Eurasian Journal of Educational Research 93 (2021) 337-356

Eurasian Journal of Educational Research
www.ejer.com.tr

An Investigation into Weighting Problem in Norm-Referenced Grading System *

Neşe ÖZTÜRK GÜBEŞ[1]

ARTICLE INFO

Article History:
Received: 09 Jan. 2020
Received in revised form: 28 Dec. 2020
Accepted: 29 Mar. 2021
DOI: 10.14689/ejer.2021.93.16

Keywords
standard scores, norm-referenced grading, agreement

ABSTRACT

Purpose: In grading, one of the most common errors is made in combining two or more different test scores. This study aimed to investigate the agreement of grades calculated by weighting raw scores and standard scores. **Research Methods:** In this simulation study, data were simulated for midterm and final measurements. Nine conditions [3 (class level: poor, average, good) x 3 (standard deviation (SD) difference (0, 10, and 20 units)] were considered. The sample size for each measurement was taken as 60 and the replication number set to 100. The weight for midterm and final measurements was respectively taken as 40% and 60%. The students' norm-referenced grades were calculated in two ways: (1) based on T scores of weighted raw success scores (T_{WRSS}) and (2) based on T scores of weighted standardized success scores (T_{WSSS}). The agreement between T_{WRSS} and T_{WSSS} grades was calculated with the simple percentage agreement, extended (±1 grade) percentage agreement and kappa coefficient. The agreement between grades was compared by conducting two-way ANOVA. **Findings:** The results showed that the SD main effect was a significant effect on the agreement between grades. The maximum agreement was provided when midterm and final measurements had equal SD. The minimum was provided when the SD difference was at the highest level. **Implications for Research and Practice**: It was recommended that scores should be standardized before combined in the norm-referenced grading system. The effects of the shape of data (skewness or kurtosis) on norm-referenced grades could be investigated in the further studies.

*This study was presented at the 6th International Eurasian Educational Research Congress in Ankara, 19-22 June, 2019

[1] Burdur Mehmet Akif Ersoy University, TURKEY,e-mail: nozturk@mehmetakif.edu.tr, ORCID: 0000-0003-0179-1986

Introduction

Grades and grading are indispensable parts of the educational system. The grades are used for many actions and decisions in education. For instance, grades are used for attending upper classes in a school, for admission to some programs or departments, for getting scholarship aid, for admission to the college or for graduating from a school (Thorndike & Hagen, 1977), informing parents about their children's school progress (Linn & Gronlund, 1995) or reporting a student's educational progress to him/her, to his/her parents, to future teachers, and possible employers (Ebel, 1965). Grades also have an important role in increasing students' motivation towards the course. The highest grades should be given to those students who performed best or who gained the objective of course at the highest level (Ebel, 1965; Nitko & Brookhart, 2007). Grades can fulfill their functions if they are reliable and valid.

Grading is one of the most challenging duties of teachers. While assigning grades, teachers have to consider lots of factors and have to take many decisions at the same time. The grading system used by teachers can be classified under three categories: letter grades, quantitative grades and verbal descriptions (Popham, 2011). At universities, letter grades are commonly used. In the letter grading system, a single letter grade (A, B, C, D, F) for each lesson is assigned. While the "A" grade represents the maximum level of achievement, the "F" grade represents the minimum level of achievement. Sometimes by adding letters or ± signs besides grades, the scales of grade can be increased 5 to 15 levels. In the quantitative grading system, an achievement scale is defined based on chosen numbers (e.g., 100, 10, 5 or 4) and in the verbal grading system, verbal descriptions as "poor", "average" and "good" are used. An example of a quantitative and verbal grading system used in secondary schools in Turkey is given in Table 1 (Ministry of Education, 2016).

Table 1

An Example for Quantitative and Verbal Grades.

Degree	Score Range
Very Good	85.00-100.00
Good	70.00-84.99
Moderate	60.00-69.99
Pass	50.00-59.99
Fail	.00-49.99

The grading systems are also classified based on criteria: criterion-referenced grading and norm-referenced grading. While in the criterion-referenced grading system, each student is evaluated based on predetermined absolute standards, in the norm-referencing grading system, a prespecified percentage of students would have a definite grade (Lok, McNaught, & Young, 2016). Norm-referenced grading is also known as "grading on the curve" (Thorndike & Hagen, 1977, p. 599). The "curve" is substituted for "normal distribution". As Pontes (2018) reported, the normal distribution was first presented by Abraham de Moivre (1667-1754) in 1738. Moivre

named this curve normal because "its mean represents the norm, that is, things should all be like the mean; so that everything that deviates from this average is considered an error, where equivalence" (p. 29). In other words, most humans respectively similar to each other (for example, average height), with a minority of people are at extreme points very tall or very short (Mertler, 2007).

In grading on the curve, the system supposes that there are five letters (A-B-C-D-F) used to express achievement levels. A grade of "C" represents average achievement; "B" represents above-average achievement; "A" represents outstanding achievement; "D" represents below-average achievement and "F" indicates failure (Ebel, 1965). In this system, most students get "C" grade, somewhat lower, but equal number of students get "B" and "D" grades, a few but an equal number of students get "F" and "A" grades (Thorndike et al., 1991, p. 182). In the norm-referenced system, students are ranked based on their scores and this system assigns a fixed percentage for determining grades (Chan, 2014). For example, 7% of students get "A", 24% of students get "B", 38% of students get "C", 24 % of students get "D" and 7% of students get "F" (Thorndike & Hagen, 1977). Another approach in the norm-referenced system is assigning grades regarding the standard deviations (SD) from the mean. For example, the scores more than 1.5 SD above from the mean may get "A", the scores between .5 and 1.5 SD above the mean may get "B" and so on (Chan, 2014; Ebel, 1965).

Teachers may use many ways to assess the performance of students, such as test scores, performance on homework assignments, projects and class participation. After the component of grades, teachers decide how much each component is to be weighted. Generally, the weighting of components is pointed out in percentages. For example, midterm scores will constitute 40% and final scores will constitute 60% of a student's grade (Thorndike et al., 1991). One of the most common errors in grading is made while combining two or more different test scores (Francis, 2006; Linn & Gronlund, 1995; Nitko & Brookhart, 2007). This is called a weighting problem. The reason for this error comes from the spread of students' test scores in different tests. Test score variability is measured by the standard deviation (Francis, 2006). As Kelley and Zarembka (1968) stated, "…if two tests possess the same number of total points, and even the same mean score, but possess different "spreads", then the test which has the greater spread is more important determining the students' grade." (p. 160). Nitko and Brookhart (2007) pointed out that the component which has the larger standard deviation, the more influences the final ranking of students. In other words, the test which has a larger standard deviation or variability in the total scores would have more weight on the composite score (Cunningham, 2005).

In education, while a norm-referenced grading system is used, the scores which have more weight should affect the students' rank more. This principle is violated when different test scores multiply by a certain percentage or ratio (Nitko & Brookhart, 2007). If two tests result in the same number of scores, the tests are weighted the same. However, this assumption is only true if both distributions of test scores are on the same scale (Thorndike et al., 1991). As Francis (2006) stated, if we do not use a standard unit of measurement while combining different test scores in grading, we make the Mars Climate Orbiter Miscalculation. In December 1998, The Mars Climate Orbiter,

which cost $125 million launched by NASA for studying Martian Climate, Martian atmosphere, and surface changes. After about 10 months, the Mars Climate Orbiter burned and broke up. This was because of not using the right units. While the navigation team at the Jet Propulsion Laboratory (JPL) used the metric system (millimeters and meters) in its calculations, the designer and builder team Lockheed Martin Astronautics in Denver, Colorado, used the English system of inches, feet and pounds. JPL team did not pay attention to the unit conversation (Harish, 2019).

We cannot say two tests are on the same scale unless they have the same mean and standard deviation. The best way is to convert raw scores to standard scores which, means that they have the same mean and standard deviation. Scores can be standardized by a calculated z-score. If x is a score from a distribution that has mean μ and standard deviation σ, the z score of x is (Moore, McCabe, & Craig, 2009):

$$z = \frac{x-\mu}{\sigma} \qquad (1)$$

The z-score tells us how many standard deviations that the value x falls away from the mean and which direction. If values of x are larger than the mean, we have positive z-scores and values of x that are smaller than the mean, we have negative z-scores. If x equals the mean, then x has a z score of 0. Another way to express z-scores is by creating a standard normal distribution that has a mean zero and standard deviation of 1.0. Using z-scores is not desired because of two matters: (1) z-scores can get negative values and (2) z-scores can get values with decimals. Interpreting negative or decimal values in education may be quite difficult. Multiplying every z-score by a constant, such as 10 and by adding a convenient constant amount, such as 50, we can get out of negative and decimal z-scores (Thorndike et al., 1991). The name of this process is a linear conversion of z-scores. For instance, a z-score can be converted to a T-score, which has a mean of 50 and a standard deviation of 10:

$$T= 10\,(z) +50 \qquad (2)$$

In Turkey, while some of the universities use a criterion-referenced grading system, some of them use the norm-referenced grading system (Nartgün, 2007). In many universities, criterion-referenced or norm-referenced grading systems are applied using the grade interval, which is based on the T score (Atalmış, 2019). In grading systems, while combining different assessments to calculate a composite score, raw scores of midterm and final multiplied by a ratio. Although midterm and final assessments had different mean and standard deviations, it is assumed that they are on the same scale. The intended weight of a score component and its actual impact on grades may be very different. In grading, the combination of different assignments should be valid and properly weighted according to the instructor's intentions for the course (Brookhart, 1999). The weighting problem in grading is significant because it affects the fairness of the grading system and the hortative effects of grades. In the literature, most of the studies (Atalmış, 2019; Atılgan, Yurdakul, & Öğretmen, 2012; Başol-Göçmen, 2004; Kaya & Semerci, 2017; Lok, McNaught & Young, 2016; Nartgün, 2007; Sayın, 2016) have focused on comparing norm-referenced assessment with criterion-referenced assessment system. A few studies (Kelley & Zarembka, 1968;

Tinkelman, Venuti, & Schain, 2013) focused on the weighting problem in the norm-referenced grading system. These studies were conducted with actual data sets and the results of these studies showed that the choice of different grade combination methods was highly affecting individual students' grade (Tinkelmen et al., 2013) and the standard deviation variations of different tests had important effects on letter grades (Kelley & Zarembka, 1968). Unlike previous studies (Kelley & Zarembka, 1968; Tinkelmen et al., 2013), this study is a simulation study that investigates the effects of both mean and standard deviation on grades while weighting different assignments. It is thought that this research would draw the attention of instructors and decision-makers to improve the grading systems for weighting different assignments properly.

Aim of this Research

This study aims to investigate the agreement between grades calculated by weighting raw scores and standard scores. For this main aim, the research questions are:

1. How do the variation of midterm and final standard deviations and class level affect the agreement between grades which are calculated by weighting raw scores and standard scores?
2. What is the relationship between students' ranking of final-midterm scores and total scores calculated based on weighted raw scores and weighted standardized scores?

Method

Research Design

This study was a simulation study designed to investigate that the effects of standard deviation difference and class level on the agreement between grades, which were calculated by weighting raw scores and standard scores, compared under various conditions.

Data

This study was conducted using simulated data. Within the scope of this study, data were simulated for midterm and final measurements based on three achievement levels of class: poor, average, and good. For every level of class, three conditions were considered. For each condition, the means for midterm and final measurements were kept equal, but the standard deviation of midterm was changed. At the first condition, standard deviations of midterm and final measurements were kept equal; at the second condition, the standard deviation of midterm measurement was increased 10 units higher than the final measurement's standard deviation; at the third condition, the midterm measurement's standard deviation was increased 20 units higher than the standard deviation of final measurement. The means for class levels were chosen based on middle points of grade levels of the Uşak University Grading System. In the Uşak University Grading System, the mean for a good class is between 57.5-62.5, the mean for an average class is between 47.5-52.5 and the mean for a poor class is smaller than

42.5. Therefore, the values of 60, 50, and 40 were chosen for good, average and poor classes. While choosing a standard deviation value of 10 for the first condition, the standard deviation border for calculating T-scores in some universities (such as Akdeniz University) was considered. For instance, in the Grading System of Akdeniz University, the T score calculating formula is changing based on standard deviation smaller than 10 or larger than 10. The increasing value of 10 units for standard deviation was selected arbitrarily. This study aims to show the effect of standard deviation on weighting midterm and final scores on grades. The intended weight for the final assignment is 60% and for the midterm, 40%. To show how midterm assignment could affect grades because of larger standard deviation than final assignment although it has a smaller ratio, the standard deviations of midterm measures were chosen larger than final measures. The conditions used in simulating data are shown in Table 2.

Table 2

Data Simulation Conditions

	Condition 1				Condition 2				Condition 3			
	Midterm		Final		Midterm		Final		Midterm		Final	
Class Level	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
Poor	40	10	40	10	40	20	40	10	40	30	40	10
Average	50	10	50	10	50	20	50	10	50	30	50	10
Good	60	10	60	10	60	20	60	10	60	30	60	10

As seen in Table 2, nine conditions [3 (class level) x 3 (standard deviation difference)] were considered in this study. The data were simulated based on symmetric normal distribution using the "rsnorm" command in the "fGarch" (Wurtz et al., 2017) R package. The sample size for each measurement was taken as 60 and the replication number for each condition was set to 100.

Data Analysis

For this study, the weight for midterm and final measurements in a student's general achievement grade were respectively taken as 40% and 60%. The T score scale of the University of Uşak was used while determining grades. The students' grades were calculated in two ways. At first, by taking 40% of the midterm measurement and 60% of the final measurement, weighted raw success scores (WRSS) were calculated. Then, T scores [10z+50] for WRSS were calculated and the grades that corresponded to these T_{WRSS} were determined depending on the class' WRSS average. Second, T scores for midterm measurement and final measurement were calculated separately. Then, by taking 40% of the midterm's T score and 60% of the final's T score, the weighted standardized success score (WSSS) was calculated. After that, the grades which equal these T_{WSSS} were determined to depend on the class's WRSS average.

After calculating grades, the agreement between grades (which was determined based on T_{WRSS} and T_{WSSS}) was calculated using the simple percentage of agreement, extended percentage (±1 grade) agreement and kappa coefficient. The "irr" (Gamer, 2015) R package was used in computing agreement of grades. The simple percentage

agreement was a measure of consistency between two observers about the score (or grade). The simple percentage agreement was computed by summing the grades where two observers rated the same, dividing that total number of counts and then multiplying the result by 100. The advantage of this approach was easy to calculate and understand. The extended percentage agreement was the percentage of the observers agreed to within one point. In the extended percentage of agreement, the tolerance value was taken as 1 grade; in other words, 1 grade higher or 1 grade lower (±1 grade) agreement was calculated. A simple and extended percentage of the agreement did not account for the agreement with occurred by chance. Cohen (1960) proposed a statistic called kappa (κ) and kappa considers the chance agreement between observers. The equation for kappa is (as cited in Goodwin, Sands, & Kozleski, 1991):

$$K = \frac{P_o - P_c}{1 - P_c} \qquad (3)$$

In the equation above, P_o the observed proportion of agreement and P_c equals the proportion of agreement expected by chance. Theoretically, kappa can take values -1.00 to +1.00, where 1.00 means perfect agreement, .00 means observed agreement equals by chance and negative values indicate agreement less than chance. Vieara and Garett's (2005) criteria, which is shown in Table 3, was considered reporting kappa values in this study:

Table 3

Interpretation of Kappa

Kappa	Agreement
< .00	Less than chance agreement
.01 - .20	Slight agreement
.21 - .40	Fair agreement
.41 - .60	Moderate agreement
.61 - .80	Substantial agreement
.81 - 1.00	Almost perfect agreement

At the last step of data analyses, the relationship between students' ranking of final-midterm scores and total scores, which were calculated based on weighted raw scores and weighted standardized scores, were calculated using Spearman's rank-order correlation coefficient. The agreement between grades of T_{WRSS} and T_{WSSS} was compared by conducting a two-way analysis of variance (ANOVA) with the achievement level of class and different standard deviations of midterm measurement.

Results

To answer the first research question, the two-way ANOVA was applied. The effects of the standard deviation and class level on the agreement between grades based on T score of weighted raw success score (T_{WRSS}) and grades based on T score of weighted standardized success score (T_{WSSS}) were examined with considering three

agreement indexes (simple percentage agreement, extended percentage agreement, and kappa). The descriptive statistics for simple percentage agreement, extended percentage agreement and kappa coefficient are given in Table 4.

Table 4

Descriptive Statistics for Simple Percentage Agreement, Extended Percentage Agreement and Kappa Values

		Poor Class		Average Class		Good Class	
		Mean	SD	Mean	SD	Mean	SD
Simple Percentage Agreement	C1	59.52	8.43	58.45	7.53	60.52	8.81
	C2	46.48	8.25	45.13	8.84	47.55	8.73
	C3	36.95	7.36	38.00	7.19	36.22	6.56
Extended Percentage Agreement	C1	99.32	1.34	99.37	1.22	99.27	1.28
	C2	94.22	3.95	93.33	4.50	93.85	4.11
	C3	84.68	5.63	86.23	5.34	86.12	6.36
Kappa Values	C1	.51	.10	.50	.09	.53	.10
	C2	.36	.10	.34	.10	.37	.10
	C3	.25	.09	.26	.08	.24	.08

Note. C1: Condition 1; C2: Condition 2; C3: Condition 3

As seen in Table 4, for all agreement indexes at each class level (poor, average and good), while the maximum agreement was provided at Condition 1 when midterm and final measurements had an equal standard deviation, the minimum ones provided at Condition 3 when the standard deviation gaps between midterm and final measurements at the highest level (20 units).

The Results related to Simple Percentage Agreement

The two-way ANOVA results for simple percentage agreement are presented in Table 5. As seen in Table 5, there was a significant main effect of standard deviation [$F(2,899)=594.65$, $p<.05$]. In other words, when the class level was ignored, the standard deviation difference between midterm and final measurements affected the simple percentage agreement between T_{WRSS} and T_{WSSS} grades. As seen in Table 4, for all class levels (poor, average and good), while the maximum simple percentage agreement was provided when midterm and final measurements had an equal standard deviation, the minimum ones provided when the standard deviation gaps

between midterm and final measurements at the highest level (20 units). To analyze the effects of standard deviation on the simple percentage agreement of T_{WRSS} and T_{WSSS} grades, Tukey's post hoc test was used among the three conditions of standard deviation. The agreement between T_{WRSS} and T_{WSSS} grades for Condition 1 (where the standard deviations of midterm and final measurements were equal) was significantly higher than Condition 2 (where the standard deviation of midterm 10 units higher than final) and Condition 3 (where the standard deviation of vise 20 units higher than final). Also, there was a significant difference between Condition 2 and Condition 3. The agreement between T_{WRSS} and T_{WSSS} grades for Condition 2 was higher than Condition 3.

Table 5

The Two-Way ANOVA Results for Simple Percentage Agreement

Source	Sum of Squares	df	Mean Squares	F	p	Sig. Dif.
A (Std. Dev.)	76235.11	2	38117.56	594.65	.000*	1-2,1-3,2-3
B (Class Level)	121.53	2	60.77	.95	.388	
AxB	546.26	4	136.56	2.13	.075	
Error	57113.34	891	64.10			
Total	134016.25	899				

*p<.05

Table 5 also showed that there was not a significant main effect of the class level [F(2,899)= .95, p>.05]. We can say that when we ignored the standard deviation difference between midterm and final measurement, the class level did not influence the simple percentage agreement between T_{WRSS} and T_{WSSS} grades. Finally, the interaction between the effects of standard deviation and the class level was not statistically significant [F(4,899)=2.13, p>.05]. In other words, the effects of standard deviation on the agreement between T_{WRSS} and T_{WSSS} grades were not different for each class level.

The Results related to Extended Percentage Agreement

The agreement between grades which determined based on T_{WRSS} and T_{WSSS} was calculated using an extended percentage agreement. In the extended percentage of agreement, the tolerance value was taken as 1 grade; in other words, 1 grade higher or 1 grade lower (±1 grade) agreement was calculated. The two-way ANOVA results for an extended percentage agreement are shown in Table 6.

Table 6

The Two-Way ANOVA Results for an Extended Percentage Agreement

Source	Sum of Squares	df	Mean Squares	F	p	Sig. Dif.
A (Std. Dev	28240.68	2	14120.34	801.98	.000*	1-2,1-3,2-3
B (Class Level)	18.22	2	9.11	.52	.596	
AxB	170.71	4	42.68	2.42	.047*	
Error	15687.65	899	17.61			
Total	44117.26	900				

*$p<.05$

As seen in Table 6, there was a significant main effect of standard deviation [F(2,899)= 801.98, p<.05] on the extended agreement between T_{WRSS} and T_{WSSS} grades. In other words, when we ignored class level, the standard deviation difference between midterm and final measurements affected the extended percentage agreement between T_{WRSS} and T_{WSSS} grades. As seen in Table 4, for all class levels (poor, average and good), while the maximum extended agreement was provided when midterm and final measurements had an equal standard deviation, the minimum agreement was provided when the standard deviation difference between midterm and final measurements at the highest level (20 units).To analyze the effects of standard deviation on the extended percentage agreement of T_{WRSS} and T_{WSSS} grades, Tukey's post hoc test was used among the three conditions of standard deviation. The extended agreement between T_{WRSS} and T_{WSSS} grades for Condition 1 (where the standard deviations of midterm and final measurements were equal) was significantly higher than Condition 2 (where the standard deviation of midterm 10 units higher than final) and Condition 3 (where the standard deviation of vise 20 units higher than final). Also, there was a significant difference between Condition 2 and Condition 3. The agreement between T_{WRSS} and T_{WSSS} grades for the Condition 2 was higher than Condition 3.

Lastly, as seen in Table 6, there was not a significant main effect of class level [F(2,899)= .52, p>.05) on the extended agreement of T_{WRSS} and T_{WSSS} grades. We can say that when we ignored the standard deviation difference between midterm and final measurement, the class level did not influence the extended percentage agreement between T_{WRSS} and T_{WSSS} grades. Finally, there was a statistically significant interaction between the effect of standard deviation and class level [F(4,899)=2.42, p<.05]. However, the effect size for the interaction was very small ($\eta 2$=0.01) and also the interaction graph in Figure 1 showed parallel lines. Although there was an increase of agreement in Condition 3 and there was a small drop in Condition 2 after poor class, we can say that the effects of standard deviation (Condition 1, Condition 2, and Condition 3) were similar at all class levels (poor, average, and good) (see Figure 1). Also, as shown in Figure 1, at all class levels, an agreement between T_{WRSS} and T_{WSSS} grades of Condition 1 (where standard deviations of midterm and final measurements were equal) were higher than Condition 2 (where the standard deviation of midterm

10 units higher than final) and Condition 3 (where the standard deviation of midterm 20 units higher than final). In addition, the extended agreement of Condition 2 was higher than Condition 3.

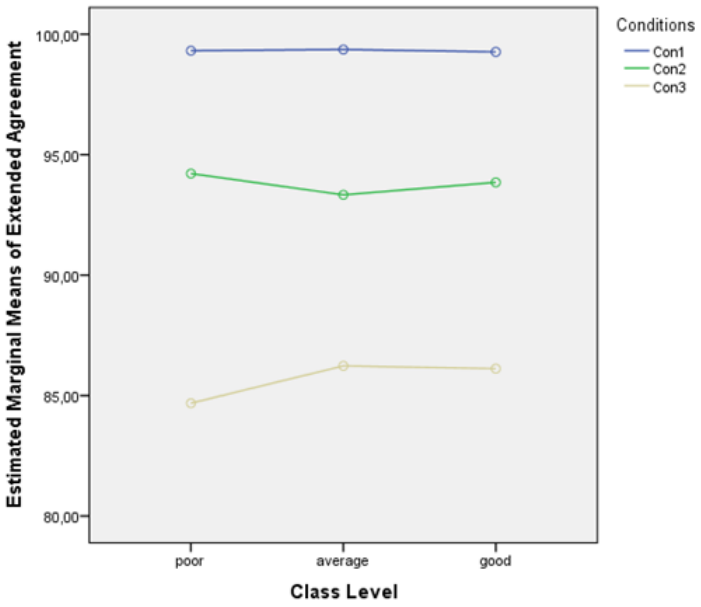

Figure 1. *Graph of Interaction between Standard Deviation Difference and Class Level for Extended Agreement*

The Results related to Extended Percentage Agreement

The two-way ANOVA results for kappa are shown in Table 7. The two-way ANOVA results for kappa values were similar to the simple percentage and extended percentage agreement results. As seen in Table 7, while there was a statistically significant main effect of standard deviation [$F(2,899)=596.06$, $p<.05$] on T_{WRSS} and T_{WSSS} grades agreement, there was not a statistically significant main effect of class level [$F(4,899)=.90$, $p>.05$] and interaction between the effect of standard deviation and class level [$F(4,899)=.02$, $p>.05$]. Based on a statistically significant standard deviation main effect, we can say that when we ignored class level, the standard deviation difference between midterm and final measurements affected the kappa agreement between T_{WRSS} and T_{WSSS} grades.

Table 7

The Two-Way ANOVA Results for Kappa Coefficient

Source	Sum of Squares	df	Mean Squares	F	p	Sig. Dif.
A (Std. Dev.)	10.79	2	5.39	596.06	.000*	1-2,1-3,2-3
B (Class Level)	.02	2	.01	.90	.407	
AxB	.08	4	.02	2.14	.074	
Error	8.06	891	.01			
Total	18.95	899				

*$p<.05$

As seen in Table 7, for all class levels (poor, average and good), while the maximum agreement was provided when midterm and final measurements had an equal standard deviation, the minimum agreement was provided when the standard deviation difference between midterm and final measurements was at the highest level (20 units). According to Viera and Garret's (2005) criteria for Kappa, when the standard deviations of midterm and final measurement were equal, there was a moderate agreement at all class levels, but as the standard deviation difference between midterm and final measurements increased the agreement decreased to a fair agreement. Tukey's post hoc test was used among the three conditions of the standard deviation to analyze the effect of standard deviation on the kappa agreement of TWRSS and TWSSS grades. There was a statistically significant difference between (1) Condition 1 - Condition 2, (2) Condition 1 - Condition 3 and (3) Condition 2- Condition 3.

The Results related to Spearman's Rank-Order Correlations

To answer the second research question of this study, Spearman's rank-order correlation between students' rank of midterm-final scores and weighted raw success scores (WRSS) and weighted standardized success scores (WSSS) were calculated. The results are presented in Table 8.

Table 8

Correlations between Students' Rank of Midterm-Final Scores and Composite Scores

	Class Level								
	Poor			Average			Good		
Correlations	Con 1	Con 2	Con 3	Con 1	Con 2	Con 3	Con 1	Con 2	Con 3
Midterm-WRSS	.53	.77	.88	.55	.78	.88	.53	.78	.87
Final-WRSS	.81	.58	.41	.81	.56	.45	.81	.58	.42
Midterm-WSSS	.53	.53	.52	.54	.53	.54	.53	.54	.51
Final - WSSS	.81	.81	.81	.81	.81	.82	.81	.82	.81

Note. Con1= condition 1; Con 2 = condition 2; Con 3 = condition 3

In this study, students' composite score was calculated by taking 40% of the midterm measurement and 60% of the final measurement. As seen in Table 8, for all class levels at Condition 1 (where the standard deviation of midterm and final measurement were equal), Spearman's rank-order correlations between final and WRSS were higher than correlations between midterm and WRSS. For instance, if we considered the poor class and Condition 1, the correlation between midterm-WRSS was .53 and the correlation between final and WRSS was .81 (see Table 8). We can say that when final and midterm measurements had equal standard deviations, the final measurement (which has a bigger weight in total raw scores) affected students' rank more. However, when the difference between standard deviations of midterm and final measurements was increased as in Condition 2 and Condition 3 at all class levels, correlations between midterm measurement and WRSS were higher than correlations between final measurement and WRSS. For example, if we considered the poor class and Condition 2, the correlation between midterm-WRSS was .77 and the correlation between final and WRSS .58 (see Table 8). In this case, we can say that, although the final measurement has more weight on student's composite scores, midterm measurement affected students' rank more.

As seen in Table 8, at all conditions and all class levels, the Spearman's rank-order correlations between midterm and WSSS were lower than the correlations between final and WSSS. For example, for the poor class at Condition 2, the correlation between midterm-WSS was .53 and the correlation between final-WSS was .81 (see Table 8). We can say that the final measurement that has more weight on WSSS affected students' rank more than midterm measurement at all class levels and conditions.

Discussion, Conclusion and Recommendations

In grading, one of the most common errors is made in combining two or more different test scores. While norm-referenced grading system used, the scores which have more weight should affect the students' rank more. However, this principle is violated when different test scores multiply by a certain percentage or ratio. The intended weight of a score component and its actual impact on grades may not be the same. The test which has a larger standard deviation would have more weight on the composite score. This study focused on the weighting problem in the norm-referenced grading system. In this study, the effects of standard deviation difference and class level on the agreement of grades were investigated. Nine simulation conditions [3 standard deviation difference (0, 10 units, 20 units) x 3 class level (poor, average, good)] were considered. Grades were calculated in two ways: (1) T scores based on weighted raw success scores (WRSS), which we called T_{WRSS} grades; (2) T scores based on weighted standardized success scores (WSSS), which we called T_{WSSS} grades. Three agreement statistics were used in this study: simple percent agreement, extended percent agreement and kappa.

The two-way ANOVA results for all agreement coefficients showed that the standard deviation main effect was a significant effect on the agreement between

TWRSS and TWSSS grades. It could be concluded that if the class level was ignored, the standard deviation difference between midterm and final measurements affected agreement statistics (simple percent agreement, extended percent agreement, and kappa). Descriptive statistics showed that for all class levels (poor, average and good), while the maximum agreement was provided when midterm and final measurements had an equal standard deviation, the minimum agreement was provided when the standard deviation difference between midterm and final measurements were at the highest level (20 units).

As mentioned before, if we do not use a standard unit of measurement to combine scores that got from different scales, we make the Mars Climate Orbiter Miscalculation (Francis, 2006. The findings in many studies (Francis, 2006; McLachlan & Whitten, 2000; Miller, Imrie & Cox, 1998, Thorndike et al., 1991) suggest that test scores should be converted to a common score scale before aggregation. Scores from different scales were not on the same scale unless they have the same mean and standard deviation. The results of this study showed that combining test scores by only taking the simple percentage of raw scores (TWRSS grades) and standardized midterm and final scores (T_{WSSS} grades) yielded different grades. Also, the results showed that as the standard deviation difference between midterm and final measurement increased, the agreement between T_{WRSS} grades and T_{WSSS} grades decreased. For example, in the poor class, while extended agreement (or ±1 grades) between grades was about 99% in Condition 1 (where the standard deviations of midterm and final measurements were equal), it decreased to about 84% in Condition 3 (see Table 7). Another result of this study showed that there was not a significant main effect of class level on all types of agreement between T_{WRSS} and T_{WSSS} grades. It is concluded that when we ignored the standard deviation difference between midterm and final measurement, the achievement level or the mean of the class did not affect the agreement between grades. The study results also showed that the interaction between the effects of standard deviation and the class level were not statistically significant. We concluded that the effects of standard deviation were similar at all class levels.

In the literature, it was mentioned by many researchers (Francis, 2006; Linn & Gronlund, 1995; Nitko & Brookhart, 2007; Tinkelman et al., 2013) that when using norm-referenced grading, the larger standard deviation of one score component's affected more the final ranking of students when scores simply combined. The results of this study have confirmed this finding. The results showed that when final and midterm measurements had equal standard deviations, the final measurement (which has a bigger weight in total raw scores) affected students ranking more. However, when the difference between standard deviations of midterm and final measurements were increased, midterm raw scores have larger Spearman's rank-order correlations with the total score than the final measurement. But, if the midterm and final measurements were combined after standardized, at all condition final measurement had larger Spearman's rank-order correlations with the total score.

Based on the findings obtained in study, it was recommended that scores should be standardized before combined in the norm-referenced grading system. Otherwise, as study results showed, grades would not have the intended weight and the score

component, which had a large standard deviation, would affect students' ranking more. This study was limited using symmetrically distributed data; the effect of the shape of data (skewness or kurtosis) on norm-referenced grades could be investigated in the further studies.

References

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

Atılgan, H., Yurdakul, B., & Öğretmen, T. (2012). A research on the relative and absolute evaluation for determination of students achievement. *İnönü University Journal of the Faculty of Education, 13*(2), 79-98. Retrieved from https://dergipark.org.tr/en/download/article-file/92231

Başol-Göçmen, G. (2014, July). *Değerlendirmeye genel bir bakış: Kriter-referans (mutlak) ya da norm-referanslı (bağıl) değerlendirme. [A general look to evaluation: Criterion-referenced (absolute) or norm-referenced (relative) evaluation.* Paper presented at the XIII. National Educational Sciences Congress, Malatya. Retrieved from https://www.pegem.net/dosyalar/dokuman/212.pdf

Brookhart, S. M. (1999). *The art and science of classroom assessment: The missing part of pedagogy.* ASHE-ERIC Higher Education Report (Vol. 27, No: 1). Washington, DC: The George Washington University, Graduate School of Education and Human Development.

Chan, W. (2014). A better norm-referenced grading using the standard deviation criterion. *Teaching and Learning in Medicine, 26*(4), 364-365. doi: 10.1080/10401334.2014.945031

Cunningham, G. K. (2005). *Assessment in the classroom: Constructing and interpreting tests.* Bristol: Falmer Press.

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

Francis, R. W. (2006, Fall). Common errors in calculating final grade. *Thought and Action,* 15-22. Retrieved from https://eric.ed.gov/?id=EJ1070708

Gamer, M. (2015). irr: various coefficients of interrater reliability and agreement. R package version 0.84.

Goodwin, L. D., Sands, D. J., & Kozleski, E. B. (1991). Estimating interinterviewer reliability for interview schedules used in special education research. *The Journal of Special Education, 25*(1), 73-89. doi:10.1177/002246699102500105

Harish, A. (2019, March, 21). When NASA lost a spacecraft due to a metric math mistake [Web log comment]. Retrieved from https://www.simscale.com/blog/2017/12/nasa-m Mars climate-orbiter-metric/

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. doi: 10.16992/asos.12321

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

Linn, R. L. ve Gronlund, N. E. (1995). *Measurement and assessment in teaching* (7th Edition). New Jersey: Prentice Hall

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

McLachlan, J. C. & Whiten, S. C. (2000). Marks, scores and grades: Scaling an aggregating student assessment outcomes. *Medical Education, 34*(10), 788-797. doi: 10.1046/j.1365-2923.2000.00664.x

Mertler, C. A. (2007). *Interpreting standardized test scores: Strategies for data-driven instructional decision making.* California: Sage Publications, Inc.

Miller, A. H., Imrie, B. W., & Cox, K. (1998). *Student assessment in higher education: A handbook for assessing performance*. London: Routledge.

Ministry of Education. (2016). *Ministry of Education Secondary Education Regulation*. Retrieved from http://ogm.meb.gov.tr/meb_iys_dosyalar/2016_11/03111224_ooky.pdf

Moore, D. S., McCabe, G. P., & Craig, B. A. (2009). *Introduction to the practice o statistics* (7th Edition). New York: W. H. Freeman and Company.

Nartgün, Z. (2007). An investigation of criterion referenced and norm referenced evaluation practices based on same scores causing a difference on grades or not. *Ege Journal of Education, 8*(1), 19-40.

Nitko, A. J. & Brookhart, Susan M. (2007), *Educational assessment of students* (5th Edition). New Jersey, Pearson Education Inc.

Pontes, E. A. S. (2018). A brief historical overview of the gaussian curve: From Abraham De Moivre to Johann Carl Friederich Gauss. *International Journal of Engineering Science Invention, 7*(6), 28-34. Retrieved from http://www.ijesi.org/papers/Vol(7)i6/Version-5/D0706052834.pdf

Popham, W. J. (2011). *Classroom assessment: What teachers need to know* (6th Edition). Boston: Pearson Education Inc.

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. doi: 10.11114/ jets.v4i9.1571

Tinkelman, D., Venuti, E., & Schain, L. (2013). *Global Perspectives on Accounting Education, 10*, 61-80. Retrieved from https://gpae.wcu.edu/Vol10/Methods%20of%20Combining%20Test%20a d%20Assignment%20Scores.pdf

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L. ve Hagen, E. P. (1991). *Measurement and evaluation in psychology and education* (5th Edition). New York: Macmillan Publishing Company.

Thorndike, R. L. & Hagen, E. P. (1977). *Measurement and evaluation in psychology and education* (4th Edition). New York: John Wiley & Sons, Inc.

Viera, A. J. & Garrett, J. (2005). Understanding interobserver agreement: the kappa statistics. *Family Medicine, 37*(5), 360-363.

Wuertz, D., Setz, T., Chalabi, Y., Boudt, C., Chausse, P., & Miklovac, M. (2017). fGarch: Rmetrics-autoregressive conditional heteroskedastic modelling. R package version 3042.83.1.

Norm-Dayanaklı Not Verme Sisteminde Ağırlıklandırma Probleminin İncelenmesi

Atıf:

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

Özet

Problem Durumu: Not vermede en önemli hatalardan biri ağırlıklandırmada yapılmaktadır. Bu hatanın kaynağı öğrencilerin farklı test puanlarındaki değişkenliğidir. Öğrencilerin puanlarının değişkenliği bir diğer deyişle puanların standart sapması bir testten diğerine farklılık gösterdiği zaman, daha yüksek standart sapmaya sahip puan final notunu daha fazla etkilemektedir. Norm-dayanaklı not verme sistemi kullanıldığında, gruptaki öğrencilerin final sıralamasını en çok etkileyen not bileşeni en fazla ağırlığa sahip olan olmalıdır. Ancak, bu ilke not bileşeni puanı basitçe belirli bir yüzdelikle ya da oranla çarpıp toplamı alındığında ihlal edilmektedir. Bunun sebebi toplam puana göre sıralamayı toplam puanı oluşturan bileşenlerin standart sapmasının etkilemesidir (Nitko ve Brookhart, 2007).

Notlar, öğrencilerin derse yönelik motivasyonlarını arttırmada önemli bir role sahiptir (Ebel, 1979; akt. Nitko ve Brookhart, 2007). Notlar bu işlevi, öğrenci başarısının geçerli ve güvenilir yansıtıcısı olması durumunda yerine getirebilir. Literatür incelendiğinde çoğu araştırmanın (Atalmış, 2019; Atılgan, Yurdakul & Öğretmen, 2012; Başol-Göçmen, 2004; Kaya & Semerci, 2017; Lok, McNaught & Young, 2016; Nartgün, 2007; Sayın, 2016) norm dayanaklı değerlendirme ile mutlak değerlendirmeyi

karşılaştırmaya odaklandığı söylenebilir. Norm dayanaklı not verme sisteminde ağırlıklandırmada yapılan hatayı inceleyen sınırlı sayıda araştırma (Kelley & Zarembka, 1968; Tinkelman, Venuti & Schain, 2013) bulunmaktadır.

Araştırmanın Amacı: Bu araştırmanın amacı; öğrenci başarısını belirlemede birden fazla not bileşenin kullanıldığı bağıl değerlendirmede, öğrencilerin final notunu belirlerken standart puan kullanılarak ve kullanılmadan yapılan ağırlıklandırmanın harf notları üzerindeki etkisini incelemektir.

Araştırmanın Yöntemi: Bu araştırma simülasyon verileri kullanılarak yürütülmüştür. Araştırmanın amacı doğrultusunda vize ve final ölçmelerine ilişkin ortalama açısından kötü, orta ve iyi olmak üzere üç sınıf düzeyinde veri türetilmiştir. Her bir düzeyi için üç koşul dikkate alınmıştır. Her bir koşulda vize ve final ölçmelerinin ortalamaları eşit tutulmuş, vize sınavının standart sapması farklılaştırılmıştır. Birinci koşulda vize ve final sınavlarının standart sapmaları eşit, ikinci koşulda vizenin standart sapması finalden 10 birim yüksek, üçüncü koşulda ise vizenin standart sapması finalden 20 birim yüksek alınmıştır. Araştırma kapsamında toplam dokuz [3 sınıf düzeyi x 3 standart sapma farklılaşması] koşul dikkate alınmıştır. Her bir ölçme için örneklem büyüklüğü 60 ve her bir koşul için tekrar sayısı 100'dür. Araştırmada vize sınavının öğrencinin genel başarı notuna etkisi %40 ve final sınavının katkısı %60 olarak alınmıştır. Veri analizinin ilk aşamasında vizenin %40'ı ve finalin %60'ı alınarak öğrencinin ağırlıklandırılmış ham başarı puanı (AHBP), bu puana karşılık gelen T puanı (T_{AHBP}) ve T_{AHBP} puanına ait harf notu hesaplanmıştır. İkinci aşamada, vize sınavının T puanı ve final sınavının T puanları hesaplandıktan sonra ağırlıklandırılıp toplanarak öğrencinin ağırlıklandırılmış standartlaştırılmış başarı puanı (T_{ASBP}) ve bu puana karşılık gelen harf notu hesaplanmıştır. Üçüncü aşamada ise elde edilen T_{AHBP} ve T_{ASBP} harf notları arasındaki uyum basit yüzde, genişletilmiş yüzde (±1 not) ve kappa katsayısı ile hesaplanmıştır. Veri analizinin son aşamasında ise standart sapma farklılaşması ve sınıf düzeyine göre öğrencilerin T_{AHBP} ve T_{ASBP} harf notları arasındaki uyumun farklılaşıp farklılaşmadığı iki yönlü varyans analizi (ANOVA) yapılarak incelenmiştir. Araştırmada ayrıca öğrencilerin vize puanı ve final puanları ile basit yüzdelik ile çarpılarak elde edilen toplam ham puanları (THP) arasındaki ilişki; vize ve final puanları ile vize ve final puanları standartlaştırıldıktan sonra elde edilen toplam standart puanları (TSP) arasındaki ilişki Sperman sıra farkları korelasyonu ile incelenmiştir.

Bulgular: Araştırmanın iki-yönlü ANOVA bulguları, tüm uyum indeksi değerlerinde standart sapma ana etkisinin istatistiksel olarak anlamlı etkiye sahip iken sınıf düzeyi ana etkisi ile sınıf düzeyi-standart sapma etkileşiminin anlamlı etkiye sahip olmadığını göstermiştir. Betimsel istatistikler, her bir sınıf düzeyinde notlar arasında en yüksek uyumun vize ve final ölçmelerinin standart sapmalarının eşit olduğu koşulda, en düşük uyumun ise standart sapma farkının 20 birim olduğu koşulda elde edildiğini göstermiştir. Araştırma bulguları, standart sapmaların eşit olduğu koşullarda öğrencilerin final puanı ile THP puanları arasındaki ilişkinin, vize puanları ile THP arasındaki ilişkiden daha yüksek olduğunu dolayısıyla öğrenci sıralamasını ağırlığı yüksek olan final sınavının etkilediğini göstermiştir. Ancak, vize ve finalin standart sapmalarının farklılaştığı koşullarda vize sınavının THP ile arasındaki ilişki final

sınavının THP ile arasındaki ilişkiden yüksek bulunmuş ve öğrenci sıralamasını ağırlığı düşük fakat standart sapması yüksek olan vize sınavının etkilediği görülmüştür. Öğrencilerin vize ve final sınavı puanları ile TSP arasındaki ilişki incelendiğinde ise her sınıf düzeyi ve koşulda öğrencilerin final sınavı puanları ile TSP puanları arasındaki ilişkinin daha yüksek olduğunu ve öğrencilerin sıralamasını ağırlığı fazla olan final sınavının daha çok etkilediğini görülmüştür.

Araştırmanın Sonuçları ve Öneriler: Araştırmanın sonucunda standart sapmanın; ham puanların ağırlıklandırması ile hesaplanan notlar ile standartlaştırıldıktan sonra ağırlıklandırılan notlar arasındaki uyumu etkilediği, vize ve final ölçmelerine ait standart sapmalar arasındaki fark arttıkça her iki yöntemle hesaplanan notlar arasındaki uyumun azaldığı görülmüştür. Ayrıca vize ve finalin standart sapmaları eşit olduğu koşullarda, öğrencilerin sıralamasını genel notta ağırlığı daha fazla olan final notu etkiler iken puanlar standartlaştırılmadan birleştirildiğinde öğrencilerin sıralamasını ağırlığı küçük fakat standart sapması büyük olan vize sınavının etkilediği görülmüştür. Puanlar standartlaştırıldıktan sonra birleştirildiğinde ise vize ve final standart sapmaları arasındaki fark her ne olursa olsun öğrencilerin sıralamasını ağırlığı daha yüksek olan final sınavının etkilediği görülmüştür.

Araştırmanın sonuçlarına dayalı olarak, eğer norm –dayanaklı not verme sistemi kullanılacak ise farklı testlere ilişkin puanların standart puanlara çevrildikten sonra birleştirilmesi önerilmektedir. Aksi takdirde, puanların genel nottaki hedeflenen ağırlığı ile gerçekte oluşan ağırlığı farklı olabilir. Öğrencinin sıralamasını ağırlığı yüksek olan sınav değil standart sapması yüksek olan sınav etkileyecektir. Bu araştırma vize ve final puanlarının normal simetrik dağılıma uygun olarak türetilmesi ile sınırlıdır. Gelecekte yapılacak olan araştırmalarda not bileşeni puanlarının çarpık ya da basık dağılım göstermesinin notlara etkisi incelenebilir.

Anahtar Kelimeler: Standart puanlar, norm-dayanaklı not verme, uyum.