4	75 – 79	73 – 79	73 – 79
2.50	CB	70 – 79	66.5 – 73.4	70 – 74	66 – 72	66 – 72
2.00	CC	60 – 69	59.5 – 66.4	60 – 69	60 – 65	60 – 65
1.50	DC	55 – 59	52.5 – 59.4	50 – 59	55 – 59	55 – 59
1.00	DD	50 – 54	45.5 – 52.4	40 – 49	50 – 54	50 – 54
0.50	FD	40 – 49	34.5 – 45.4	30 – 39	–	–
0.00	FF	0 – 45	0 – 34.4	0 – 29	0 – 39	0 – 49

Norm-Referenced Assessment System Applying T Score Conversion and Quantiles

This assessment system is based on the number of students participating in the norm-referenced assessment. Upon examining the norm-referenced assessment regulations of the universities applying this assessment system, only T score conversion is conducted in cases when the number of students participating in the norm-referenced assessment is 30 or over just as in Table 2, while the letter grades are based on quantiles as in Table 4 when the number of the students is between 10 (11 in some of the universities) and 29 (30 in some of the universities).

Table 4. The Calculation of Letter Grades Depending on Quantiles

Class Level	RSG average of the class	AA (4)	BA (3.5)	BB (3)	CB (2.5)	CC (2)	DC (1.5)	DD (1)	FF (0)
Outstanding	$70<\mu\leq100$	24(24)	15.2(39.2)	22.8(62)	11.6(73.6)	17.4(91)	4.8(95.8)	3.2(99)	1(100)
Excellent	$62.5<\mu\leq70$	18(18)	14.4(32.4)	21.6 (54)	12.8 (66.8)	19.2 (86)	7.2(93.2)	4.8(98)	2(100)
Very Good	$57.5<\mu\leq62.5$	14(14)	12.8 (26.8)	19.2 (46)	14.4 (60.4)	21.6 (82)	9(91)	6(97)	3(100)
Good	$52.5<\mu\leq57.5$	10(10)	11.6(21.6)	17.4(39)	14.8(53.8)	22.2 (76)	12(88)	8(96)	4(100)
Satisfactory	$47.5<\mu\leq52.5$	7(7)	9.6 (16.6)	14.4 (31)	15.2 (46.2)	22.8 (69)	14.4 (83.4)	9.6(93)	7(100)
Sufficient	$42.5<\mu\leq47.5$	4(4)	8(12)	12(24)	14.8 (38.8)	22.2(61)	17.4 (78.4)	11.6(90)	10(100)
Poor	$\mu<42.5$	3(3)	6 (9)	9(18)	14.4 (32.4)	21.6 (54)	19.2 (73.2)	12.8(86)	14(100)

* The values in parentheses indicate the percentage of the cumulative percentages.

First, the percentage of the students participating in the norm-referenced assessment is calculated while determining their letter grades and then their letter grades are identified through using Table 4. For instance, in a class where the RSG average is 60, the letter grade of a student in the top 10% is AA, while that of a student in the top 30% is identified as BB.

Norm-Referenced Assessment System Implementing T-Score Conversion, Quantiles, and Standard Deviation

Within this system, the RSGs' standard deviation of some students participating in the norm-referenced assessment in some universities is calculated in addition to the T-score conversion and the percentile method. If the standard deviation is below a certain value, a criterion-referenced assessment system is used. This value ranges between 4 (such as Yalova University) and 8 (Selçuk University) based on the regulations of the universities.

Standard Deviation Based Norm-Referenced Assessment System

This assessment system holds criterion or norm-referenced assessment systems by focusing on the standard deviation of the grade distribution in the classroom as well as the average of the student group (class) participating in the norm-referenced assessment. As indicated in the study conducted by Nartgün (2007), the effectiveness of the criterion or norm-referenced assessment systems depends on the standard deviation. In this system, moreover, criterion or norm-referenced assessment is applied depending upon the lower limit of the number of students participating in the norm-referenced assessment. This varies between 15 and 20 according to the regulations of the universities. Nevertheless, in this assessment system, the criterion-referenced assessment system is a prerequisite when the standard deviation of the grades in the class is below 8 (grade distributions are close to each other). In applying this system, RSG is calculated as the sum of 40% of the student's mid-term scores and 60% of their final scores.

Table 5 presents the letter grade determination table of the İstanbul University which first applied this norm-referenced assessment system.

Table 5. İstanbul University Letter Grade Calculation through Norm-Referenced Assessment System

Letter Grade	Very Poor: μ<44	Poor: 44≤μ<50	Below: 50≤μ<56	Average: 56≤μ<63
AA	[μ+1.881σ, 100]	[μ+1.645σ, 100]	[μ+1.476σ, 100]	[μ+1.227σ, 100]
BA	[μ+1.405σ, μ+1.881σ)	[μ+1.175σ, μ+1.645σ)	[μ+0.994σ, μ+1.476σ)	[μ+0.739σ, μ+1.227σ)
BB	[μ+0.706σ, μ+1.405σ)	[μ+0.524σ, μ+1.175σ)	[μ+0.358σ, μ+0.995σ)	[μ+0.126σ, μ+0.739σ)
CB	[μ+0.332σ, μ+0.706σ)	[μ+0.126σ, μ+0.524σ)	[μ-0.075σ, μ+0.358σ)	[μ-0.358σ, μ+0.126σ)
CC	[μ-0.176σ, μ+0.332σ)	[μ-0.468σ, μ+0.126σ)	[μ-0.772σ, μ-0.075σ)	[μ-0.878σ, μ-0.358σ)
DC	[μ-0.643σ, μ-0.176σ)	[μ-0.878σ, μ-0.468σ)	[μ-1.126σ, μ-0.772σ)	[μ-1.227σ, μ-0.878σ)
DD	[μ-1.175σ, μ-0.643σ)	[μ-1.405σ, μ-0.878σ)	[μ-1.645σ, μ-1.126σ)	[μ-1.751σ, μ-1.227σ)
FF	[35, μ-1.175σ)	[35, μ-1.405σ)	[35,μ-1.645σ)	[35,μ-1.751σ)
Letter Grade	Above Average: 63≤μ<71	Good: 71≤μ<80	Very Good: μ≥80	*For the absence of conflicts, the intervals are shown as"[" indicating "included" and closed from the left side, and ")" referring to "excluded" and open from the right side. In the table, μ suggests the average of the RSG values and σ shows the standard deviation of these values.
AA	[μ+0.915σ, 100]	[μ+0.583σ, 100]	[μ+0.440σ, 100]	
BA	[μ+0.385σ, μ+0.915σ)	[μ+0.100σ, μ+0.583σ)	[μ-0.100σ, μ+0.440σ)	
BB	[μ-0.075σ, μ+0.385σ)	[μ-0.305σ, μ+0.100σ)	[μ-0.496σ, μ-0.100σ)	
CB	[μ-0.524σ, μ-0.075σ)	[μ-0.739σ, μ-0.305σ)	[μ-0.915σ, μ-0.496σ)	
CC	[μ-0.994σ, μ-0.524σ)	[μ-1.126σ, μ-0.739σ)	[μ-1.282σ, μ-0.915σ)	
DC	[μ-1.341σ, μ-0.994σ)	[μ-1.476σ, μ-1.126σ)	[μ-1.645σ, μ-1.282σ)	
DD	[μ-1.881σ, μ-1.341σ)	[μ-2.054σ, μ-1.476σ)	[μ-2.326σ, μ-1.645σ)	
FF	[35,μ-1.881σ)	[35,μ-2.054σ)	[35, μ-2.326σ)	

Here μ represents classroom average and σ signifies the standard deviation of the distribution in the class. Formulas of μ and σ are presented as follows (Field, 2009).

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}, \qquad \sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}}$$

In this formula, X_i represents the RSG of a student participating in the norm-referenced assessment, while *N* shows the total number of the students participating in this assessment.

Table 3 indicates the criterion-referenced assessment letter intervals in İstanbul University when the norm-referenced assessment system is not applied. Criterion-referenced assessment is applied in cases where the number of students in the class is less than 10, and/or the standard deviation of the grade distribution in the class is below 8.

The Significance and Aim of Research

Unlike previous studies focusing a particular norm-referenced system in Turkey (Atılgan et al., 2012; Duman, 2011; Nartgün, 2007), this research has compared aforementioned four different norm-referenced assessment systems that are widely used in Turkey. In this regard, Aksaray University, which uses norm-referenced assessment for each grade level among the universities that apply only T-score conversion, and the norm-referenced assessment system of Akdeniz University (for the classes with over 60 RSG mean) that applies criterion-referenced assessment systems for the upper-level classes have been chosen. Among the universities that use T-score conversion and quantile method, the sample of the norm-referenced assessment system of Karadeniz Technical University has been chosen; Selçuk University's norm-referenced assessment system has been preferred as an example for the university using T score conversion, quantile, and standard deviation. Finally, İstanbul University norm-referenced assessment system has been chosen as it does not use T-score conversion and only uses a standard deviation-based conversion.

In order to achieve this goal, answers to the following questions have been sought:

1. Is there a statistically significant difference between students' letter grades calculated through norm-referenced assessment and criterion-referenced assessment?
2. Is there a statistically significant difference between students' letter grades calculated by different norm-referenced assessment systems?

METHOD

This study can be considered a causal-comparative research approach, seeking to determine differences between groups by examining differences in the experiences of group members (Lodico, Spaulding, & Voegtle, 2010). In this study, the groups of individuals are students whose letter grades calculated through different norm-referenced assessment and criterion-referenced assessment. This section holds information regarding the research sample, process, and data analysis.

Research Sample

The research data have been collected through the midterm and final grades of the students studying at different faculties and vocational colleges in a state university during the fall semester of 2014-2015, and the total of 19,574 students' RSGs have been considered during data analysis.

Process

Students' RSGs have been calculated by taking into account 40% of the midterm score and 60% of the final score. Afterward, the RSGs of the students have been calculated as letter grades by adapting them depending on the above-mentioned assessment systems of Aksaray University, Akdeniz University, Karadeniz Technical University, Selçuk University, and İstanbul University. Besides, the RSGs of the students have been converted into letter grades considering the absolute assessment table of these universities as displayed in Table 2. Thus, each student's letter grade calculated by means of both norm-referenced and criterion-referenced assessment has been determined and then converted to the

grade equivalent to 4- point grading system. Given the students took more than one course in the 2014-2015 fall semester, the same process has been performed for each course's RSG taken by each student. In other words, the analysis in this research has been conducted for 19,574 students' 157,983 letter grade.

Data Analysis

In order to identify the difference between norm-referenced and criterion-referenced assessment, which is the first research question, the paired t-test was used to analyze the difference between the two groups. For the second research question, the differences between the letter grades obtained by the norm-referenced assessment system of each of the universities mentioned above have been analyzed through one-way analysis of variance (ANOVA).

RESULTS

Figure 1 depicts the distribution of students' RSGs, are not converted scores through norm-referenced/criterion-referenced assessment.

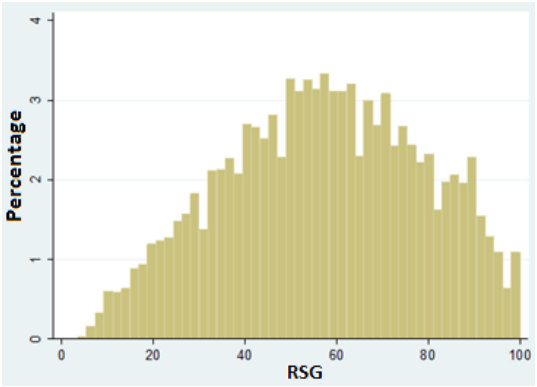

Figure 1. Students' End-of-Term Raw Success Grade (RSG) Distribution

Upon examining Figure 1, the skewness coefficient of the students' RSGs distribution was identified to be -0.109 and this value was considered normal because it is between -1 and +1 (Büyüköztürk, 2009).

In the next stage, norm-referenced and criterion-referenced assessment systems used by Aksaray University, Akdeniz University, Karadeniz Technical University, Selçuk University, and İstanbul University have been applied for these RSGs.

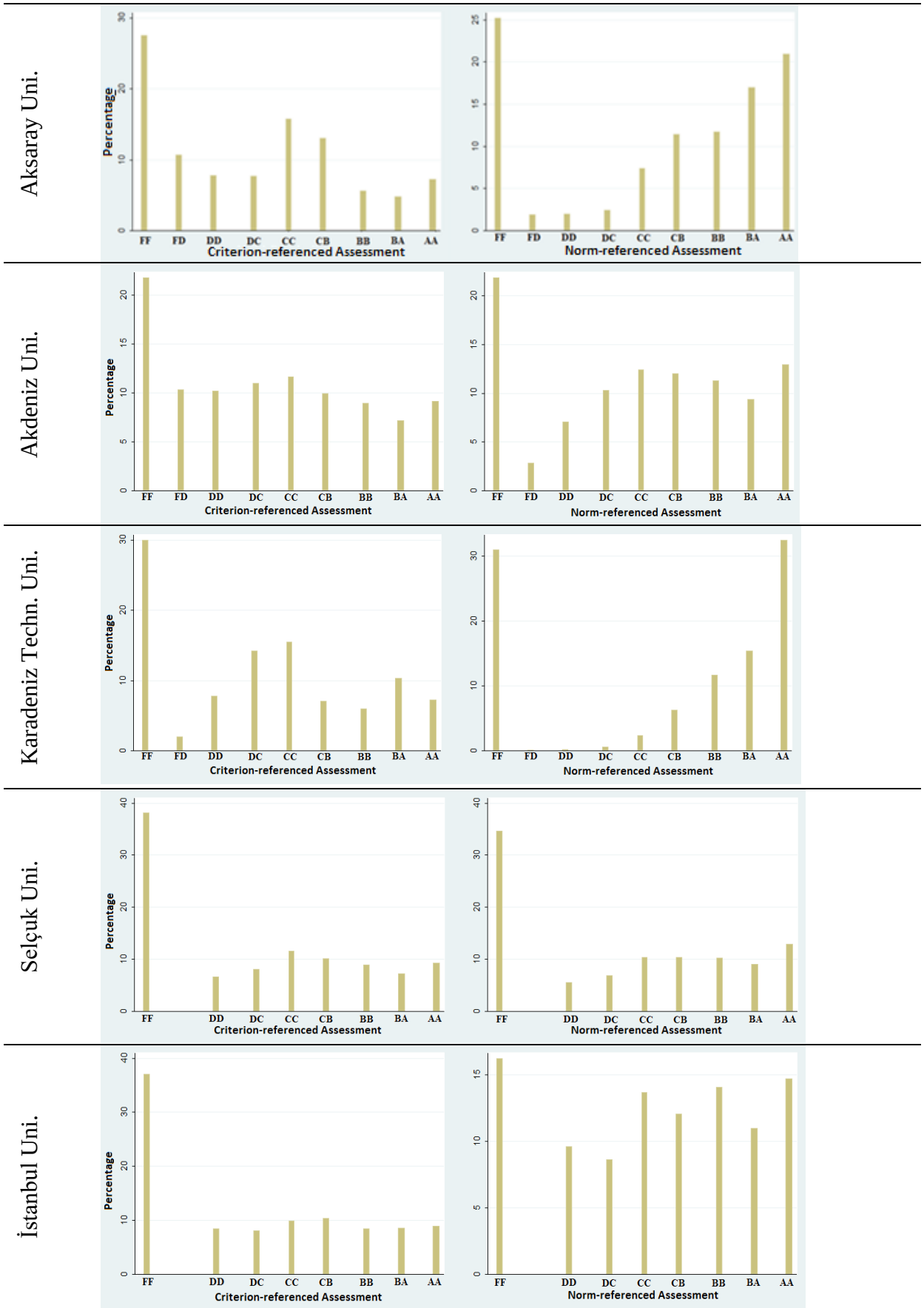

Figure 2. Criterion-Referenced Norm-Referenced Assessment Grade Distribution of the Universities

Figure 2 suggests that norm-referenced assessment generally increases letter grades of the students contrary to criterion-referenced assessment. As a result of the paired t-test, the norm-referenced assessment mean score of the students ($\bar{x}$=2.21, SD=1.51) has been noted to statistically differ from that of the criterion-referenced assessment ($\bar{x}$=1.59, SD=1.37) (t(789914)=-590.92, p<.05). Table 6 displays t-test results conducted for each university.

Table 6. t-Test Results Regarding Criterion-Referenced and Norm-Referenced Assessment for Different Universities

Universities	N	Criterion-referenced assessment		Norm-referenced assessment		df	*T*
		$\bar{x}$	SS	$\bar{x}$	SS		
Aksaray Uni.	157983	1.51	1.30	2.28	1.54	157982	-298.60*
Akdeniz Uni.	157983	1.68	1.33	1.97	1.37	157982	-271.74*
Karadeniz Technical Uni.	157983	1.62	1.34	2.41	1.69	157982	-335.21*
Selçuk Uni.	157983	1.57	1.44	1.77	1.50	157982	-129.00*
İstanbul Uni.	157983	1.58	1.44	2.20	1.32	157982	-374.76*

*p<.05

As shown in Table 6, the differences between the criterion-referenced and norm-referenced assessment scores for each university have been determined to be statistically significant and the scores increase in direction of norm-referenced assessment. ANOVA has been applied to explore the difference between the scores obtained from different norm-referenced assessment systems, which is related to other research question. Accordingly, a statistically significant difference has been determined ($F_{(4, 789910)}$=4632.88, p<.05). As a result of the post-hoc test (LSD), the scores calculated with the norm-referenced assessment score of each university are statistically different from those of the other university. The results are presented in Table 7.

Table 7. Comparison of Norm-Referenced Difference Scores among Universities

University (I)	University (J)	Difference (I-J)
Aksaray Uni.	Akdeniz Uni.	.312*
	Karadeniz Tech. Uni.	-.123*
	Selçuk Uni.	.516*
	İstanbul Uni.	.087*
Akdeniz Uni.	Karadeniz Tech. Uni.	-.434*
	Selçuk Uni.	.205*
	İstanbul Uni.	-.225*
Karadeniz Tech. Uni.	Selçuk Uni.	.634*
	İstanbul Uni.	.209*
Selçuk Uni.	İstanbul Uni.	-.430*

*p<.05

According to Table 7, the university which applies the most advantageous norm-referenced assessment system has been determined to be Karadeniz Technical University, which is followed by Aksaray University, Istanbul University, Akdeniz University, and Selçuk University, respectively.

RESULTS AND DISCUSSION

This research aims to explore whether students' letter grades differ across norm-referenced and criterion-referenced assessment methods and how this difference varies across universities. In this regard, the end-of-term raw achievement scores of 19,574 students who study at a state university during the fall term of 2014-2015 academic year have been converted into letter grades and 4-point

grading systems through use of both norm-referenced and criterion-referenced assessment regulations of the universities. After applying the paired samples t-test, the letter grades calculated via the norm-referenced assessment have been identified to be statistically significant and high compared to those calculated through the criterion-referenced assessment. In the following stage, ANOVA has been used in order to determine the difference between the letter grades obtained by using the norm-referenced assessment systems of the universities and the result has been found to be statistically significant. Specifically, students with the same RSG appear to have different letter grades in different universities.

The research findings have notably shown that the students' letters grades decrease as standard deviation in the norm-referenced assessment systems and low cut-off scores in criterion-referenced assessment systems are used. To exemplify, considering the lower grades obtained through Selçuk University norm-referenced assessment system, criterion-referenced assessment system are used for classes with RSG mean of over 70 and/or standard deviation below 8. Likewise, criterion-referenced assessment is used for the classes whose RSG mean is over 60 in Akdeniz University norm-referenced assessment system, while the same system is used for classes whose RSG mean is over 90 and/or standard deviation is below 8 in İstanbul University. In the norm-referenced assessment system of Aksaray University, criterion-referenced assessment is not applied depending on the RSG mean, whereas criterion-referenced assessment is used for the classes with 80 and over RSG meaning the norm-referenced assessment system applied by Karadeniz Technical University. However, 80-89 scores in the system applied by this university refer to BA and the scores above 90 signify AA letter grade, which increases students' letter grades.

When the research results are considered in general terms, the norm-referenced assessment has been determined to be much more in favor of students' letter grades compared to the criterion-referenced assessment. A similar result has emerged in the study conducted by Sayın (2016); however, different results have been found by Atılgan et al. (2012). This may result from the different use of the norm-referenced assessment algorithms and the small size sample group.

Based on the results of this research, two different evaluations can be made in terms of students and academic staff. On the basis of the student's perspective, students have higher grade point averages with the norm-referenced assessment than criterion-referenced assessment. This paves the way for the fact that norm-referenced assessment will lead to grading inflation as indicated by Başol (2015) in the related studies. Besides, different norm-referenced assessment systems reveal that the same RSG has been converted into different letter grades, meaning that the universities applying norm-referenced assessment system are more advantageous compared to the others. Both the difference between the norm-referenced and criterion-referenced grading system and the difference between the norm-referenced assessment systems may cause injustice. The concrete indicator of this situation occurs when the students apply to graduate programs. The degree effect of the GPA is used up to 40% in some universities is the evidence of the injustice of assessment systems used in the universities.

Upon analyzing the research results in terms of academic staff, the norm-referenced assessment system will be able to reduce the grading errors that will arise from the structure of the tests prepared by the academic staff. Considering that the academic staff may be lacking preparing a sufficient test or question technique, errors emerging due to the structure of the test, such as misinterpretation of the question or not being included in the current program, will cause students to get poor grades. However, since the findings of the previous studies (Nartgün, 2007; Sayın, 2016) and this study provide the conclusion that the norm-referenced assessment increases the grades of students, the norm-referenced assessment is likely to convert these lower grades, especially those of academic staff, into higher letter grades. The point to be noted here is that some students pass through the class without deserving it or being in high letter ranges. Thus, it is preferable to have an adequate level of the students included in the norm-referenced assessment and to determine the letter intervals rigorously.

What is more, it is of high significance to decide whether to use norm-referenced or criterion-referenced assessment depending on the purpose, structure, and results of the exams in terms of efficient measurement and assessment. Turgut and Baykul (2015) have noted that the scores will be

distributed symmetrically when there is a normal distribution; otherwise, the grades of the students will be largely affected by the other people in the class. Because the extreme values in both the right skewed and left skewed grades will change the mean of the distribution as well as increasing the standard deviation (Turgut & Baykul, 2015).

Besides, the use of criterion-referenced assessment will lead to different results in the case of the skewness of the raw score distributions. As the grade point averages will be high in the left-skewed distributions, many people in the class will pass with high letter grades, while in a class with a right-skewed distribution, the grades will be lower, thus the letter grades will be low and many students may fail. Given the skewness of the grades derives from the fact that the exam questions are too difficult or too easy, it is probable that the academic staff does not prepare qualified questions in terms of measurement and assessment. Thorndike (2005) draws attention to the fact that while preparing a qualified test, 25% of the questions should be difficult, 50% of them are at a medium level and 25% easy. Thus, the distribution of the grades will be closer to the normality.

Yücel (2015) has stated that the national exams in Turkey such as university entrance for which norm-referenced assessment is used measure the objectives at the level of remembering, understanding and applying levels. The use of blueprint in the exams prepared by the academic staff may affect the distribution of grades. In particular, writing the questions that will measure the cognitive gains of the students in all levels will determine how much the student has learned. The most commonly used type of question, open-ended questions and project-based assignments, which are the most commonly used type of questions, will provide the students with the objectives that need to be gained in the upper levels, namely, analysis, synthesis, and evaluation. It is expected that the number of questions will be higher in the exams prepared in this direction, as a result of which both the scope validity of the exam will increase and the grades obtained as a result of the application of the questions will be expected to be distributed normally. In other words, the exam consisting of questions related to all cognitive levels (knowledge, comprehension, application, analysis, synthesis, and evaluation) will undoubtfully affect the difficulty of the questions, and as Thorndike (2005) stated, it will cause the questions to be distributed normally in terms of the degree of difficulty. In short, the preparation of an exam within the framework of measurement and evaluation will affect the grade distribution, and consequently, exams can be evaluated through criterion-referenced assessment without the need for norm-referenced assessment.

REFERENCES

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Atılgan, H., Yurdakul, B., & Oğretmen, T. (2012). A research on the relative and absolute evaluation for determination of students achievement. *Inonu University Journal of the Faculty of Education, 13*(2), 79-98.

Başol, G. (2013). *Eğitimde ölçme ve değerlendirme*. Ankara: Pegem Akademi Yayınları.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Büyüköztürk, Ş. (2009). Sosyal bilimler için veri analizi el kitabı: İstatistik, araştırma deseni, SPSS uygulamaları ve yorum (9. baskı). Ankara: Pegem Yayınları.

Çelen, Ü., & Aybek, E. C. (2013). Öğrenci başarısının öğretmen yapımı bir testle Klasik Test Kuramı ve Madde Tepki Kuramı yöntemleriyle elde edilen puanlara göre karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *4*(2), 64-75.

Demir, P. & Atalmış, E.A. (2017). Öğretmen Yazılı Sınav Sorularının Hess Bilişsel Zorluk Matrisine Göre İncelenmesi. Köse, Selçuk, & Atalmış (Eds.), *Sosyo Ekonomik Stratejiler III – Eğitim* içinde (s. 43-73). Londra: IJOPEC Publication.

Duman, B. (2011). The views of classroom teachers related to norm-referenced assessment. *Education Sciences, 6*(1), 536-548.

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education, 10*(2), 133-143. DOI 10.1007/s10459-004-4019-5.

Field, A. (2009). *Discovering statistics using SPSS*. Thousand Oaks, CA: Sage publications.

Güler, N. (2017). *Eğitimde ölçme ve değerlendirme*. Ankara: Pegem Akademi Yayınları.
Haladyna, T.M. & Rodriguez, M.C (2013). *Developing and validating test items*. New York: Routledge.
Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research, 48*(1), 1-47.
Johnson, V. E. (2003). Grade inflation: a crisis in college education. New York, NY: Springer.
Lodico, M. G., Spaulding, D. T., & Voegtle, K. H. (2010). *Methods in educational research: From theory to practice*. San Francisco, CA: John Wiley & Sons.
Martin, I. G., & Jolly, B. (2002). Predictive validity and estimated cut score of an objective structured clinical examination (OSCE) used as an assessment of clinical skills at the end of the first clinical year. *Medical education, 36*(5), 418-425. https://doi.org/10.1046/j.1365-2923.2002.01207.x
Masters, J. C., Hulsmeyer, B. S., Pike, M. E., Leichty, K., Miller, M. T., & Verst, A. L. (2001). Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. *Journal of Nursing Education, 40*(1), 25-32. https://doi.org/10.3928/0148-4834-20010101-07
Mehrens, W.A. & Lehmann, I.J. (1991*). Measurement and evaluation in education and psychology.* New York: Harcourt Brace.
Nartgün, Z., (2007). Aynı puanlar üzerinden yapılan mutlak ve bağıl değerlendirme uygulamalarının notlarda farklılık oluşturup oluşturmadığına ilişkin bir inceleme. *Ege Eğitim Dergisi, 8* (1), 19- 40.
Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education, 30*(2), 175-194. https://doi.org/10.1080/0260293042000264262
Sayın, A. (2016). The Effect of using relative and absolute criteria to decide students' passing or failing a Course. *Journal of Education and Training Studies*, *4*(9), 1-9. http://dx.doi.org/10.11114/jets.v4i9.1571
Selvi, K. (1998). Üniversitelerde uygulanan başarı değerlendirme yaklaşımları. *Kurgu Dergisi*, *15*, 336-345
Tan, Ş. (2015). Öğretim hedeflerinin belirlenmesi. Şeref Tan (Ed.), *Öğretim ilke ve yöntemleri* içinde *(s.38-76)*. Ankara: Pegem Akademi Yayınları.
Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse education in practice*, *6*(6), 354-363. DOI:10.1016/j.nedt.2006.07.006
Thorndike, R.M. (2005). *Measurement and Evaluation in Psychology and* Education (7th Ed.). Upper Saddle River, NJ: Pearson Education.
Turgut, M. F., & Baykul, Y. (2015). *Eğitimde ölçme ve değerlendirme*. Pegem Akademi Yayıları.
Yorke, M. (2011). Summative assessment: dealing with the 'measurement fallacy'. *Studies in Higher Education, 36*(3), 251-273. https://doi.org/10.1080/03075070903545082
Yücel, C. (2015). Sınıf İçi Değerlendirme ve Not Verme. Emin Karip (Ed.), *Ölçme ve değerlendirme* içinde (s. 324-361). Ankara: Pegem Akademi Yayınları.

Türkiye'de Yükseköğretimde Kullanılan Bağıl Değerlendirme Sistemlerinin İstatistiksel Olarak Karşılaştırılması

Giriş

Eğitim ve öğretimin amacı, kazandırılmak istenen bilişsel, duyuşsal ve psikomotor becerileri öğrenciye planlı ve programlı bir şekilde sunmaktır. Bunu sağlamak için öncelikle dört temel öğeden oluşan öğretim programının belirlenmesi gerekmektedir. Öğretim programının bu öğeleri (a) hedef ve davranışların belirlemek, (b) içeriği bu hedef ve davranışlara tutarlı ve öğrencilerin hazırbulunuşluklarına uygun olarak yapılandırmak, (c) her öğrenci farklı öğrenir düşüncesiyle öğrenme ve öğretme aktiviteleri oluşturmak ve (d) anlamlı bir ölçme ve değerlendirme yapmaktır (Tan, 2015). Özellikle programda belirlenen hedef ve davranışların öğrencilerin hazırbulunuşluklarına ne derece sahip olduğu ve yine öğrenme ve öğretme aktivitelerinin hedef ve davranışlara ne derece uygun olduğunu belirlemede ölçme ve değerlendirmenin önemi göz ardı edilememektedir.

Ölçme ve değerlendirme kavramları sürekli beraber kullanılmasına rağmen, birbirinden farklı ve hatta birbirini tamamlayan kavramlardır. Özellikle ölçme bir değişkeni veya bir nesneyi sembollerle ifade ederken, değerlendirme ise ölçmeden elde edilen sonuçları bir ölçütle kıyaslayarak bu sonuçların anlamlaştırılmasını sağlamaktadır. Özellikle değerlendirmenin önemli bir öğesi olan ölçütü belirlemek oldukça karmaşık ve problemli bir süreç olduğu kabul edilmektedir. Ölçüt belirlemenin öğretmen kanısına, sınıftaki öğrenci başarı dağılımına, öğrenci yeteneğine, öğrencinin programdaki erişisine

(programın başındaki ve sonundaki öğrenme farkı) ve programın hedeflerine göre değişeceği önceki çalışmalarda belirtilmektedir (Turgut & Baykul, 2015; Martin & Jolly, 2002; Yorke, 2011). Bu durumlar göz önüne alındığında, doğru bir değerlendirme yapabilmenin en önemli unsurlarından bir tanesi uygun bir ölçütün seçilmesidir. Zaten değerlendirme kullanılan ölçüte göre sınıflandırılmaktadır. Mutlak ölçütün kullanıldığı diğer bir ifade ile grup ve grup özelliklerine bakılmaksızın herkes tarafından aynı şekilde kabul edilen değerlendirmeye mutlak değerlendirme (ölçüt dayanaklı) adı verilmektedir. Grup ve özellikle grubun başarı ortalamasına bağlı olarak seçilen ölçüte bağıl ölçüt ve yapılan değerlendirmeye de bağıl değerlendirme (norm dayanaklı) denilmektedir.

Türkiye'de devlet üniversitelerine bakıldığında yarıdan fazla üniversitede not verme sistemi olarak bağıl değerlendirme kullanıldığı görülmektedir. Mutlak değerlendirme kullanan üniversitelere bakıldığında farklı geçme ve kalma notlarına sahip olduğu ve harf aralıklarının farklı şekilde belirlendiği görülmektedir. Türkiye' deki bağıl değerlendirme sistemi kullanan üniversitelerin yönergeleri incelendiğinde sınıftaki öğrenci sayısı, sınıf not ortalaması, yüzdelik dilimleri ve notların standart sapmasına göre farklı yöntem ve algoritmalar kullandığı görülmektedir. Bazı üniversiteler CC notunu belirlerken sınıf ortalamasının yanında öğretim elemanlarının müdahale etmesine de müsaade etmektedir (Ankara Üniversitesi ve İstanbul Teknik Üniversitesi). Bunun yanı sıra bağıl değerlendirme kullanan birçok üniversitede bağıl değerlendirme sistemi uygulanırken öğrencilerin T puanının hesaplanarak sınıftaki not aralığına göre mutlak ya da değerlendirme sistemi kullanılmaktadır (Akdeniz Üniversitesi, Aksaray Üniversitesi, Bartın Üniversitesi, Bitlis Eren Üniversitesi, Bursa Teknik Üniversitesi, Bülent Ecevit Üniversitesi, Ege Üniversitesi, Fırat Üniversitesi, Hitit Üniversitesi, Karamanoğlu Mehmet Bey Üniversitesi, Kırıkkale Üniversitesi, Kilis 7 Aralık Üniversitesi, Mehmet Akif Ersoy Üniversitesi, Muğla Sıtkı Koçman Üniversitesi, Muş Alparslan Üniversitesi, Tunceli Üniversitesi, Uşak Üniversitesi gibi). Yine bu T puanı hesaplama yöntemini uygulayan üniversitelerin bazılarında ise öğrencilere verilen harf notları öğrencilerin sınıf içerisindeki yüzdelik dilimleri ile beraber düşünülerek verilirken (Artvin Çoruh Üniversitesi, Atatürk Üniversitesi, Balıkesir Üniversitesi, Celal Bayar Üniversitesi, Cumhuriyet Üniversitesi, Çankırı Karatekin Üniversitesi, İzmir Katip Çelebi Üniversitesi, Kafkas Üniversitesi, Karadeniz Teknik Üniversitesi, Marmara Üniversitesi, Namık Kemal Üniversitesi, Niğde Üniversitesi, Ondokuz Mayıs Üniversitesi, Süleyman Demirel Üniversitesi, Trakya Üniversitesi gibi) bir kısmında ise bunlara ilave olarak sınıfın standart sapması göz önünde bulundurularak verilmektedir (Selçuk Üniversitesi, Yalova Üniversitesi gibi). T puan yöntemini kullanmayan bazı üniversitelerde ise sınıftaki öğrenci sayısı ve sınıf not ortalamasının yanında sınıftaki not dağılımının standart sapmasına göre mutlak ya da bağıl değerlendirme sistemi uygulanmaktadır (İstanbul Üniversitesi, Çukurova Üniversitesi, Harran Üniversitesi, İnönü Üniversitesi, Kahramanmaraş Sütçü İmam Üniversitesi gibi). Yine bu üniversitelerde harf aralıklarının alt ve üst sınırları üniversite yönetimi tarafından belirlenmektedir.

Önceki çalışmalardan farklı olarak mevcut çalışmada Türkiye'de yaygın olarak kullanılan ve yukarıda bahsedilen 4 farklı bağıl değerlendirme sistemi karşılaştırılmıştır. Bu bağlamda sadece T puan dönüşümü uygulayan üniversiteler arasından her sınıf düzeyi için bağıl değerlendirme kullanan Aksaray Üniversitesi ve üst düzey sınıflar için mutlak değerlendirme sistemi uygulayan Akdeniz Üniversitesi'nin bağıl değerlendirme sistemi (ham başarı puanların (HBP) ortalaması 60 üstü sınıflar için) örneği seçilmiştir. T puan dönüşümü ve yüzdelik dilim yöntemi kullanılan üniversiteler arasından Karadeniz Teknik Üniversitesi'nin bağıl değerlendirme sistemi örneği seçilirken; T puan puan dönüşümü, yüzdelik dilim ve standart sapma kullanan üniversiteye örnek olarak ise Selçuk üniversitesi bağıl değerlendirme sistemi örneği seçilmiştir. Son olarak T puan dönüşümü kullanmayıp sadece standart sapma tabanlı bir dönüşüm kullanan üniversite olarak ise İstanbul bağıl değerlendirme sistemi örneği seçilmiştir.

Bu bağlamda aşağıdaki sorulara cevap aranacaktır:

1. Bağıl değerlendirme ve mutlak değerlendirme ile hesaplanan öğrenci harf notları arasında istatistiksel olarak fark var mıdır?
2. Farklı bağıl değerlendirme sistemleri ile hesaplanan öğrenci harf notları arasında istatistiksel olarak anlamlı bir fark var mıdır?

Yöntem

Araştırmada veri seti olarak 2014-2015 güz dönemindeki bir devlet üniversitesindeki farklı fakülte ve yüksekokullarındaki tüm öğrencilerin vize ve final notları kullanılmış, toplam 19,574 öğrencilerin HBP'leri veri analizi için göze önüne alınmıştır.

Öğrencilerin HBP'leri ara sınavın %40' ı final sınavının %60' ı alınarak hesaplanmıştır. Ardından öğrencilerin HBP'leri yukarıda bahsedilen Aksaray Üniversitesi, Akdeniz Üniversitesi, Karadeniz Teknik Üniversitesi, Selçuk Üniversitesi ve İstanbul Üniversitesi'nin uygulandığı bağıl değerlendirme sistemine göre uyarlanarak harf notu olarak hesaplanmıştır. Ayrıca bu çalışmada öğrencilerin HBP'leri yine bu üniversitelerin mutlak değerlendirme yönergelerine göre yeniden harf notuna dönüştürülmüştür. Her bir öğrencinin hem bağıl değerlendirmeyle hesaplanan hem de mutlak değerlendirme ile hesaplanmış harf notu ardından 4'lük sistemdeki not karşılığına çevrilmiştir. Öğrencilerin 2014-2015 güz döneminde birden fazla ders aldığı düşünüldüğünde aynı işlem her bir öğrencinin aldığı tüm derslerin HBP'ları için yapılmıştır. Diğer bir ifade ile bu süreç 19,574 öğrencinin 157,983 adet harf notu için yapılmıştır.

Öğrencilerin mutlak ve bağıl harf notlarının 4'lük sistemindeki karşılıkları hesaplandıktan sonra, ilk araştırma sorusu olan bağıl ve mutlak değerlendirme arasındaki farkı bulmak için tekrarlı ölçümlerindeki değişimi araştıran eşleştirilmiş iki grup arasındaki farkların testi (paired t-test) yöntemi kullanılmıştır. İkinci araştırma sorusu için yukarıda adı geçen herbir üniversitenin bağıl değerlendirme sistemiyle elde edilen harf notları arasındaki farklar ise tek yönlü varyans analizi (ANOVA) yöntemi ile bulunmuştur.

Sonuç ve Tartışma

Bu çalışmanın amacı öğrencilerin harf notlarının bağıl ve mutlak değerlendirme kullanılarak farklılık gösterip göstermediğini ve farklı bağıl değerlendirme sistemleri ile hesaplanan öğrenci harf notları arasında istatistiksel olarak anlamlı farklılık olup olmadığını bulmaktır. Bu bağlamda bir devlet üniversitesinde okuyan 19,574 öğrencinin her bir ders için dönem sonu ham başarı puanları farklı bağıl ve mutlak değerlendirme sistemine göre harf notuna ve ardından 4'lük sistemdeki not karşılığına çevrilmiştir. Araştırma sonunda, bağıl değerlendirme ile hesaplanan harf notların mutlak değerlendirme ile hesaplanan notlara göre istatistiksel olarak anlamlı ve yüksek olduğu elde edilmiştir. Ayrıca farklı bağıl değerlendirme sistemleri ile hesaplanan öğrenci harf notları arasında istatistiksel olarak anlamlı bir farklılık olduğu bulunmuştur.

Özellikle çalışmanın bulgularından üniversitelerin bağıl değerlendirme sistemlerinde standart sapma kullanımının yanında mutlak değerlendirmenin kullanılacağı kesme puanları düştükçe öğrenci harf notlarının düştüğü görülmektedir. Örneğin, Selçuk üniversitesi bağıl değerlendirme sistemi ile elde edilen notların düşük olması göz önüne alındığında, bu sistemin HBP ortalaması 70 üstü ve/veya standart sapması 8'in altında olan sınıflar için mutlak değerlendirme kullanıldığı görülmektedir. Yine Akdeniz üniversitesi bağıl değerlendirme siteminde HBP ortalaması 60 üstü sınıflar içinde mutlak değerlendirme kullanılırken, İstanbul üniversitesinde ise HBP ortalaması 90 üstü ve/veya standart sapması 8'in altında olan sınıflar için mutlak değerlendirme kullanıldığı görülmektedir. Aksaray üniversitesi bağıl değerlendirme sisteminde ise mutlak değerlendirme uygulaması HBP ortalamasına göre uygulanmamakta, Karadeniz Teknik üniversitesinin uyguladığı bağıl değerlendirme sisteminde ise HBP ortalaması 80 üstü sınıflar içinde mutlak değerlendirme kullanılmaktadır. Ancak bu üniversitenin uyguladığı sistemde 80-89 puanlar BA ve 90 üstü puanlar ise AA harf notuna dönüşmekte ve bu durum öğrencilerin harf notlarını artırmaktadır.

Genel olarak elde edilen sonuçlar düşünüldüğünde, bağıl değerlendirmenin mutlak değerlendirmeye göre harf notu olarak öğrenci lehine çalıştığını ortaya çıkmaktadır. Aynı sonuç yakın zamanda Sayın (2016) yılında yapılan çalışma ile desteklenmesine rağmen Atılgan ve diğerlerinin (2012) bulduğu sonuç ile farklılık göstermektedir. Bu durum Atılgan ve diğerlerinin (2012) çalışmasında kullandığı bağıl değerlendirme algoritması ile mevcut çalışmadaki kullanılan bağıl değerlendirme

algoritmalarından farklı ve daha az bir örneklem ile yapılmasından kaynaklanabileceği şeklinde açıklanabilir.

Bu çalışmanın sonuçlarından hareket ederek biri öğrenci açısından diğeri ise üniversitedeki öğretim elemanı açısından iki farklı yorum yapılabilir. Öğrenci açısından bakıldığında mutlak değerlendirmeye göre bağıl değerlendirme ile öğrenciler daha yüksek not ortalamasına sahip olurlar. Bu durum daha önceki çalışmalarda Başol (2015)' un ifade ettiği gibi bağıl değerlendirmenin not enflasyonuna sebep olacağı görüşünü desteklemektedir. Yine bu çalışmanın bulgularından hareketle farklı bağıl değerlendirme sistemleri aynı HBP'yi farklı harf notuna dönüştüğünü ortaya çıkarmakta, bu durum bağıl değerlendirme sistemi uygulayan bazı üniversitelerin diğer üniversitelere göre daha avantaj sağladığı düşünülebilir. Gerek mutlak ve bağıl değerlendirme not sistemi farklılığı, gerekse kullanılan bağıl değerlendirme sistemleri arasındaki farkın adaletsizliğe sebep olabilmektedir. Bu durumun somut göstergesi özellikle üniversite mezuniyet sonrasında öğrencilerin lisansüstü eğitime başvurularda ortaya çıkmaktadır. Bu başvurularda lisans mezuniyet ortalamasının etki derecesi bazı üniversitelerde %40'ı bulduğu düşünüldüğünde mezun olunan üniversitede kullanılan değerlendirme sistemlerinin adaletsizliğe ne derece sebep olduğu görülmektedir.

Çalışmanın sonuçlarına öğretim elemanları açısından bakıldığında bağıl değerlendirme sistemi öğretim elemanlarının hazırladıkları testlerin yapısından kaynaklanacak not hatalarını azaltabilecektir (Duman, 2011). Özellikle öğretim elemanlarının yeteri düzeyde test ya da soru hazırlama tekniğinden yoksun olabileceği düşünülürse, sorunun yanlış anlamlandırılması ya da mevcut programda olmaması gibi testin yapısından kaynaklanan hatalar her bir öğrencinin notunu düşürecektir. Ancak gerek önceki çalışmaların (Nartgün, 2007; Sayın, 2016) gerekse bu çalışmanın bulguları bağıl değerlendirmenin öğrenci notlarını artırdığı sonucunu desteklediğinden, bağıl değerlendirme özellikle öğretim elemanlarından kaynaklanan bu düşük notları daha yüksek harf notlarına dönüştürebilmesi muhtemeldir. Ancak bağıl değerlendirme ile bu düşük notların harf aralıkları öğrenciler lehine olacaktır. Burada dikkat edilmesi gereken nokta, bazı öğrencilerin hak etmedikleri halde dersten geçmeleri ya da yüksek harf aralığına düşmeleridir. Bu sebepten dolayı bağıl değerlendirmeye giren öğrencilerin yeteri düzeyde olması ve harf aralıklarının titizlikle belirlenmesi tercih edilmektedir.

Bunların yanı sıra herşeyden önce yapılan sınavın amacı, yapısı ve sonuçlarına bağlı olarak mutlak değerlendirme mi yoksa bağıl değerlendirme mi kullanılmasına karar vermek doğru bir ölçme ve değerlendirme açısından oldukça önemlidir. Turgut ve Baykul (2015) özellikle bağıl değerlendirme kullanılan dağılımın normal dağılım olması durumunda verilecek notların da simetrik olarak dağılacağını, aksi durumda öğrencilerin aldığı notların sınıftaki diğer kişilerden fazla etkileneceğini ifade etmektedir. Çünkü gerek sağa çarpık gerekse sola çarpık notlardaki uç değerleri hem dağılımın ortalamasını değiştirecek hem de standart sapmayı artıracaktır.

Yine ham puan dağılımlarının çarpık olması durumunda mutlak değerlendirme kullanılması ise farklı sonuçları doğuracaktır. Sola çarpık dağılımlarda sınıf not ortalaması yüksek olacağından sınıftaki birçok kişi yüksek harf notları ile geçerken, sağa çarpık dağılıma sahip bir sınıfta ise sınıftaki notlar düşük olacağından verilen harf notları da düşük olacak hatta birçok kişi dersten başarısız olabilecektir. Notların çarpık dağılım göstermesi sınav sorularının çok zor ya da çok kolay sorular sorulmasından kaynaklanacağı göz önüne alındığında öğretim elemanının ölçme ve değerlendirme adına nitelikli soruların hazırlanmadığı düşünülebilir. Thorndike (2005) nitelikli bir test hazırlarken soruların güçlük derecelerin dengeli olmasına yani soruların %25' inin zor, %50' sinin orta zorlukta ve %25' inin kolay olmasına dikkat çekmektedir. Böylelikle notlar dağılımı normale daha da yaklaşabilecektır.

Yücel (2015)'in ifade ettiği Türkiye'de bağıl değerlendirmenin kullandığı üniversite giriş sınavı gibi ulusal sınavlara bakıldığında bilişsel düzey bakımından alt düzey yani bilgi, kavrama ve uygulama düzeyindeki hedefleri ölçtüğünü söylenebilir. Özellikle sınıf içinde değerlendirmelerde öğretim elemanların hazırladıkları sınavları belirtke tablosu kullanması sınav sonucunda oluşacak not dağılımını etkileyebilecektir. Özellikle öğrencilerin bilişsel kazanımları ölçecek soruların tüm basamakları kapsayacak şekilde yazılması öğrencinin hangi hedefi ne derece öğrendiğini diğer bir ifade ile verilen bilgide ne kadar derinleştiğini ölçecektir. Bunun içinde en çok kullanılan soru tipi olan çoktan seçmeli yerine açık uçlu sorular ve proje tabanlı ödevler verilerek öğrenciye kazandırılması gereken hedefler üst basamaklara diğer bir ifade ile analiz, sentez ve değerlendirme basamağına

çıkaracaktır. Bu doğrultuda hazırlanan sınavlardaki soru sayısı fazla olması beklenip, bunun sonuncunda sınavın hem kapsam geçerliliğinin artması hem de soruların uygulanması sonucunda elde edilen notların da normal olarak dağılması beklenecektir. Daha açık bir ifade ile sınavın öğrencilerin bilişsel kazanımlardaki tüm basamakları (bilgi, kavrama, uygulama, analiz, sentez ve değerlendirme) kapsayacak şekildeki sorulardan oluşması soruların güçlük dereceleri etkileyecek ve Thorndike (2005)' in ifade ettiği gibi soruların güçlük derecesi bakımından dengeli bir şekilde dağılmasına sebep olacaktır. Kısacası bir sınavın ölçme ve değerlendirme çerçevesinde hazırlanması not dağılımını etkileyecek ve bunun sonucunda bağıl değerlendirmeye ihtiyaç duyulmayarak mutlak değerlendirme ile sınavlar değerlendirilebilecektir.