International Journal of Assessment Tools in Education
2023, Vol. 10, No. 4, 796–812
https://doi.org/10.21449/ijate.1266808
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

The Complexity of the Grading System in Turkish Higher Education

Recep Gur[1*], **Mustafa Koroglu**[2]

[1]Eskişehir Osmangazi University, Faculty of Education, Department of Measurement and Assessment, Eskişehir, Türkiye
[2]Erzincan Binali Yıldırım University, Faculty of Education, Department of Educational Sciences, Erzincan, Türkiye

ARTICLE HISTORY
Received: Mar. 17, 2023
Revised: Nov. 08, 2023
Accepted: Dec. 13, 2023

Keywords:
Higher education,
Grading system,
Fleiss' kappa,
Criterion-referenced assessment,
Norm-referenced assessment.

Abstract: Based on the academic performance grades of university students, various high-stakes decisions are made, including determinations of pass/fail status, the awarding of diplomas, and eligibility for placement in graduate education programs. According to the criteria used, the types of assessment are divided into two assessment, criterion-referenced assessments and norm-referenced assessments. When the grading system of state universities in Turkish higher education is examined, it has been observed that some universities use criterion-referenced assessment, some use norm-referenced assessment, and some use both assessment systems. The purpose of this research is to examine whether inter-university grading systems show significant concordance in the context of university students' letter grades or not. In other words, it is to reveal whether there are skew in the grading systems of public universities. In this context, 250 individuals were simulated in a way that their class/group achievement level would show a normal distribution. Among the public universities in the 2021-2022 Academic Performance Ranking of Universities (URAP), four state universities were determined in the first quarter, second quarter, third quarter, and last quarter. The letter grades of each student's academic success grade in the relevant universities were determined and it was examined whether there was a significant concordance between the letter grades of the students. In the study, it was concluded that in the context of university students' letter grades, inter-university grading systems generally do not show significant concordance. The findings are expected to contribute to the work of the Council of Higher Education and the University Education Commissions.

1. INTRODUCTION

Teaching practices across various levels of the education system aim to equip individuals with skills in cognitive, emotional, and psychomotor domains through activities tailored to specific programs. In the planning of educational and training processes, four fundamental elements of educational programs are considered. These elements include setting objectives, defining content, organizing educational scenarios (learning-teaching processes), and designing measurement and evaluation activities (Demirel, 2007). Measurement and evaluation are crucial for assessing whether the objectives outlined in the educational program align with

*CONTACT: Recep GÜR verianalizirehberi@gmail.com Eskişehir Osmangazi University, Faculty of Education, Department of Educational Sciences, Ekişehir, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

individuals' readiness levels, and whether the strategies, methods, and techniques employed in determining their educational status are consistent with their achievements (Atalmış, 2019).

One of the objectives of measurement and evaluation studies in education is to grade students based on their academic achievements in schools (Kubiszyn & Borich, 2000). For this purpose, grades serve as academic performance indicators and play an indispensable role in decisions made during the measurement and evaluation processes. These grades, given to assess students' achievement levels in relation to targeted learning outcomes, can be described as metrics for making judgments about students (Turgut & Baykul, 2015).

These grades can be defined in different ways (4-point system; 100-point system; AA, BA, ..., FF or A1, A2, ..., F3). Higher Education Institutions determine the grading systems they want to implement according to the decisions taken by the university senate. In many countries, while the 100-point system is frequently used in pre-higher education levels, different forms of application are observed in higher education levels. In Turkey's higher education system, lecturers evaluate students' academic achievement through various components. These include test scores, in-class and out-of-class performance, project tasks, and active participation in learning and teaching processes. Lecturers or university senate decides can decide how much weight (weighting percentage) each component will have in grading. For example, 40% of a student's grade is usually midterm exam scores and 60% final scores. It is observed that the regulations regarding the grading of higher education institutions in Turkey are examined, and the grades are generally given in the 4-point system. In addition, some universities include both 100-point and 4-point grading systems in students' transcripts (Özkan, 2016).

Various high-risk decisions (pass/fail, awarding diploma, placement in postgraduate education, etc.) are made based on the results of university students' academic achievement grades. Morever, academic achievement grades are used for admission to a school/programme, promotion to higher grades, graduation and feedback (academic progress for students, quality of education for teachers, status of their children for parents) (Ebel, 1965; Thorndike & Hagen, 1977). Therefore, academic achievement grades can also be effective in shaping the careers and future life of students (Moses & Nanna, 2007). On the other hand, the grades given to the students, regardless of the type of grading system, are the target of some criticisms because they cannot be measured directly. Considering the psychological effects (anxiety, anxiety) that grading systems have on stakeholders all over the world, regardless of letters or numbers, it is quite difficult to say that academic achievement grades are excellent in terms of reliability and validity in determining the level of achievement of individuals (Finkelstein, 1913). The high grades of the students who show the best performance or achieve the aim of the course at the highest level, if they are reliable and valid, can also serve the purpose of increasing students' motivation towards the course (Nitko & Brookhart, 2007). Therefore, grading systems that can evaluate students' performance with minimum error, make precise measurements and provide the opportunity to compare the resulting data should be preferred (Özkan, 2016).

1.1. Criterion-Based Assessments

In universities, assessments conducted to assign letter grades are referred to as "summative assessments," which is one category of assessment based on its intended purpose. On the other hand, the teacher's opinion, the ability of the individual (student), the objectives of the programme, the success level of the group to which the student belongs, the norms developed throughout the country, etc. can be used as criteria in evaluation processes (Turgut, 1983; Airasian, 1994; Haladyna, 1999). One of the critical points of the assessment process is that setting criteria is a very complex and problematic process. The choice of criterion and the amount (level) of the chosen criterion affect the decisions made (Kaya et al., 2017). Determining the criteria may vary according to the teacher's opinion, the success level of the student group who took the exam, the student's ability level, the student's gain development

between the end of the program and the beginning of the program, and the program learning outcomes (Martin & Jolly, 2002).

Criterion-based assessments fall into two categories. These are criterion referenced (absolute) assessment and norm-referenced assessment. In criterion-referenced assessment, before the measurement process is performed, the proficiency standard is determined and the success of the student is evaluated independently of the group's performance; in norm-referenced assessment, after the measurement process is made, the success of the student is evaluated according to the relative criterion obtained based on the success grades of the class/group (Kubiszyn & Borich, 2000; Atılgan et al., 2011). It is aimed to evaluate student performances norm-referenced to each other by determining norm-referenced criteria in line with the arithmetic mean and standard deviation scores based on the performances of the students participating in the exam. As a result of the assessment, students can get an idea about their learning levels and meeting the expectations of the students in this process can motivate them (Kaysi et al., 2017).

In the norm-referenced assessment, since the success of the student is determined by the success position among the other students in the class, a student with a good position in a group with a low average success rate may receive a high letter grade or not fail the course. As for criterion-reference assessment, regardless of whether the group's performance is low or high, if the student does not meet the proficiency standard or standard set determined before the measurement process, the student is considered unsuccessful (Kubiszyn & Borich, 2000; Özçelik, 1992). In this system, considering the standards set by the lecturers or institutions, students who reach the relevant standards are considered successful (Mandernach, 2003). Academic achievement grades determined within the scope of criterion-referenced assessment may depend on some factors arising from the student and the lecturer. Among the factors affecting students' academic achievement grades are the faculty members' ability to convey information, their ability to create a suitable psychological environment for students, their psychological state during the evaluation process, etc. Some student-dependent variables, likewise, the achievement level of the students, the approaches towards the lecture, and the lecturer, can also significantly affect the academic achievement grades of the students.

1.2. Differences in the Evaluation System Between Universities

Regulations regarding the assessment systems/grading system used by universities may differ. Beyond the differences in the assessment system between universities, there may be differences even in different faculties of the relevant university (Atılgan et al., 2012). Some universities may use complex scoring systems by including lettering and percentage systems etc. together in their regulations. Although the usage methods of grading systems differ on the basis of countries, the Higher Education Law No. 2547 is taken into account when determining the education, examination and grading systems in universities in Turkey. Within the scope of the law, it is stated that "The education and training carried out according to the characteristics and needs of the establishment in higher education institutions and the principles related to the diplomas which are awarded based on this are specified in the education and examination regulations to be prepared by each university". Higher education institutions based on this law make judgments about student achievements by using different grading systems and passing grades (Özkan, 2016). In many universities, norm-referenced or criterion-referenced grading systems are implemented using a grade range based on T score (Atalmış, 2019). Öztürk-Gübeş (2021) states that in grading systems, when calculating the composite score by combining different evaluations, multiplying the midterm and final raw scores with a ratio is a separate problem. Although the final and midterm assessments have different standard deviations and means, they are assumed to be on the same scale. The target weight of a score component and its real impact on grades can be very different. The stated reasons prevent the grades of

individuals who graduated from different higher education institutions from being comparable, but they may also cause measurement bias for students. Bias is defined as a systematic error that leads to the advantage or disadvantage of one group (Reynolds, Livingston, & Wilson, 2006). For example, considering two different university grading systems where it is easy and difficult to get an AA grade, when the grades are not comparable, this situation causes injustice and measurement bias problems arising from grading systems. The lack of a standard system in higher education institutions shows that the grades of university students who continue their education in different higher education institutions, which provide indicators of their academic success, can find direction in different values. The fact that universities and the Higher Education Council have different grade conversion tables-based on different passing grades and evaluation practices-complicates the matching of grades both within and between institutions. This lack of standardization creates issues in comparing the academic success levels of students, even if they graduate from the same programs.

Although the objectives such as determining the level of access of individuals to the attainments required to be achieved in the curriculum throughout the world, removing the individuals who fall below the specified standards from the system, and increasing the quality of education, make criterion-referenced assessment more common, higher education institutions in Turkey prefer to use norm-referenced assessment more. There are studies in the literature on how to use criterion-referenced or norm-referenced systems. Many studies in the literature focused on comparing criterion-referenced and norm-referenced rating systems (Başol- Mandernach, 2003; Göçmen, 2004; Nartgün, 2007; Duman, 2011; Lok et al., 2016; Özkan, 2016; Sayın, 2016; Atalmış, 2019). In addition to these, there are several studies examining the errors made in the weighting of the activities that are the subject of passing grades in the norm-referrenced assessment (Kelley & Zarembka, 1968; Tinkelman et al., 2013; Öztürk-Gübes, 2021). Although there are studies in the literature in which different grade-taking systems are compared and the advantages and disadvantages of criterion-referenced and norm-referrenced assessment systems are stated, the class averages and standard deviations are controlled with simulative data equally, and the letter of each student's academic achievement raw score in different quarters according to the URAP ranking of the universities. No study has been found on the concordance between the grading systems in higher education institutions/letter grades of students. The aim of this research is to investigate the level of concordance between different universities' grading systems, particularly concerning students' letter grades. In other words, it is to reveal whether there are skews in the grading systems of public universities.

2. METHOD

2.1. Research Model

This research aims to determine the letter grade equivalents of each student's academic achievement grade generated by the Monte Carlo method in the relevant universities and to examine the coefficients of concordance between the letter grades of the students. Therefore, this research is a Monte Carlo simulation study that seeks to answer the question "What would happen if each student's academic achievement grade was like this?" (Dooley, 2002).

2.2. Generating Data

Ethics committee decision is not required in this study, since the analysis was conducted on simulative data. In order to apply a norm-referrenced assessment, the grade distribution of the group should show a normal distribution. Due to the difficulty of achieving a normal distribution of academic grades in a small sample size, this study used simulated data for 250 individuals. The data were modeled to have a normal distribution in class/group achievement level with a mean (*M*) of 62.55 and a standard deviation (*SD*) of *12.53.*

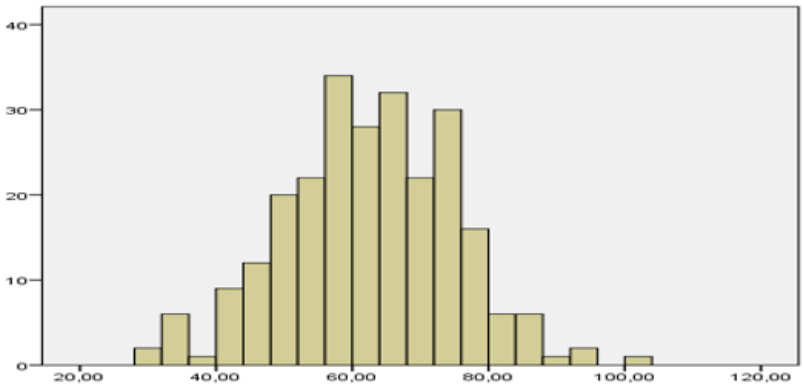
When the class achievement levels are examined according to the limit values of the raw achievement grade point averages (GPA) in the grading systems of the universities, the arithmetic average is around 62.50, since the class raw achievement GPA range of 60-65 is generally defined as the "middle" class level. Thus, in the Monte Carlo simulation process, care was taken to ensure that the simulated data represented real-world scenarios by using parameters to simulate 250 individuals with *M=62.55*, *SD=12.53*, whose class/group achievement level was normally distributed. In Monte-carlo simulation studies, a data set is created in accordance with the conditions specified by the researcher. Monte-carlo simulation studies diversify the data sets and provide an effective and fast comparison between grading systems etc. Simulative data were produced by using SimulCAT (Han, 2011) software. The descriptive statistics regarding the academic achievement scores of the students are given below.

Table 1. *Descriptive statistics on students' academic achievement scores.*

N	M	Median	Mode	SD	Min	Max	Skewness	SE_{skewness}	Kurtosis	SE_{Kurtosis}
250	62.55	62.26	56.35	12.53	28.71	100	-.08	.15	.03	.31

When Table 1 is examined, there are 250 students and their academic success grades are distributed between *28.71* and *100* (*M=62.55*, *SD=12.53*). In addition, it can be stated that the arithmetic mean, median and mode values are close to each other, the kurtosis and skewness coefficients are close to zero, the ratio of the skewness and kurtosis coefficients to the standard error is close to zero, the data show a normal distribution curve and the basic basis of norm-referenced assessment is provided. The distribution of students' academic achievement scores is given in Figure 1.

Figure 1. *Distribution of students' academic achievement scores*



When Figure 1 is examined, it shows the normal distribution curve characteristic of students' academic achievement scores.

2.3. Application Process

There are 120 state and 59 foundation universities among 179 universities in the Academic Performance Ranking of Universities in Turkey 2021-2022 (URAP). Among the URAP 2021-2022 public universities, Istanbul University (5th), Harran University (58th), Bartın University (90th), Kırklareli University (152nd) and four state universities were determined in the first quarter, second quarter, third quarter and last quarter, respectively (URAP, 2022).

In the grading system at Istanbul University, students with a raw achievement score (RAS) below 35 are automatically given an FF. On the other hand, students with an RAS of 100 are excluded from the Norm-Referenced Assessment System and receive an AA directly.Therefore, after examining the simulated data set between *28.71* and *100* (*M=62.55*, *SD=12.53*, *N=250*)

regarding student academic achievement grades, the letter grades of the students whose RAS were below *35* were directly FF; The arithmetic mean and standard deviation were calculated and found (*M=63.38, SD=11.27*) in order to calculate the T scores of the other students, without directly defining the letter grades of the students with an RAS of 100 and not including the raw grades of the related students in the Norm-Referenced Assessment System. Table 2 shows that class achievement level at Istanbul University (*M=63.38*) was defined as "Above Average" ($63 \leq \mu < 71$) according to the limit values of raw grade point averages. Since there are $N \geq 20$, *SD* (σ) ≥ 8, the assessment system at Istanbul University uses variable intervals based on the mean and standard deviation below (Istanbul University, 2022).

Table 2. *Variable intervals method in Istanbul University norm referenced assessment system.*

Letter Grade	Very Poor/Very Low: $\mu<44$	Poor/Low: $44 \leq \mu < 50$	Below Average: $50 \leq \mu < 56$
AA	$[\mu+1{,}881\sigma, 100]$	$[\mu+1{,}645\sigma, 100]$	$[\mu+1{,}476\sigma, 100]$
BA	$[\mu+1{,}405\sigma, \mu+1{,}881\sigma)$	$[\mu+1{,}175\sigma, \mu+1{,}645\sigma)$	$[\mu+0{,}994\sigma, \mu+1{,}476\sigma)$
BB	$[\mu+0{,}706\sigma, \mu+1{,}405\sigma)$	$[\mu+0{,}524\sigma, \mu+1{,}175\sigma)$	$[\mu+0{,}358\sigma, \mu+0{,}994\sigma)$
CB	$[\mu+0{,}332\sigma, \mu+0{,}706\sigma)$	$[\mu+0{,}126\sigma, \mu+0{,}524\sigma)$	$[\mu-0{,}075\sigma, \mu+0{,}358\sigma)$
CC	$[\mu-0{,}176\sigma, \mu+0{,}332\sigma)$	$[\mu-0{,}468\sigma, \mu+0{,}126\sigma)$	$[\mu-0{,}772\sigma, \mu-0{,}075\sigma)$
DC	$[\mu-0{,}643\sigma, \mu-0{,}176\sigma)$	$[\mu-0{,}878\sigma, \mu-0{,}468\sigma)$	$[\mu-1{,}126\sigma, \mu-0{,}772\sigma)$
DD	$[\mu-1{,}175\sigma, \mu-0{,}643\sigma)$	$[\mu-1{,}405\sigma, \mu-0{,}878\sigma)$	$[\mu-1{,}645\sigma, \mu-1{,}126\sigma)$
FF	$[35, \mu-1{,}175\sigma)$	$[35, \mu-1{,}405\sigma)$	$[35, \mu-1{,}645\sigma)$
Letter Grade	Average: $56 \leq \mu < 63$	Above Average: $63 \leq \mu < 71$	Good/High: $71 \leq \mu < 80$
AA	$[\mu+1{,}227\sigma, 100]$	$[\mu+0{,}915\sigma, 100]$	$[\mu+0{,}583\sigma, 100]$
BA	$[\mu + 0{,}739\sigma, \mu + 1{,}227\sigma)$	$[\mu + 0{,}385\sigma, \mu + 0{,}915\sigma)$	$[\mu + 0{,}100\sigma, \mu + 0{,}583\sigma)$
BB	$[\mu+0{,}126\sigma, \mu+0{,}739\sigma)$	$[\mu-0{,}075\sigma, \mu+0{,}385\sigma)$	$[\mu-0{,}305\sigma, \mu+0{,}100\sigma)$
CB	$[\mu- 0{,}358\sigma, \mu + 0{,}126\sigma)$	$[\mu-0{,}524\sigma, \mu-0{,}075\sigma)$	$[\mu-0{,}739\sigma, \mu-0{,}305\sigma)$
CC	$[\mu-0{,}878\sigma, \mu-0{,}358\sigma)$	$[\mu-0{,}994\sigma, \mu-0{,}524\sigma)$	$[\mu-1{,}126\sigma, \mu-0{,}739\sigma)$
DC	$[\mu-1{,}227\sigma, \mu-0{,}878\sigma)$	$[\mu-1{,}341\sigma, \mu-0{,}994\sigma)$	$[\mu-1{,}476\sigma, \mu-1{,}126\sigma)$
DD	$[\mu-1{,}751\sigma, \mu-1{,}227\sigma)$	$[\mu-1{,}881\sigma, \mu-1{,}341\sigma)$	$[\mu-2{,}054\sigma, \mu-1{,}476\sigma)$
FF	$[35, \mu-1{,}751\sigma)$	$[35, \mu-1{,}881\sigma)$	$[35, \mu-2{,}054\sigma)$
Letter Grade	Very Good/Very High: $\mu \geq 80$		
AA	$[\mu+0{,}440\sigma, 100]$		
BA	$[\mu- 0{,}100\sigma, \mu + 0{,}440\sigma)$		
BB	$[\mu-0{,}496\sigma, \mu-0{,}100\sigma)$		
CB	$[\mu-0{,}915\sigma, \mu-0{,}496\sigma)$		
CC	$[\mu-1{,}282\sigma, \mu-0{,}915\sigma)$		
DC	$[\mu-1{,}645\sigma, \mu-1{,}282\sigma)$		
DD	$[\mu-2{,}326\sigma, \mu-1{,}645\sigma)$		
FF	$[35, \mu-2{,}326\sigma)$		

When the grading system of Harran University is examined, students whose raw achievement score is below 35 are given FF directly; Students with an RAS of 90 and above are not included in the Norm-Referenced Assessment System by taking AA directly. When the grading system of Harran University (Harran University, 2022) is examined, the raw achievement score (RAS) is directly FF; Students with an RAS of 90 and above are not included in the Norm-Referenced Assessment System by taking AA directly. Therefore, the simulated data set between *28.71* and *100* (*M=62.55, SD=12.53, N=250*) related to student academic achievement scores was examined and the letter grades of the students whose RAS was below *35* were directly FF; The letter grades of the students with an RAS of *90* and above were directly defined as AA, and the arithmetic mean and standard deviation were calculated (*M=63.11, SD=10.76*) in order to calculate the T scores of the other students without including the raw achievement scores of the related students in the Norm-Referenced Assessment System. Table 3 indicates that, class achievement level at Harran University (*M=63.11*) was defined as "*Good*" ($60 \leq \mu < 70$) according to the limit values of raw grade point averages. Since there are $N \geq 20$, $SD \geq 8$, the assessment system at Harran University uses variable intervals based on the mean and standard deviation below (Harran University, 2022):

Table 3. *Variable intervals method in Harran University norm referenced assessment system.*

Letter Grade	Poor ($\mu < 50$)	Average ($50 \leq \mu < 60$)	Good ($60 \leq \mu < 70$)
AA	[μ + 3.00, 100]	[μ + 2.30σ, 100]	[μ + 1.50σ, 100]
BA	[μ + 2.60σ, μ + 3.00σ)	[μ + 1.90σ, μ + 2.30σ)	[μ + 1.10σ, μ + 1.50σ)
BB	[μ + 2.20σ, μ + 2.60σ)	[μ + 1.50σ, μ + 1.90σ)	[μ + 0.70σ, μ + 1.10σ)
CB	[μ + 1.30σ, μ + 2.20σ)	[μ + 0.80σ, μ + 1.50σ)	[μ + 0.30σ, μ + 0.70σ)
CC	[μ +0.40σ, μ + 1.30σ)	[μ + 0.10σ, μ + 0.80σ)	[μ – 0.10σ, μ + 0.30σ)
DC	[μ– 0.30σ, μ+0.40σ)	[μ – 0.80σ, μ +0.10σ)	[μ – 1.30σ, μ – 0.10σ)
DD	[μ – 1.00σ, μ – 0.30σ)	[μ – 1.70σ, μ – 0.80σ)	[μ – 2.50σ, μ – 1.30σ)
FF	<μ – 1.00σ	<μ – 1.70σ	<μ – 2.50σ

Letter Grade	Very Good ($70 \leq \mu < 80$)	Excellent ($\mu \geq 80$)
AA	[μ + 1.00σ, 100]	[μ + 0.50σ, 100]
BA	[μ + 0.65σ, μ + 1.00σ)	[μ + 0.20σ, μ + 0.50σ)
BB	[μ + 0.30σ, μ + 0.65σ)	[μ– 0.10σ, μ + 0.2σ)
CB	[μ– 0.05σ, μ + 0.30σ)	[μ – 0.40σ, μ–0.10σ)
CC	[μ – 0.40σ, μ– 0.05σ)	[μ – 0.70σ, μ – 0.40σ)
DC	[μ – 1.70σ, μ – 0.40σ)	[μ – 2.10σ, μ – 0.70σ)
DD	[μ – 3.00σ, μ – 1.70σ)	[μ – 3.50σ, μ – 2.10σ)
FF	<μ – 3.00σ	<μ – 3.50σ

250 individuals were simulated with a normal distribution of grade achievement level (M=*62.55, SD=12.53; Min=28.71, Max=100*). In line with these data, when the grading system of Bartın University (Bartın University, 2022) is examined, students with a RAS below 15 are not included in the Norm-Referenced Assessment System. Since the minimum RAS in the simulated study group was 28.71, the RAS of all students was included in the Norm-Referenced Assessment System. Therefore, the location parameters (*M=62.55, SD=12.53*) did not change in order to calculate the T scores of the students. In addition, as stated in the grading system of Bartın University, the RAS is directly defined as FF for students whose RASis below threshold limit 45. Table 4 shows that class achievement level at Bartın University (*M=62.55*) was defined as '*Average*' ($50 \leq \mu < 65$) according to the limit values of raw grade point averages. It is assumed that students' end-of-term (final) exam raw academic scores are at least 50. "Limit

scores of the criteria applied in determining the class achievement level" and "RAS limit values of letter grades according to class achievement level at Bartın University" are given below (Bartın University, 2022):

Table 4. *Limit scores of the criteria applied in determining the class achievement level.*

Criterion	Class Achievement Level				
	Very Poor	Poor	Average	Good	Very Good
Class RAS average lower limit	0	35	50	65	85
Class RAS average upper limit	34.99	49.99	64.99	84.99	100
Limit of inclusion in assessment	15	15	15	20	20
Threshold limit of RAS	45	45	45	50	50
End-of-term (final) exam RAS limit	50	50	50	60	60

Table 5. *RAS limit values of letter grades according to class achievement level at Bartın University.*

Letter Grades	Class Achievement Level				
	Very Poor	Poor	Average	Good	Very Good
AA	75	80	85	90	95
BA	70	70	75	80	85
BB	65	65	65	70	80
CB	60	60	60	65	75
CC	55	55	55	60	70
DC	50	50	50	55	65
DD	45	45	45	50	50
FF	<45	<45	<45	<50	<50

When the grading system of Kırklareli University (Kırklareli University, 2022) is examined, the norm-referenced assessment limit is 20 and students who fall below 20 are not included in the norm-referenced assessment system. 250 individuals were simulated with a normal distribution of grade achievement level (M=*62.55, SD=12.53; Min=28.71, Max=100*). In line with these data, since the minimum RAS in the simulated study group was *28.71*, the RAS of all students was included in the norm-referenced assessment system. Therefore, the location parameters (*M=62.55, SD=12.53*) did not change in order to calculate the T scores of the students. In addition, as stated in the grading system of Kırklareli University, students who score below the lower limit of success grade 40 points will be considered unsuccessful, and students who fall below 40 are directly defined as FF. Table 6 indicates that at class achievement level Kırklareli University (*M=62.55*) is defined as *'Very good'* ($62.5 \leq \mu < 70$) according to the limit values of raw grade point averages. Norm referenced assessment system Kırklareli Unıversity is given below (Kırklareli University, 2022):

Table 6. *Norm referenced assessment system of Kırklareli Unıversity.*

Class Level	Intervals over 100 (Class average)		Lower limits of norm refernced grades according to T-scores								
	Lower Limit	Upper Limit	AA (4.00)	BA (3.50)	BB (3.00)	CB (2.50)	CC (2.00)	DC (1.50)	DD (1.00)	FD (0.50)	FF (0.00)
Overachievement	80.00	100.00	57	52	47	42	37	32	27	22	<22
Excellent	70.00	79.99	59	54	49	44	39	34	29	24	<24
Very Good	62.50	69.99	61	56	51	46	41	36	31	26	<26
Good	57.50	62.49	63	58	53	48	43	38	33	28	<28
Above Average	52.50	57.49	65	60	55	50	45	40	35	30	<30
Average	47.50	52.49	67	62	57	52	47	42	37	32	<32
Poor	42.50	47.49	69	64	59	54	49	44	39	34	<34
Bad	0	42.49	71	66	61	56	51	46	41	36	<36

Although the class averages and standard deviations are simulated, the study finds variations in the grading systems of different universities. Specifically, the terms “Above Average”, “Good”, “Average”, and “Very Good” are defined differently across the four state universities examined.

2.4. Data Analysis

The letter grades of each student's academic achievement score/ RAS at the relevant universities were determined and whether there was a significant concordance between the letter grades of the students was examined by Cohen's kappa coefficient and Fleiss' kappa coefficient. While calculating the coefficient of agreement between the two evaluators/universities, Cohen's kappa coefficient is used; Fleiss' kappa coefficient is used in cases where the agreement between more than two raters is measured (Fleiss, 1971). The STATA 14 program was used to calculate Cohen's kappa coefficient between two universities and Fleiss' kappa coefficient between four universities. Kappa coefficients are suggested to be interpreted as follows (Landis & Koch, 1977; Fleiss, 1981):

Table 7. *The value ranges for Kappa coefficients.*

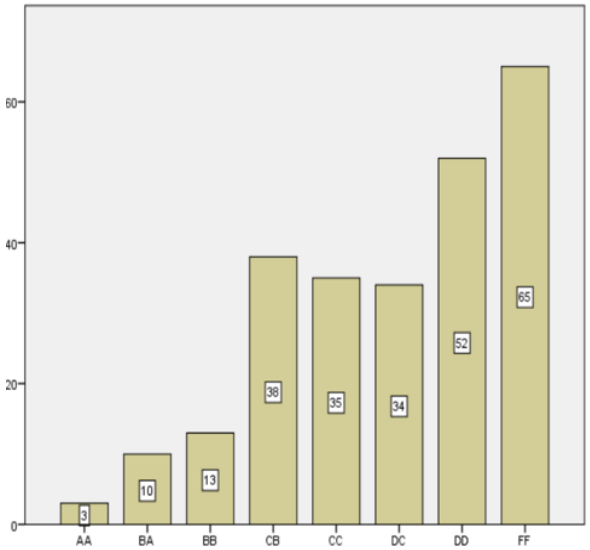
Coefficient	Interpretation
<.00	Poor
.00 to .20	Slight
.21 to .40	Fair
.41 to .60	Moderate
.61 to .80	Substantial
.81 to 1.00	Almost Perfect

Kappa coefficients are interpreted as “*poor*”, “*slight*”, “*fair*”, “*moderate*”, “*substantial*” and “*almost perfect*” respectively.

3. FINDINGS

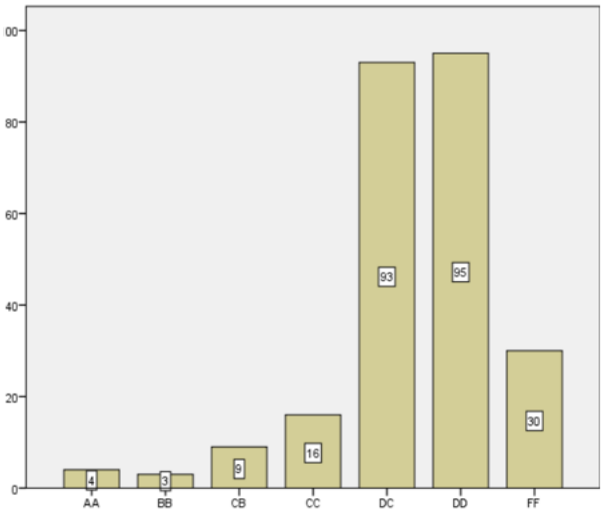
The distribution of the raw achievement score of each simulated student according to the grading system of Istanbul University regarding the letter grades of the students is presented in Figure 2.

Figure 2. *Distribution of students' letter grades according to the Istanbul University grading system.*



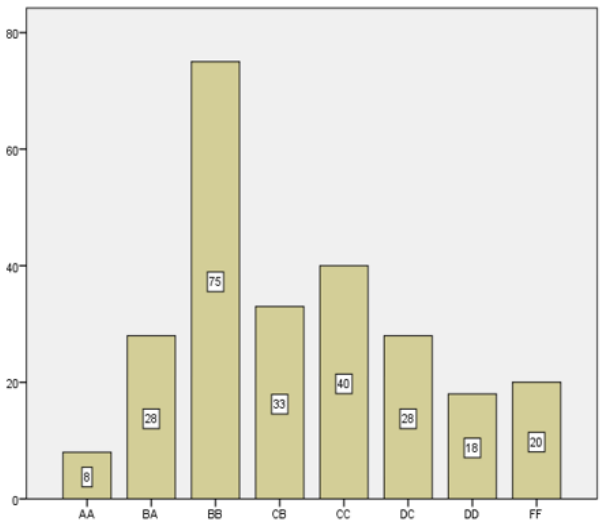
According to the grading system of Istanbul University, the number of students who received high letter grades AA (*f*=3), BA (*f*=10), and BB (*f*=13) is low. Most students have received DD (*f*=52) and FF (*f*=65) grades. In other words, student letter grades generally piled up to unsuccessful/low letter grades. The distribution of each student's raw achievement score according to the grading system of Harran University regarding the letter grades of the students is presented in Figure 3.

Figure 3. *Distribution of students' letter grades according to the Harran University grading system.*



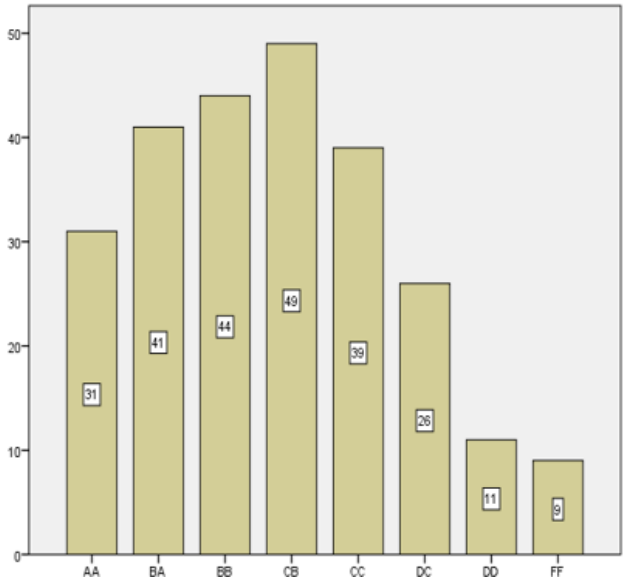
When the letter grade equivalents of each student's raw achievement score are examined according to the grading system of Harran University, it is seen that the number of students who receive AA (*f=4*), BB (*f=3*), CB (*f=9*) and CC (*f=16*) is low. It has been observed that students generally piled up on letter grades of DD (*f=93*), DD (*f=95*) and FF (*f=30*). In other words, student letter grades and letter grades have generally piled up on unsuccessful letter grades. Additionally, despite a class of 250 students, when the letter grade levels of the students are examined, it is notable that the grades of the students de-escalated (scree) from AA level to BB level, and that there was no student who received a "BA" letter grade. The distribution of the student's raw achievement score according to the grading system of Bartın University regarding the letter grades of the students is given in Figure 4.

Figure 4. *The distribution of hidden letter grades according to Bartın University grading system.*



When the letter grade equivalents of each student's raw achievement score are examined according to the grading system of Bartın University, it can be stated that while the number of students with a letter grade of BB and above is high, the number of students with a letter grade of DC or below is low. In other words, student letter grades are generally piled up successful/high letter grades. The distribution of each student's raw achievement score according to the grading system of Kırklareli University regarding the letter grades of the students is given in Figure 5.

Figure 5. *Distribution of students' letter grades according to the Kırklareli University grading system.*



When the letter grade equivalents of each student's raw achievement score are examined according to the grading system of Kırklareli University, it can be stated that while the number of students with a letter grade of CB and above is high, the number of students with a letter grade of DC or below is low. In other words, student letter grades generally piled up on successful/high letter grades. Distribution of students' letter grades according to the grading system of four different state universities (Istanbul University Q1, Harran UniversityQ2, Bartın University Q3, Kırklareli University Q4) ranked in the first quarter, second quarter, third quarter and last quarter, respectively, among URAP 2021-2022 public universities given in Table 8.

Table 8. *Distribution of students' letter grades according to the grading system of universities.*

	Istanbul University	Harran University	Bartın University	Kırklareli University
AA	3	4	8	31
BA	10	3	28	41
BB	13	-	75	44
CB	38	9	33	49
CC	35	16	40	39
DC	34	93	28	26
DD	52	95	18	11
FF	65	30	20	9

When the letter grade equivalents of each student's raw achievement score are examined according to the grading system of the universities, the number of students who received AA letter grades increased as we move across Istanbul University to Kırklareli University (from Q1 to Q4); The number of students who received FF letter grades decreased. For example, while

there are *3* students with AA and *65* students with FF at Istanbul University; Kırklareli University has *31* students with AA and *9* students with FF. In other words, student letter grades at Bartın and Kırklareli University have generally piled up on successful/high letter grades. When the letter grade coefficients of the universities are examined, they equal to *AA-4.00, BA-3.50, BB-3.00, CB-2.50, CC-2.00, DC-1.50, DD-1.00, FF-0.00* respectively. The difference in letter grade coefficients between universities is given in Table 9.

Table 9. *The difference in letter grade coefficients between universities.*

Letter Grade Coefficient Differences	Istanbul-Harran	Istanbul-Bartın	Istanbul-Kırklareli	Harran-Bartın	Harran-Kırklareli	Bartın-Kırklareli
-2.50	-	-	-	-	-	-
-2.00	-	-	19 %7.60	4 %1.60	55 %22	-
-1.50	-	50 %20	88 %35.20	111 %44.40	104 %41.60	-
-1.00	35 %14	125 %50	121 %48.40	75 %30	62 %24.80	11 %4.40
-0.50	1 %0.4	47 %18.80	10 %4.00	28 %11.20	16 %6.40	123 %49.20
0.00	111 %44.40	28 %11.20	12 %4.80	32 %12.80	13 %5.20	116 %46.40
0.50	55 %22	-	-	-	-	-
1.00	48 %19.20	-	-	-	-	-

Although the class averages and standard deviations are equal, it has been observed that there are differences of up to 2 coefficients between the letter grades of the students in different universities with the same raw score. For example, when the letter grade equivalents of students between raw achievement scores from *70.137* to *71.535* are examined, it corresponds to the letter grades of *CC* in Istanbul University, *DC* in Harran University, *BB* in Bartın University and *BA* in Kırklareli University. It is notable that the letter grades corresponding to the raw achievement score are different from each other, although the raw achievement score is between *70.137* and *71.535*, the class mean and standard deviations of all four state universities are controlled. The kappa coefficients showing the concordance between the grading systems of the universities are presented in Table 10.

Table 10. *Kappa coefficients between the grading systems of universities.*

Universities	Kappa	*p*
İstanbul-Harran (Q_1 x Q_2)	.325[a]	.00[*]
İstanbul-Bartın (Q_1 x Q_3)	-.002[a]	.92
İstanbul-Kırklareli (Q_1 x Q_4)	-.059[a]	.00[*]
Harran-Bartın (Q_2 x Q_3)	.034[a]	.07
Harran-Kırklareli (Q_2 x Q_4)	-.031[a]	.06
Bartın-Kırklareli (Q_3 x Q_4)	.374[a]	.00[*]
Q_1 x Q_2 x Q_3 x Q_4	.08[b]	.00[*]

a : Cohen Kappa coefficient
b : Fleiss' Kappa coefficient
[*] : $p<.05$

When Table 5 is examined, it is seen that in terms of letter grades of university students, inter-university grading systems generally do not show significant concordance and Kappa coefficients are poor (below .00), slight (between .00 and .20) or fair (between .20 and .40) level was found.

4. DISCUSSION and CONCLUSION

In this Monte Carlo simulation study, it was analysed whether there is a significant concordance between the letter grade equivalents of students' raw academic achievement scores. The study focuses on four public universities Istanbul University, Harran University, Bartın University, and Kırklareli University ranked in different quarters of the 2021-2022 URAP Academic Performance Ranking. As a result of the research, as one moves from the first quarter to the last quarter among the 2021-2022 URAP state universities, the number of students with *AA* letter grades increased; it was concluded that the number of students who received *FF* letter grades decreased.

Although the class averages and standard deviations of the students in different universities are controlled by simulative data according to the limit values of the raw grade point averages in the grading systems of the universities, they are classified as "*Above Average*", *"Good"*, *"Average"* and "*Very Good"*, respectively. All four of the four state universities have different definitions of the respective class achievement level. In addition, although the class averages and standard deviations are equal, the observation of differences up to *2* coefficients between the letter grades of students in different universities with the same raw score reveals the skewness between the grading systems of state universities. In the context of university students' letter grades, it has been concluded that inter-university grading systems generally do not show significant concordance and kappa coefficients are poor, slight or fair level.

Research findings show that there is generally no concordance between the grading systems of public universities. Although it is seen that this situation arises from the difference in the norm-referenced assessment algorithms used by the universities discussed in the study, it can be thought that it creates a bias in favor of the students studying at some universities that use thenorm-referencedassessment system. The difference between the criterion-referenced and norm-referenced assessment systems, as well as the injustice caused by the differences between the norm-referenced assessment systems in practice, directly affect the applications for graduate education or lateral transfer applications of graduate university students with their undergraduate graduation averages.

According to the general results of the study, the finding of differences between student letter grades stemming from the grading systems in the universities discussed in the research in the process of determining letter grades is similar to the studies in which students' letter grades are compared using different systems (criterion-referencedor norm-referenced assessment) (Mandernach, 2003; Başol-Göçmen, 2004; Nartgün, 2007; Duman, 2011; Lok *et al.*, 2016; Özkan, 2016; Sayın, 2016; Atalmış, 2019). While Lok *et al.* (2016) emphasize that criterion-referenced and norm-referencedassessment systems should be compatible and complementary; Kaya and Semerci (2017) receive opinions from lecturers about the positive and negative aspects of criterion-referenced and norm-referenced assessment systems. Sayın (2016) and Atalmış (2019) reached the conclusion that norm-referencedassessment received higher letter grades than criterion-referenced assessment, and that measures should be taken against the negativities of using the norm-referenced assessment system in Başol Göçmen (2004) and Mandernach (2003) studies. Duman (2011) stated as a result of the research that prospective classroom teachers have negative perceptions towards norm-referenced assessment. Differing from these findings, Atılgan *et al.* (2012) and Nartgün (2007) have stated that grading by using the criterion-referenced assessment system is a more accurate practice, however, Atılgan *et al.*

(2012) concluded that in a study where the norm-referenced assessment and criterion-referenced assessment are used to compare the obtained letter grades, in the case of a norm-referencedassessment, student letter grades will be approximately *40%* less than the letter grades obtained as a result of the criterion-referenced assessment.

When the studies showing that norm-referencedassessment increases student grades/causes grade inflation are examined, it can be explained that using norm-referenced assessment is advantageous in terms of instructors' inability to prepare questions in accordance with the principles of assessment and assessment, reducing the errors caused by the assessment tool, and not punishing the student for failure that may arise from lack of teaching. Turgut and Baykul (2015) state that in cases where the group distribution in the norm-referenced assessment is normal, the letter grades to be obtained will also be symmetrical, and in other cases, the letter grades to be taken will be more affected by the extreme values. Thorndike (2005) emphasizes that while preparing tests with appropriate psychometric properties, the difficulty levels of the items should be balanced and *25%* of the items should be above medium difficulty, *50%* moderate and *25%* below medium difficulty.

There have also been several studies (Kelley & Zarembka, 1968; Öztürk-Gubes, 2021; Tinkelman *et al*., 2013) examining errors in weighting of activities that are subject to passing grades in norm-referenced assessment. Öztürk-Gübes (2021) emphasized that the agreement between the grade values obtained by weighting according to the raw scores and the grade values that were weighted after standardization changed, and the fit between the grade values calculated by both methods decreased as the difference between the standard deviations of the midterm and final measurements increased. As for Özkan (2016), in the study titled 'Chaos in university graduation grades and conversion tables' concluded that The Council of Higher Education's grade conversion table provides a transformation in favor of the students in the universities with 50 passing grades, and against the students in the universities where 60 and 70 passing grades are applied. Özkan (2016) stated that the determination of the passing grades and the systems of the students who graduated from different higher education institutions cause problems in the grade conversion. He stated that the determination of the passing grades and the systems of the students who graduated from different higher education institutions cause problems in the grade conversion.

In line with these discussions, it shows that the grading systems of higher education institutions, which are applied in different ways based on the education and examination regulations specified in the higher education law numbered 2547 and prepared by the relevant commissions of different universities, and the determination of passing grades cause various problems. The lack fit between the grading systems of universities prevents the grades of individuals who graduated from different higher education institutions from being comparable, but also undermines the validity of the measurement results as it will cause measurement bias in favor of students in some universities. Students with an equivalent bachelor's degree are expected to practice an equivalent profession. Nevertheless, it is demonstrated as a crucial problem that the skewness caused by the grading systems of the universities from which the students graduated need to be considered and solved.

According to the academic achievement grades of university students, various high-risk decisions such as pass/fail, awarding diplomas, placement in graduate education, acceptance for transfer, etc. are made. Considering that these grades are effective in shaping students' careers and future lives, it is expected that there should be a standard grading system to ensure fairness among Higher Education Institutions. In order to eliminate the skewness between the grading systems of universities, the Education Commissions, which regulate the grading systems of universities, under the leadership of the Council of Higher Education, organize workshops, panels, etc., it is recommended to organize programs and make the necessary

arrangements for the standard grading system to serve its purpose, taking into account the qualifications expected from the graduates of the relevant faculty. The findings are expected to contribute to the work of the Council of Higher Education and the education commissions that regulate the grading systems of universities.

Acknowledgments

The authors would like to thank the blind reviewers for their valuable suggestions. While this study was under development, an abstract of some parts of the study was presented as an oral (abstract) paper titled "*The complexity of grading system in Turkish higher education*" by Gür, R. and Köroğlu M. at the VIIIth International TURKCESS Education and Social Sciences Congress held in Hatay on 21 July 2022.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Erzincan Binali Yıldırım University, Human Research and Educational Sciences Ethics Committee, 30/12/2022-12/06.

Authorship Contribution Statement

Recep Gur: Investigation, Resources, Methodology, Visualization, Software, Formal Analysis, and Writing-original draft. **Mustafa Koroglu:** Literature Review, Methodology, Supervision, and Writing-original draft.

Orcid

Recep Gur https://orcid.org/0000-0002-3686-4199
Mustafa Koroglu https://orcid.org/0000-0001-9610-8523

REFERENCES

Airasian, P.W. (1994). *Classroom assessment*. Mc Graw Hill. Inc. New York.

Atalmış, E.H. (2019). A statistical comparison of norm-referenced assessment systems usepar in higher education in Turkey. *Journal of Measurement and Evaluation in Education and Psychology, 10*(1), 12-29. https://doi.org/10.21031/epod.487335

Atılgan, H., Yurdakul, B., & Öğretmen, 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.

Atılgan, H., Kan, A., & Doğan, N. (2011). *Eğitimde ölçme ve değerlendirme* [*Measurement and evaluation in education*]. Anı Yayıncılık.

Bartin University (2022). Bartın Üniversitesi Bağıl Değerlendirme Sistemi Uygulama Yönergesi [Bartin University Relative Assessment System Implementation Instruction]. https://kms.kaysis.gov.tr/Home/Kurum/85269548

Basol Gocmen, G. (2004). *Değerlendirmeye genel bir bakış: Kriter-referanslı (mutlak) ya da norm-referanslı (bağıl) değerlendirme* [*An overview of evaluation: Criterion-referenced (absolute) or norm-referenced (relative) evaluation*]. XIII. Ulusal Eğitim Bilimleri Kurultayı'nda sunulmuş bildiri. İnönü Üniversitesi Eğitim Fakültesi, Malatya.

Demirel, Ö. (2007). *Eğitimde program geliştirme* [*Program development in education*]. Pegem A Yayıncılık.

Dooley, K. (2002). Simulation research methods. In J. Baum (Ed.), *Companion to organizations* (pp. 829-848). Blackwell.

Duman, B. (2011). Sınıf öğretmeni adaylarının bağıl değerlendirmeye ilişkin görüşleri [Opinions of classroom teacher candidates regarding relative assessment]. *NWSA-E Journal of New World Sciences Academy, 6*(1), 536-548

Ebel, R.L. (1965). *Measuring educational achievement*. Prentice-Hall, Inc.

Finkelstein, I.E. (1913). The marking system in theory and prac-tice. *Baltimore.* https://babel.hathitrust.org/cgi/pt?id=uc1.$b264457;view=1up;seq=9

Fleiss J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 7*, 378-382.

Fleiss, J.L. (1981) *Statistical methods for rates and proportions*. John Wiley, 38-46.

Haladyna, T.M. (1999). *A Complete guide to student grading*. Allyn and Bacon. A Viacom Company.

Han, K.T. (2011). *User's manual: SimulCAT*. http://www.umass.edu/remp/software/simcata/simulcat/SimulCAT_Manual.pdf.

Harran University (2022). Harran Üniversitesi Bağıl Değerlendirme Yönergesi [Harran University Relative Assessment Instruction]. http://ogrenci.harran.edu.tr/assets/uploads/other/files/ogrenci/files/Ba%C4%9F%C4%B1l_De%C4%9Ferlendirme_Sistemi_Ekim_2019_web.pdf

Istanbul University (2022). İstanbul Üniversitesi Önlisans, Lisans Ölçme ve Değerlendirme Esasları [Istanbul University Associate Degree, Undergraduate Measurement and Evaluation Principles]. https://cdn.istanbul.edu.tr/FileHandler2.ashx?f=olcme-degerlendirme.pdf.

Kaya, Ü., & Semerci, Ç. (2017). The opinions about relative and absolute assessment of teaching staff in the higher education. *The Journal of Academic Social Science, 5*(47), 457-467. https://doi: 10.16992/asos.12321

Kaysi, F., Bavli, B., & Gürol, A. (2017). Educational connoisseurship and criticism: evaluation of a cooperation model between university and the sector on vocational education. *Journal of Education and Practice, 8*(6), 25-35.

Kelley, A.C., & Zarembka, P. (1968). Normalization of student test scores: An experimental justification. *The Journal of Educational Research*, *62*(4), 160-164. https://www.jstor.org/stable/27532173

Keskin M., & Ertan H. (2001). *İstanbul Üniversitesi'nin bağıl değerlendirme sistemi kitapçığı* [*Istanbul University's relative assessment system booklet*]. İstanbul: İstanbul Üniversitesi.

Kırklareli University (2022). Kırklareli Üniversitesi Sınav ve Başarı Değerlendirme Yönergesi [Kırklareli University Examination and Success Assessment Instruction]. https://oidb.klu.edu.tr/Yardimci_Sayfalar/183-yonergeler.klu

Kubiszyn, T., & Borich, G. (2000). *Educational testing and measurement*: *Classroom application and practice* (6th ed.). John Wiley & Sons, Inc.

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

Lok, B., McNaught, C., & Young, K. (2016). Criterian-referenced and norm-reference assessments: Compatibility and complementarity. *Assessment & Evaluation in Higher Education, 41*(3), 450-465. https://doi.org/10.1080/02602938.2015.1022136

Mandernach, B.J. (2003). Effective grading strategies. *Park University Faculty Development Quick Tips*. https://www.park.edu/center-for-excellence-in-teaching-and-learning/

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

Moses, M.S., & Nanna, M.J. (2007). The testing culture and the persistence of high-stakes testing reforms. *Education and Culture*, 55-72.

Nartgün, Z. (2007). Aynı puanlar üzerinden yapılan mutlak ve bağıl değerlendirme uygulamalarının notlarda farklılık oluşturup oluşturmadığına ilişkin bir inceleme [An examination of whether absolute and relative evaluation practices based on the same scores create differences in grades]. *Ege Eğitim Dergisi, 8*(1), 19-40.

Nitko, A.J., & Brookhart, S.M. (2007), *Educational assessment of students*. Pearson Education Inc.

Özçelik, D.A. (1992). *Okullarda ölçme ve değerlendirme* [*Measurement and evaluation in schools*]. ÖSYM.

Özkan, Y.Ö. (2016). Üniversite mezuniyet notları ve dönüşüm tablolarında yaşanan kaos [The chaos in university graduation grades and conversion tables]. *Journal of Higher Education and Science, 6(1),* 71-76.

Öztürk-Gübeş, N. (2021). An investigation into weighting problem in norm-referenced grading system. *Eurasian Journal of Educational Research, 93,* 337-356. https://doi.org/10.14689/ejer.2021.93.16

Reynolds, C.R., Livingston, R.B., & Wilson, W. (2006). *The problem of bias in educational assessment in measurement and assessment in education*. Pearson Education.

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

Thorndike, R.M. (2005). *Measurement and evaluation in psychology and education*. Pearson Education.

Thorndike, R.L., & Hagen, E.P. (1977). *Measurement and evaluation in psychology and education*. John Wiley & Sons, Inc.

Tinkelman, D., Venuti, E., & Schain, L. (2013). Disparate methods of combining test and assignment scores into course grades. *Global Perspectives on Accounting Education, 10,* 61-80. https://gpae.wcu.edu/disparate-methods-of-combining-test-and-assignment-scores-into-course-grades/

Turgut, M.F. (1983). Program değerlendirme. *Cumhuriyet Döneminde Eğitim* [Program evaluation. Education in the Republican Era]. 215-234.

Turgut, M.F., & Baykul, Y. (2015). *Eğitimde ölçme ve değerlendirme* [*Measurement and evaluation in education*]. Pegem Akademi.

URAP (2022). University Ranking by Academic Performance. https://newtr.urapcenter.org