BUEFAD | 2021, Volume 10, Issue 2, 309-324
Bartın University Journal of Faculty of Education
dergipark.org.tr/buefad
DOI: 10.14686/buefad. 878504

Scoring Methods for Multiple Choice Tests: How does the Item Difficulty Weighted Scoring Change Student's Test Results?

Metin Yaşar[a], Seval Kula Kartal[b], & Eren Can Aybek[c]

a, Dr., Pamukkale University, (http://orcid.org/0000-0002-7854-1494) * myasar@pau.edu.tr

b, Dr., Pamukkale University, (http://orcid.org/0000-0002-3018-6972)

c, Dr., Pamukkale University, (http://orcid.org/0000-0003-3040-2337)

Research Article	Received: **11.02.2021**	Revised: **01.05.2021**	Accepted: **02.05.2021**

Abstract

The past studies on weighted test scoring were focused on the correlation with unweighted test scores. Yet, it is important to investigate the effect of weighted scoring on students' pass and fail rates. In the present study, it was aimed to compare students' test scores, item and test statistics calculated based on the unweighted (1 – 0) and item difficulty weighted scores (Qj – 0). The study also included a proposal for converting the weighted scores into a 100-point scale system. A teacher-made 34-item multiple-choice achievement test was conducted to a group of 431 people via learning management system. As a result of the data analysis, the McDonald's Omega internal consistency coefficients that were obtained according to the 1 – 0 and (Qj – 0) methods were obtained as .725 and .721, respectively. The Pearson's product moment correlation coefficient was .916, and the Spearman's rank-order correlation coefficient was .926 between student scores obtained according to the two methods. Furthermore, a criterion-based evaluation was made based on the two criteria (test scores of 50 and 60), and the numbers of the students who were successful and unsuccessful in the course were determined according to both scoring methods. Accordingly, it was found that more students would be considered unsuccessful in the course in the (Qj-0) scoring method; however, it was understood that this method could reveal differences among individuals more than the unweighted scoring method.

Keywords: Teacher-made test, multiple choice tests, scoring methods

Çoktan Seçmeli Testlerde Puanlama Yöntemleri: Madde Güçlüğüne Dayalı Ağırlıklandırma Öğrencilerin Test Sonuçlarını Nasıl Değiştirir?

Öz

Ağırlıklı puanlama ile ilgili geçmiş çalışmalar incelendiğinde, genellikle ağırlıklandırılmamış puanlarla olan korelasyonların incelendiği, buna karşın ağırlıklandırmanın öğrencilerin geçme – kalma oranlarına olan etkisinin araştırılmadığı görülmüştür. Bu çalışmada öğretmen yapımı çoktan seçmeli 34 maddelik bir başarı testinin 431 kişilik bir gruba öğrenme yönetim sistemi aracılığıyla uygulanmıştır. Daha sonra ağırlıklandırılmamış (1 – 0) ve madde güçlüğüne göre ağırlıklandırılmış (Qj – 0) puanlara göre madde ve test istatistiklerinin, öğrencilerin dersten geçme ve kalma durumlarının karşılaştırılması amaçlanmıştır. Aynı zamanda ağırlıklandırılmış puanların 100'lük puan sistemine çevrilmesine yönelik bir öneri de sunulmuştur. Veri analizi sonucunda 1 – 0 ve Qj – 0 yöntemlerine göre elde edilen McDonald's Omega iç tutarlık katsayıları sırasıyla .725 ve .721 olarak elde edilmiştir. İki yönteme göre elde edilen öğrenci puanları arasında ise Pearson momentler çarpım korelasyon katsayısı .916 ve Spearman sıra farkları korelasyon katsayısı .926 olarak bulunmuştur. Aynı zamanda sırasıyla 50 ve 60 puana göre ölçüt dayanaklı bir değerlendirme yapıldığında, her iki yönteme göre dersten başarılı ve başarısız sayılan öğrenci sayıları belirlenmiştir. Buna göre Qj – 0 puanlama yöntemine göre daha çok öğrencinin dersten başarısız sayılacağı bulunmuş, ancak buna karşın bu yöntemin bireyler arasındaki farklılıkları daha iyi ortaya koyabileceği anlaşılmıştır.

Anahtar kelimeler: Öğretmen yapımı test, çoktan seçmeli testler, puanlama yöntemleri

To cite this article in APA Style:

Yaşar, M., Kula Kartal, S., & Aybek, E. C. (2021). Scoring methods for multiple choice tests: How does the item difficulty weighted scoring change student's test results? *Bartın University Journal of Faculty of Education, 10*(2), 309-324. https://doi.org/10.1016/buefad.878504

1 | Introduction

Multiple-choice tests are widely used, from classroom measurement applications to national and even international applications. Additionally, multiple-choice tests may be decisive in most of the student's success in the course, especially when they are used to measure classroom learning (Mavis, Cole, & Hoppe, 2001; McDougall, 1997). Their many superior aspects such as application on large groups, being composed of many items, ensuring more reliable measurements with the increase of the number of items (Wilson & Wang, 1995), being convenient in terms of application and scoring, and being eligible to be graded objectively (DiBattisa & Kurzawa, 2011; Roediger & Marsh, 2005; Sax, 1989) have led to the widespread use of multiple-choice items. Moreover, including a large number of items in such tests also allows increasing the content validity (Bacon, 2003). There are also criticisms directed at multiple choice tests in contrast to the superior aspects of these tests. These criticisms focus on the view that multiple-choice test items are not suitable for measuring high-level thinking skills (Clark & Linn, 2003; Heubert & Hauser, 1999; Shepard, 2000; Walsh & Seldomridge, 2006). There are also researchers who believe that higher-level thinking skills may be measured with multiple-choice items (Brookhart, 2010). However, it should be considered that, as the cognitive level desired to be measured with such items increases, item writing also becomes more difficult (Buckles & Siegfried, 2006; Palmer & Dewitt, 2007).

The traditional method of scoring multiple-choice items is to score the correct answer (Bereby-Meyer, et.al., 2002; Kurz, 1999). In this scoring, students receive 1 point for their correct answer and 0 points for their incorrect or no answer (Akkuş & Baykul, 2001; Downing & Haladyna, 2006; Gözen, 2006; Kruz, 1999; Özdemir, 2003; Sax, 1989; Turgut, 1992; Yurdugül, 2010). This scoring is also known as 1 – 0 scoring, Bernoulli weighting or unweighted scoring (Rotou, et.al., 2002; Stocking, 1996). In this method, all items in the test are considered and rated at an equal weight (Haladyna, 1990).

One of the important criticisms of the traditional scoring method is that 1 – 0 scoring may provide an estimate of the ranking of the students taking the exam, not their level of knowledge (Kurz, 1999). Another criticism directed at this scoring method is that the 1 – 0 scoring method cannot increase the validity of the item (Merwin, 1959). Additionally, in a multiple-choice item, there is a possibility that an individual will answer correctly to the item by chance, even if they do not have the qualification measured by the item. However, in this method, it is considered that the person who gets 1 point by answering the item correctly has the qualification measured by the item completely, and the person who gets 0 by answering incorrectly or does not answer at all does not have the qualification measured by the item at all. Therefore, success by chance is not taken into account while interpreting scores obtained based on the 1 – 0 scoring method (Budescu & Bar-Hillel, 1993; Frary, 1988; Kubinger, et.al., 2010).

When multiple-choice test items are scored as 1 – 0, the answers of individuals who have the qualification measured by the item fully or those who answer the item correctly by chance are classified as correct, while all other answers are classified as incorrect (Akkuş & Baykul, 2001; Gözen, 2006; Jaradat and Tollefson, 1988; Cruise, 1999; Özdemir, 2003; Sax, 1989; Yurdugül, 2010). Additionally, those who have the qualification measured in a multiple-choice item fully or partly, those who answer the item correctly, and those who answer correctly by chance receive the same item score. Similarly, it is possible to answer an item incorrectly while partially having the measured qualification or to answer the item incorrectly due to carelessness while completely having the qualification. Therefore, there are some points where the 1 – 0 scoring method is insufficient to determine the difference between an individual who has the qualification required by the item and another individual who does not.

Different scoring methods have been developed for multiple-choice items by researchers who consider the limitations of the 1 – 0 scoring method. One of these researchers, Cooms (1953) proposed the method of elimination scoring. In this method, the options are weighted, and individuals eliminate the options that they think are wrong. The scores that an individual can get from an item vary in the range of [-(n-1), (n-1)] to indicate the number of options for the items by n. In other words, for an item with four options, the

option with the falsest information is rated as -3, and the option with the most accurate information is rated as +3. Other options are also weighted by a value in this range, according to the accuracy of the information they contain.

Following Coombs (1953), different scoring methods have also been developed. Frary (1989) considered scoring methods in two basic classes: the direct response methods and examinee judgements methods. In the direct answering method, individuals select and mark which of the options they think is the correct answer. In this method, scores based on the answer, option weighting, multiple answers, and the item response theory may be applied until the correct answer is found. In the answerer decisions method, individuals mark the options or groups of options that they think are either wrong or right among all options (Özdemir, 2003). In this method, different ways of scoring may be followed, such as scoring based on the degree of trust, dividing into subsets and scoring based on the probability of answering (Akkuş & Baykul, 2001).

Another recommendation for scoring multiple-choice test items is to weigh the items based on an objective measure. Here, psychometric properties such as the difficulty and discrimination of the item may be used as the criteria. Items with high difficulty and discrimination are given more points, while items with low difficulty and discrimination are given fewer points (Budescu, 1979). In particular, the findings of the research conducted by Gözen (2006) drew attention while examining the research on weighting based on item difficulty. In their study, the 1 – 0 scoring method was compared to the 1 – P_j and [(1 – P_j)r_{jX}] scoring methods. Comparisons were made for both short-answer items and multiple-choice items. Accordingly, significant relationships on the level of .91 were obtained between the 1 – 0 and 1 – P_j scoring methods and on the level of .92 between the 1 – 0 and [(1 – P_j)r_{jX}] scoring methods. The study by Yurdugül (2010) also compared the 1 – 0 scoring method to the method of scoring by weighing with r_jX and methods based on the Item Response Theory (IRT). Accordingly, a significant relationship on the level of .99 was found between the 1 – 0 and r_{jx} – 0 scoring methods. While the study by Gözen (2006) was conducted on 316 students, the study by Yurdugül (2010) was conducted on a group of 10000 people selected from students who participated in a national exam.

Considering the weighting methods used by Gözen (2006) and the size of the sample, this study is also similar to the study pattern of Gözen (2006). However, the aforementioned studies have focused on the relationship between weighted and unweighted scores and the effect of scoring methods on the psychometric qualities of the test. These studies did not focus on the question of how students will be affected by decisions that will be made according to absolute criteria while using different scoring methods. This study differs from related studies in that it focuses on the changes that the scoring method creates on the individual level. Additionally, the related studies did not include discussions on how weighted scores can be converted to a hundred-point system. Theoretically, the benefit of such a transformation is open to debate. However, in practice, there are situations that require the use of a hundred-point system. Therefore, in cases where weighted scoring is utilized, there is a need for research that will guide implementers to convert the total score into a hundred-point system. For this reason, when the unweighted (1 – 0) and weighted scoring methods according to item difficulty were used in the study, it was aimed to compare the psychometric characteristics of the test and examine the relationship between the success scores of individuals according to both methods. Additionally, when the criteria to be taken to succeed in the course were 50 and 60 points, it was examined how the decisions made about students changed according to both scoring methods.

2 | METHOD

RESEARCH GROUP

The research group consisted of 431 students who were enrolled in the assessment and evaluation class in the summer term of 2019-2020 at Pamukkale University, School of Education, but received distance education due to the COVID-19 pandemic and took the mid-term exam. Among these students, 126 (29.2%) were male, and 305 (70.8%) were female.

DATA COLLECTION

A 34-item multiple-choice achievement test was used to collect the data. The achievement test was developed to measure the academic achievement of university students in the assessment and evaluation class in education and applied to the students as a mid-term exam. Therefore, the scope of the test included the basic concepts of measurement and evaluation, error in measurement, reliability, validity and usefulness. To determine the internal consistency of the test, the McDonald's Omega coefficient was calculated. According to the scoring method, the 1 – 0 reliability coefficient of the test was .725, which was found to be. 721 according to the scoring method weighted based on item difficulty. The results section covers more detailed discussion on the psychometric properties of the test.

The data collection process was carried out through the Moodle Learning Management System (LMS) due to the COVID-19 pandemic. The students were given 45 minutes to complete the achievement test. In order to avoid problems that may have arisen from the system while answering the test items, the students were given the right to re-enter to the system. Additionally, the items included in the test were presented to the students in groups of five. After the students answered the five items presented to them, the other five test items appeared on the screen. During the application phase, the students were able to access the item they wanted to review again within the time given to them and check their responses to the test items.

DATA ANALYSIS

Within the scope of the study, the data collected from the sample was analyzed using the ShinyItemAnalysis 1.3.4 (Martinková and Drabinová, 2016) package on R 4.0.2 (R Core Team, 2020). ShinyItemAnalysis is a software which can calculate items and test statistics. Using the ShinyItemAnalysis package, item difficulty as item statistics, the difference between the item difficulty levels for upper and lower 27% of group (gULI), item reliability (Rel), item reliability based on the item-remainder correlation (rel-drop) item-total correlation (RIT), and the item-rest correlation (RIR) statistics were calculated. Test statistics were obtained by using the Microsoft Office 365 Excel software.

Primarily, an item-score matrix was created to calculate the item-difficulty, gULI, Rel-drop, RIR and RIT statistics based on the scoring methods based on 1 – 0 and item-difficulty (the $Q_j - 0$ expression will be used later on in the report for ease of display) weighting as shown in Table 1. According to the $Q_j - 0$ scoring method, the Q Matrix given in Table 2 was created. The item-score matrix in Table 1 is two-dimensional, where the rows show answerers, and the columns show items. In the case where the item-score matrix is created with a score of 1 – 0, the row totals show the total number of correct answers or the total score ($\sum X_c$) of the answerer. The column totals ($\sum X_j$) show how many examinees answered each test item correctly.

While the equation $P_j = \frac{\sum X_j}{n}$ was used in the calculation of the difficulty level (P_j), the equation $Q_j = 1 - P_j$ was used for the wrong answering rates (Q_j) of the test items. In this study, the focus was not on the options; the weighted item statistics were calculated by considering the difficulty levels of the test items (P_j). In weighting the items, the $Q_j = 1 - P_j$ ratios of those who answered incorrectly to the relevant item were accepted as the weight ratio of the item. In the 1 – 0 scoring method, the P_j value represented

the difficulty level of the item. If this value got close to 1, the difficulty of the items would decrease; that is, the items would get easier.

According to the $Q_j - 0$ method, as the difficulty level of the item $(1 - P_j)$ gets closer to 1, the item gets more difficult. The item difficulty levels of the $Q_j - 0$ scoring method are shown in Table 2 as the matrix Q. In the Matrix Q, the rows show answerers, resulting in total scores weighted according to the $Q_j - 0$ scoring method when the row sums are added.

The conversion of the correct number of responses in the test to a score of 100 was found to be based on equation 1 compared to the 1 – 0 scoring.

$$TS_{1-0} = \frac{100}{K} d_S$$

In this equation, 100 indicates the scoring unit, K indicates the number of items in the test, and d_S indicates the number of the correct answers of the answerer. In order to calculate the total score that answerers received from the test according to the $Q_j - 0$ scoring method, the total score $\sum Q$ in the sense of raw score according to the scoring method $Q_j - 0$ was calculated using equation 2.

$$\sum Q = Q_1 + Q_2 + Q_3 + \cdots + Q_n \quad (2)$$

The standard deviation value for the answerers' Q Scores was calculated based on equation 3.

$$S_Q = \sqrt{\frac{\sum Q^2 - \frac{(\sum Q)^2}{n}}{n-1}} \quad (3)$$

In the next step, the T standard score with a mean value of 50 and a standard deviation of 10 for each answerer was calculated using values obtained from equation 2 and equation 3 by using equation 4.

$$T_Q = 50 + 10\left(\frac{Q - \bar{Q}}{S_Q}\right) \quad (4)$$

Using the calculated T standard scores, each answerer's score was converted to a distribution with the lowest value of 10 and the highest value of 100. Equation 5 was used for this conversion.

$$TS_Q = 10 + \frac{90*(T_Q - smallest\ T_Q)}{(The\ biggest\ T_Q - smallest\ T_Q)} \quad (5)$$

Table 1. Item-Score Matrix

Examinees	TEST ITEMS																																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	$\sum X_c$
1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	0	0	0	0	1	0	0	0	1	1	0	0	1	1	1	1	0	22
2	1	0	1	0	0	0	1	1	0	1	1	1	1	1	1	1	0	0	0	0	1	1	0	0	1	0	1	0	0	1	1	0	1	0	18
3	1	0	1	1	1	0	1	1	1	1	1	1	1	0	1	0	1	1	0	0	1	1	0	0	0	1	1	0	0	1	1	0	1	0	21
4	1	0	0	1	1	1	1	1	1	0	0	1	1	0	0	0	1	1	0	0	0	0	1	0	0	1	1	1	0	0	1	0	1	0	17
5	1	1	1	1	1	1	1	0	1	1	0	0	1	1	0	0	1	0	0	1	1	0	0	1	1	1	1	1	0	1	1	1	0	0	22
6	1	1	0	1	1	0	1	1	1	0	1	1	1	0	1	1	1	1	0	0	1	1	0	0	0	1	1	0	0	0	1	0	1	0	20
7	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	0	0	1	1	0	0	0	1	0	0	0	0	1	1	1	0	22
8	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	1	1	0	1	0	1	1	1	0	0	1	1	0	0	0	1	1	1	0	23
9	1	1	0	1	1	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	1	0	0	11
10	1	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	1	0	0	8
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
422	1	1	1	1	1	0	1	0	1	0	1	0	1	0	0	1	1	0	0	1	1	1	0	0	1	0	0	0	0	1	1	0	0	0	17
423	1	1	1	1	1	0	1	1	0	1	1	1	1	0	1	0	1	1	1	0	1	0	0	0	0	1	1	0	0	1	1	0	1	0	21
424	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	0	0	1	1	0	0	0	1	1	0	0	1	1	1	1	0	23
425	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	1	0	0	1	1	0	0	0	1	0	0	1	0	0	1	0	0	1	10
426	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	0	1	0	0	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	27
427	1	1	1	1	1	0	1	1	0	1	1	1	1	0	1	0	1	0	0	0	1	1	0	0	0	1	1	0	0	1	1	0	0	0	19
428	1	1	0	1	0	1	1	1	1	1	0	1	1	1	0	1	0	0	1	0	1	1	0	0	1	0	1	1	0	1	1	1	0	0	21
429	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	1	1	0	0	1	1	0	0	0	1	1	0	0	1	1	1	0	0	22
430	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	0	0	1	1	1	0	0	1	1	0	0	1	1	1	1	0	24
431	1	1	1	1	1	1	1	1	1	0	0	1	1	1	0	0	0	0	0	0	1	0	1	0	0	1	1	1	0	0	1	1	1	1	21
$\sum X_j$	312	368	353	350	312	294	401	370	281	235	217	327	390	193	295	253	355	121	138	76	302	276	76	69	149	293	274	156	44	305	408	343	348	65	
P_j	,724	,854	,819	,812	,724	,682	,930	,858	,652	,545	,503	,759	,905	,448	,684	,587	,824	,281	,320	,176	,701	,640	,176	,160	,346	,680	,636	,362	,102	,708	,947	,796	,807	,151	

Table 2. Q Matrix

Examinees	TEST ITEMS																						
	1	2	3	4	5	6	7	8	9	10	. . .	25	26	27	28	29	30	31	32	33	34	$\sum x_c$	$\sum qw$
1	.276	.146	.181	.188	.276	.318	.070	.142	.348	.000	. . .	.000	.320	.364	.000	.000	.292	053	.204	.193	.000	22	6.1625
2	.276	.000	.181	.188	.000	.000	.070	.142	.000	.455	. . .	.654	.000	.364	.000	.000	.292	053	.000	.193	.000	18	6.0906
3	.276	.000	.181	.188	.276	.000	.070	.142	.348	.455	. . .	.000	.320	.364	.000	.000	.292	.053	.000	.193	.000	21	6.4988
4	.276	.000	.000	.188	.276	.318	.070	.142	.348	.000	. . .	.000	.320	.364	.638	.000	.000	.053	.000	.193	.000	17	5.2414
5	.276	.146	.181	.188	.276	.318	.070	.000	.348	.455	. . .	.654	.320	.364	.638	.000	.292	.053	.204	.000	.000	22	7.5709
6	.276	.146	.000	.188	.276	.000	.070	.142	.348	.000	. . .	.000	.320	.364	.000	.000	.000	.053	.000	.193	.000	20	6.1300
7	.276	.146	.181	.188	.276	.318	.070	.142	.348	.000	. . .	.000	.320	.000	.000	.000	.000	.053	.204	.193	.000	22	6.4688
8	.276	.146	.181	.188	.276	.318	.070	.142	.348	.455	. . .	.000	.320	.364	.000	.000	.000	.053	.204	.193	.000	23	7.5710
9	.276	.146	.000	.188	.276	.000	.070	.000	.348	.000	. . .	.654	.000	.364	.000	.000	.000	.000	.204	.000	.000	11	3.7054
.	.	.	.	.	.	.	.	.	.	.	. . .	.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	. . .	.	.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	.	. . .	.	.	.	.	.	.	.	.	.	.	.	
422	.276	.146	.181	.188	.276	.000	.070	.000	.348	.000	. . .	.654	.000	.000	.000	.000	.292	.053	.000	0	0	17	5.1485
423	.276	.146	.181	.188	.276	.000	.070	.142	.000	.455	. . .	.000	.320	.364	.000	.000	.292	.053	.000	.193	0	21	6.6172
424	.276	.146	.181	.188	.276	.318	.070	.142	.348	.455	. . .	.000	.320	.364	.000	.000	.292	.053	.204	.193	0	23	6.4478
425	.000	.000	.000	.000	.276	.000	.070	.142	.000	.000	. . .	.654	.000	.000	.638	.000	.000	.053	.000	.000	.849	10	5.1880
426	.276	.146	.181	.188	.276	.318	.070	.142	.348	.000	. . .	.654	.320	.364	.638	.897	.292	.053	.204	.193	.849	27	9.2181
427	.276	.146	.181	.188	.276	.000	.070	.142	.000	.455	. . .	.000	.320	.364	.000	.000	.292	.053	.000	.000	.000	19	5.3851
428	.276	.146	.000	.188	.000	.318	.070	.142	.348	.454	. . .	.654	.000	.364	.638	.000	.292	.053	.204	.000	.000	21	6.7890
429	.276	.146	.181	.188	.276	.318	.070	.142	.348	.455	. . .	.000	.320	.364	.000	.000	.292	.053	.204	.000	.000	22	6.7332
430	.276	.146	.181	.188	.276	.318	.070	.142	.348	.455	. . .	.000	.320	.364	.000	.000	.292	.053	.204	.193	.000	24	7.6753
431	.276	.146	.181	.188	.276	.318	.070	.142	.348	.000	. . .	.000	.320	.364	.638	.000	.000	.053	.204	.193	.849	21	6.5779

Research ethics

The principals of research ethics were followed by the authors in the planning, data collection, data analysis, and reporting the findings phases of the current research.

3 | Findings

In order to examine whether there was difference in the item statistics calculated based on the 1 – 0 and $Q_j - 0$ scoring methods from a teacher-made test, item statistics were calculated, and these statistics are presented in Table 3.

When the item statistics given in Table 3 are examined, it is seen that there was significant differentiation between the items involved in the test in terms of their difficulty levels. When the lower – upper 27% groups method, the generalized lower – upper 27% groups method, the total matter and remaining matter correlations were examined, there was no significant difference between the item distinctiveness indices according to the 1 – 0 and $Q_j - 0$ methods.

Table 3. Item statistics obtained by 1 – 0 and $\mathbf{Q_j - 0}$ scoring methods

	Diff		SD	ULI		gULI		RIT		RIR		Omega Drop		Rel		Rel Drop	
	1-0	Q-0	1-0	1-0	Q-0	1-0	Q-0	1-0	Q-0	1-0	Q-0	1-0	Q-0	1-0	Q-0	1-0	Q-0
M1	.724	.276	.448	.322	.301	.242	.242	.341	.309	.249 [.158,.336]	.238 [.147, .325]	.725	.721	.153	.038	.111	.029
M2	.854	.146	.354	.217	.182	.205	.149	.295	.220	.221 [.129, .309]	.189 [.096, .278]	.727	.723	.104	.011	.078	.010
M3	.819	.181	.385	.294	.238	.228	.200	.362	.285	.285 [.196, .370]	.245 [.154, .332]	.723	.720	.140	.020	.110	.017
M4	.812	.188	.391	.273	.224	.195	.149	.357	.217	.278 [.188, .363]	.173 [.080, .263]	.723	.720	.139	.016	.109	.013
M5	.724	.276	.448	.399	.273	.260	.186	.371	.281	.281 [.192, .366]	.209 [.117, .298]	.723	.718	.166	.035	.126	.026
M6	.682	.318	.466	.378	.364	.298	.242	.389	.333	.296 [.207, .380]	.248 [.157, .335]	.722	.718	.181	.049	.138	.037
M7	.930	.070	.255	.168	.003	.107	.002	.359	.055	.308 [.220, .391]	.025 [-.070, .119]	.721	.717	.091	.003	.078	.001
M8	.858	.142	.349	.252	.196	.195	.140	.335	.244	.264 [.174, .350]	.215 [.123, .303]	.724	.720	.117	.012	.092	.011
M9	.652	.348	.477	.427	.315	.321	.228	.391	.290	.296 [.207, .380]	.193 [.100, .282]	.722	.719	.186	.048	.141	.032
M10	.545	.455	.499	.490	.503	.367	.330	.435	.396	.339 [.258, .420]	.269 [.179, .354]	.719	.716	.217	.090	.169	.061
M11	.503	.497	.501	.378	.259	.265	.228	.314	.238	.209 [.117, .298]	.087 [-.008, .180]	.726	.721	.157	.059	.104	.022
M12	.759	.241	.428	.385	.273	.237	.172	.397	.272	.312 [.224, .395]	.211 [.119, .299]	.721	.717	.170	.028	.134	.022
M13	.905	.095	.294	.189	.147	.112	.112	.349	.242	.290 [.201, .374]	.226 [.134, .314]	.722	.719	.102	.007	.085	.006
M14	.448	.552	.498	.112	.238	.107	.181	.108	.231	-.002 [-.096, .092]	.063 [-.032, .157]	.738	.734	.054	.063	-.001	.017
M15	.684	.316	.465	.490	.371	.330	.274	.464	.371	.377 [.293, .455]	.290 [.201, .374]	.717	.713	.216	.054	.175	.042
M16	.587	.413	.493	.252	.245	.172	.163	.265	.256	.160 [.067, .251]	.134 [.040, .226]	.729	.726	.131	.052	.079	.027
M17	.824	.176	.382	.378	.301	.256	.251	.503	.376	.436 [.356, .509]	.340 [.254, .421]	.713	.709	.192	.025	.166	.023
M18	.281	.719	.450	.182	.252	.126	.209	.189	.234	.091 [-.003, .184]	.036 [-.059, .130]	.733	.729	.085	.076	.041	.012
M19	.320	.680	.467	.336	.420	.242	.307	.291	.420	.192 [.099, .281]	.241 [.150, .328]	.728	.725	.136	.133	.089	.076
M20	.176	.824	.382	.098	.203	.065	.130	.138	.300	.054 [-.041, .148]	.111 [.017, .203]	.736	.732	.053	.094	.021	.035
M21	.701	.299	.458	.378	.434	.288	.307	.400	.399	.309 [.221, .392]	.324 [.237, .406]	.721	.716	.183	.055	.141	.044
M22	.640	.360	.480	.503	.399	.391	.298	.477	.379	.388 [.305, .465]	.283 [.194, .368]	.716	.712	.229	.065	.186	.049
M23	.176	.824	.382	.154	.266	.121	.186	.205	.321	.122 [.028, .214]	.134 [.040, .226]	.732	.729	.078	.101	.046	.042
M24	.160	.840	.367	-.028	.112	-.014	.070	.000	.158	-.081 [-.174, .014]	-.033 [-.127, .062]	.742	.739	.000	.049	-.030	-.010
M25	.346	.654	.476	.182	.301	.135	.200	.165	.277	.060 [-.035, .154]	.088 [.-.007, .181]	.735	.731	.079	.086	.029	.027
M26	.680	.320	.467	.552	.148	.386	.115	.537	.445	.457 [.379, .529]	.364 [.279, .443]	.711	.708	.250	.069	.213	.056
M27	.636	.364	.482	.580	.503	.474	.363	.530	.462	.446 [.367, .519]	.371 [.287, .450]	.712	.708	.255	.081	.215	.065
M28	.362	.638	.481	.210	.056	.158	.037	.194	.148	.088 [-.007, .181]	.092 [-.002, .185]	.733	.731	.093	.013	.042	.008
M29	.102	.898	.303	-.056	.028	-.019	.019	-.092	.016	-.158 [-.249,-.065]	-.150 [-.241,-.056]	.745	.742	-.028	.004	-.048	-.041
M30	.708	.292	.455	.273	.161	.219	.172	.249	.173	.151 [.057, .242]	.092 [-.002, .185]	.730	.726	.113	.023	.069	.012
M31	.947	.053	.225	.119	.091	.102	.084	.284	.188	.238 [.147, .325]	.181 [.088, .271]	.725	.721	.064	.002	.053	.002
M32	.796	.204	.404	.315	.301	.237	.219	.336	.305	.253 [.162, .339]	.258 [.168, .344]	.724	.720	.135	.025	.102	.021
M33	.807	.193	.395	.343	.266	.242	.214	.417	.309	.341 [.255, .422]	.266 [.176, .352]	.719	.714	.165	.024	.134	.020
M34	.151	.849	.358	.217	.259	.153	.209	.256	.365	.180 [.087, .270]	.187 [.094, .277]	.729	.724	.092	.111	.064	.057

Diff: Difficulty value obtained by dividing the mean score by the range. **gULI:** The difference between the item difficulties of the lower and upper groups. **Rel:** Item reliability, **Rel drop:** Item reliability when the item is removed **ULI:** Item differentiation according to the lower-upper groups method **RIT:** Item-total correlation **RIR:** Item-rest correlation

Fisher's Z test was intended to be used to test the significance of the difference between the item distinctiveness indices obtained by the two methods, but this test was not possible because the correlation coefficient between the item scores according to the 1 – 0 method and the item scores according to the $Q_j - 0$ method was 1.00. Instead, 95% confidence intervals were calculated for the indices, and it was decided that the confidence intervals covered each other, so, there was no significant difference between the indices. Although it seems that the distinctiveness of the items in relation to some other items was quite low, interpretation of the distinguishing indices of the items was not made as the aim of the study did not include it. The test statistics obtained according to the unweighted and weighted scoring methods are given in Table 4.

Table 4. Individual Scores and Test Statistics Obtained by the 1 – 0 and $Q_j - 0$ Scoring Methods

	Test Statistics According to 1 - 0 Scoring Method				$Q_j - 0$ Test Statistics According to the Scoring Method			
Examiners	$\sum X_C$	Z_{1-0} Standard Score	T_{1-0} Standard Score	TS_{1-0}	$\sum Q$	Z_{Q_j-0} Standard Score	T_{Q_j-0} Standard Score	TS_{Q_j-0}
1	22.00	.3798	53.7983	64.7059	6.1625	-.1819	48.18092	49.6632
2	18.00	-.5098	44.9018	52.9412	6.0906	-.2262	47.73756	48.9580
3	21.00	.1574	51.5742	61.7647	6.4988	.0255	5.25467	52.9615
4	17.00	-.7322	42.6777	5.0000	5.2414	-.7499	42.50108	4.6293
5	22.00	.3798	53.7983	64.7059	7.5709	.6866	56.86563	63.4763
6	2.00	-.0650	49.3501	58.8235	6.1300	-.2019	47.98051	49.3444
7	22.00	.3798	53.7983	64.7059	6.4688	.0070	5.06968	52.6672
8	23.00	.6022	56.0224	67.6471	7.5710	.6866	56.86625	63.4773
9	11.00	-2.0667	29.3330	32.3529	3.7054	-1.6970	33.02954	25.5648
10	8.00	-2.7339	22.6606	23.5294	3.3342	-1.9259	3.74058	21.9242
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
422	17.00	-.7322	42.6777	5.0000	5.1485	-.8072	41.9282	33.0202
423	21.00	.1574	51.5742	61.7647	6.6172	.0985	5.9848	49.0252
424	23.00	.6022	56.0224	67.6471	6.4478	-.0060	49.9402	47.1792
425	1.00	-2.2891	27.1089	29.4118	5.1880	-.7828	42.1718	33.4507
426	27.00	1.4919	64.9189	79.4118	9.2181	1,7023	67.0229	77.3683
427	19.00	-.2874	47.1259	55.8824	5.3851	-.6613	43.3872	35.5985
428	21.00	.1574	51.5742	61.7647	6.7890	.2044	52.0442	5.8974
429	22.00	.3798	53.7983	64.7059	6.7332	.1700	51.7001	5.2893
430	24.00	.8247	58.2465	7.5882	7.6753	.7509	57.5094	6.5558
431	21.00	.1574	51.5742	61.7647	6.5779	.0742	5.7424	48.5970
				$\overline{X}$ = 59.683				$\overline{X}$: 52.556
				S_D = 13.224				S_D: 15.923
				McDonald's Omega: .725				McDonald's Omega: .721

$\sum X_C$: total number of correct answers according to 1 – 0 scoring method; Z_{1-0}: Z standard score according to 1 – 0 scoring method; T_{1-0}: T standard score according to 1 – 0 scoring method; TS: Total score according to 1 – 0 scoring method; $\sum Q$: total value of Matrix Q weighted according to $Q_j - 0$ scoring method; Z_{Q_j-0}: Z standard score according to $Q_j - 0$ scoring method; T_{Q_j-0}: T standard score according to $Q_j - 0$ scoring method; TS_{Qj-0}: Total score according to $Q_j - 0$ scoring method

When the values given in Table 4 are examined, it may be seen that the mean test scores based on the 1 – 0 and $Q_j - 0$ scoring methods were 59.7 and 52.5, respectively. One-sample t-test was performed to determine the significance of the difference between scores obtained and to determine whether the difference between the scores obtained by the two methods significantly differed from zero. Accordingly, it was found that there was a significant difference between the mean scores obtained according to the two methods (t_{430} = 22.837; p<.05). The mean score obtained by the students according to the 1 – 0 scoring method was significantly higher than the mean score they obtained according to the $Q_j - 0$ scoring

method. The McDonald's Omega internal consistency coefficients, calculated according to two different scoring methods for the test, were found to be close to each other.

After the one-sample t-test conducted to examine whether the students' achievement scores differed according to the scoring methods, the scores calculated according to the 100-point scale in Table 4 were also examined individually. This examination revealed that the scores that the students received according to the 100-point scale system differed according to the scoring methods. For example, the first student in the first place received a score of 64.71 in the 1- 0 scoring method and 49.66 in the $Q_j - 0$ scoring method. Additionally, while the total scores of the students whose correct answer numbers were equal in the 1- 0 scoring method were also equal, this did not apply to the $Q_j - 0$ method because weighting was performed according to the item difficulty. For example, in Table 4, students in the first, fifth and seventh places responded correctly to 22 items, and all three scored 64.7 points according to the 1 – 0 scoring method. However, according to the $Q_j - 0$ scoring method, the scores of these students were calculated as 49.7, 63.7 and 52.7, respectively. This result showed that the test scores of the students with the same number of correct answers differed, as the differences between item difficulty levels were taken into account in the $Q_j - 0$ scoring method. In this case, it may be stated that the first student responded correctly to easier items than the fifth- and seventh students, and the fifth student responded correctly to more difficult items than the first and seventh students. As a result, although the correct answer numbers were the same, the first student received a lower score than the fifth student.

Another situation taken into account in this study was whether there was a statistically significant relationship between the scores obtained according to the 1 – 0 and $Q_j - 0$ scoring methods. Therefore, both the Pearson's product moment correlation and Spearman's rho correlation coefficients were calculated. The degree of the relationship between the success scores from both scoring methods were found to be 0.916 according to the Pearson's product moment correlation coefficient and 0.926 according to the Spearman's rho correlation coefficient. It is seen that there was a very high and statistically significant relationship between the achievement scores obtained for both methods in the positive direction.

The correlation between the scores obtained according to the two methods was very high, which means that a student who was successful according to one type of scoring was also successful according to the other method. However, when it is decided whether a student is successful – unsuccessful in a course, it is also necessary to determine how weighted and unweighted scoring methods change the outcome of the student's evaluations. Accordingly, when the criteria for being considered successful in the course were 50 and 60 points, respectively, the numbers of students who would be considered successful and unsuccessful in the course according to the scoring methods were calculated and are given in Table 5.

Table 5. *Distribution of Students who are Considered Successful and Unsuccessful according to the 1 – 0 and* $Q_j - 0$ *Methods*

	Criteria	$Q_j - 0$		Criteria	$Q_j - 0$	
	50	Successful	Unsuccessful	60	Successful	Unsuccessful
$1 - 0$	Successful	250	91	Successful	139	101
	Unsuccessful	2	88	Unsuccessful	2	189

When the values given in Table 5 are examined, it is revealed that the numbers of the students who succeeded and failed in the course differed according to the scoring method. Compared to the $Q_j - 0$

method for the values of both criteria, it seems that the number of students who were considered successful in the course was greater in the 1 – 0 scoring method. This result suggested that the scoring methods also differentiated the decisions made about students. As mentioned earlier, the differences between the items in terms of difficulty levels were not taken into account while calculating the student scores in the 1 – 0 scoring method. In contrast, the students' scores were calculated by taking into account the fact that the items differed in terms of difficulty levels in the $Q_j - 0$ scoring method. Therefore, the success rates of the students in the test varied according to the scoring method depending on the different weight scores of the test items with different difficulty levels. A scatter plot was created to better understand how this situation changed the decisions made about the students. Figure 1 shows the scatter plot obtained after converting the students' scores received according to both the 1 – 0 and $Q_j - 0$ scoring methods into a 100-point scale system.

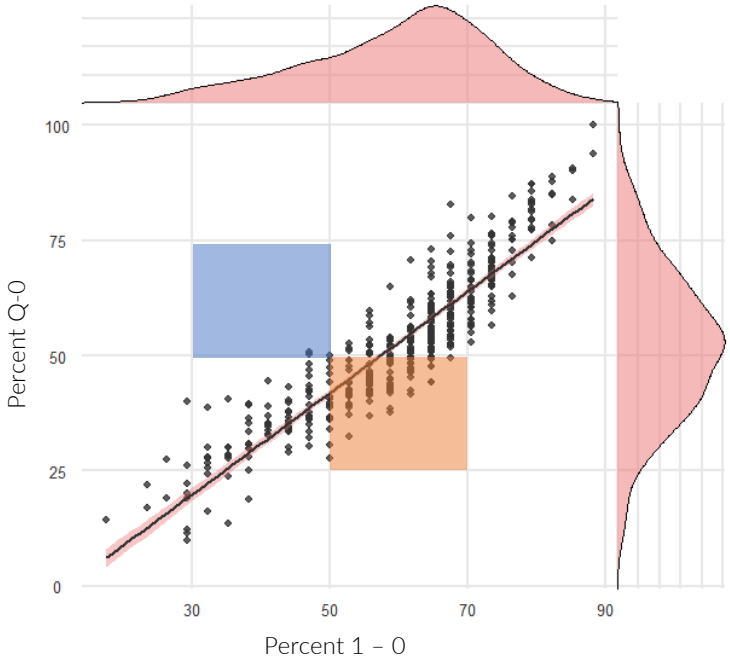

Figure 1. *Scatter plot of scores obtained according to methods 1 – 0 and Q_j - 0*

Looking at Figure 1, the difference between the areas marked with orange and blue was quite remarkable. The Orange area shows the students scoring above 50 according to the 1 – 0 scoring method but below 50 according to the $Q_j - 0$ method. The Blue Area shows those who scored below 50 according to the 1 – 0 method but above 50 according to the $Q_j - 0$ method. Although it seems that there was a high-level relationship between the two scores, it seems that while evaluating students according to absolute criteria, decisions made about students will differ based on the scoring methods. About 27% of the students who should have been considered successful according to the 1 – 0 scoring method when the criterion value was 50 points in the assessment and about 42% who should have been considered successful when the criterion value was 60 points turned out to be unsuccessful in the course according to the $Q_j - 0$ method. Additionally, when the density plots in the upper right part of Figure 1 are examined, it is understood that the scores showed a normal distribution according to the $Q_j - 0$ method, while the scores showed a skewed distribution to the left in the 1 – 0 method. In this context, it was found that the $Q_j - 0$ method could better reveal differences between individuals, given the standard deviation values of the distributions also included in Table 4.

4 | Discussion & Conclusion

In this study, it was aimed to compare the item and test statistics calculated based on the measurement results obtained according to the 1 – 0 and $Q_j - 0$ scoring methods, the test scores of the students and the decisions made about them. For this purpose, a teacher-made achievement test containing 34 multiple choice items created within the scope of the Assessment and Evaluation in Education course, which is a mandatory course in schools of education in Turkey, was applied to the study group. Based on the 1 – 0 and $Q_j - 0$ scoring methods, the students' item and test scores were calculated, and the item and test statistics obtained on the basis of these scores were compared. When the results of the item statistics were examined, it was found that there was significant difference between the items involved in the test in terms of their difficulty levels, and it was understood that the items in the test were not of equal difficulty.

According to the two scoring methods, there was a fairly high level of significant relationship between the test scores obtained. This result coincided with the research by Gözen (2006) and Yurdugül (2010), who found fairly high correlation coefficients between weighted and unweighted scores. Additionally, the study found that the McDonald's Omega reliability coefficients obtained when the reliability values were calculated based on both scoring methods were very close to each other. Supporting this finding, the study by Akkuş and Baykul (2001) stated that using different item scoring methods often does not change reliability or even increase it very little. The results of the study conducted by Yurdugül (2010) also supported the conclusion of this study on the reliability coefficient. It was found that the reliability coefficients estimated by the researcher based on the total scores obtained as a result of weighted scoring according to the item distinctiveness values and as a result of 1 – 0 scoring were very close to each other.

The results of the study on the relationship and reliability coefficients showed that there was no significant difference between the results obtained by the two scoring methods. As stated in the introduction, this study focused more on the impact of scoring methods on individuals' scores and decisions about them than on the psychometric qualities of the test. Therefore, the study sought to further examine the test scores of the individuals and the changes that occurred in the "passed – failed" decisions made about them. These reviews showed that the test scores obtained with unweighted scoring and weighted scoring based on item difficulty, the test mean and standard deviation values, and the "passed – failed" decisions made about the students differed. Budescu (1979) stated that assigning different weights to test items does not significantly affect test properties and performance. However, the results of this study revealed that when the $Q_j - 0$ scoring method was used instead of 1 – 0 scoring, there was a difference in favor of the 1 – 0 scoring method both in the test statistics and in the decisions made about the students.

In the $Q_j - 0$ scoring method, students who had the same number of correct answers on the test had different overall scores due to the different difficulty levels of the items they correctly respond to. As a result, the mean and standard deviation values of the test also changed. When the $Q_j - 0$ scoring method was used, the arithmetic mean value of the test decreased, and the standard deviation value increased. The increase in the standard deviation value indicated that weighting based on item difficulty increased the differences in the scores between individuals. This result of the study demonstrated that the $Q_j - 0$ scoring method may contribute to revealing the difference among individuals in terms of the traits are measured by the test. Similarly, Akkuş and Baykul (2001) addressed points that make it difficult to use weighted scoring methods in practice, but they emphasized that weighted scoring is useful in terms of its power to provide information about the individual.

When an absolute criterion-based assessment was performed to examine how the passed-failed decisions made about the student changed based on the scoring method, it was observed that the number

of students who succeeded or failed in the course changed according to the scoring method. About 27% of the students who should have been considered successful according to the 1 – 0 scoring method were considered unsuccessful in the course according to the $Q_j - 0$ method when the criterion value was 50 points, and about 42% of such students were considered unsuccessful in the latter method when the criterion value was 60 points.

The results of the study showed that weighted scoring based on item difficulty may contribute to revealing differences between individuals. However, another result of the study showed that the number of the students who succeeded or failed in the course changed when an absolute criterion-based assessment was performed to examine how the passed-failed decisions about the students changed according to the scoring method. It was concluded that the number of the students who failed in the course increased when weighting was performed according to item difficulty. Based on these results, as it is thought that the $Q_j - 0$ scoring method may reveal learning differences between students better than the 1 – 0 scoring method, it is recommended that weighted scoring based on item difficulty is used along with conventional scoring in order to obtain more information about the students in classroom assessments. However, at this point, it is necessary to pay attention to the fact that the number of students who will be considered unsuccessful in the course will be higher when scoring is made according to the $Q_j - 0$ method. Therefore, it is thought that the $Q_j - 0$ scoring method may be preferred to the 1 – 0 scoring method, especially in formative assessments, rather than assessment practices in which very critical decisions are made about the students. The fact that the sample was not broad enough to increase differentiation in terms of the measured characteristics was a limitation of this study. For researchers who want to conduct a similar study in a larger and wider group, it may be recommended to examine how different types of scoring lead to differences, especially on the student level, using tests that measure different content areas.

Statements of publication ethics

The ethics committee approval for present research was given by Pamukkale University Social and Humanities Ethics Committee with the issue number E-93803232-622.02-21638 and authors declare that the principals of research and publication ethics were followed.

Researchers' contribution rate

The first author contributed to the finding the problem statement and data collection. All the authors contributed to the literature review, data analysis and interpretation the results, reporting, and the checking the final form of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

Akkuş, O. & Baykul, Y. (2001) Çoktan seçmeli test maddelerinin puanlamada, seçenekleri farklı biçimlerde ağırlıklandırmanın madde ve test istatistiklerine olan etkisinin incelenmesi [An investigation on the effects of different item-option scoring methods on item and test parameters]. *Hacettepe University Journal of Education, 20*, 9-15.

Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short answer questions in a marketing context. *Journal of Marketing Education*, *25*, 31-36. doi: 10.1177/0273475302250570

Bar-Hillel, M., Budescu, D., & Attali, Y. (2005). Scoring and keying multiple choice tests: A case study in irrationality. *Mind and Society*, *4*, 3-12. doi: 10.1007/s11299-005-0001-z

Bejar, I., & Weiss, D.J., (1977) A comparison of empricial differential of inter-item correlation. *Educational and Pyschological Measurement. 37*, 335-340. doi: 10.1177/001316447703700207

Bereby-Meyer, Y., Meyer, Y., & Flascher, O. M. (2002). Prospect theory analysis of guessing in multiple choice tests. *Journal of Behavioral Decision Making*, *15*, 313–327. doi: 10.1002/bdm.417

Buckles, S., & Siegfried, J.J., (2006). Using in-depth multiple-choice questions to evaluate in-depth learning of economics. *Journal of Economics Education, 37*, 48-57. doi: 10.3200/JECE.37.1.48-57.

Budescu, D. V., & Bar-Hillel, M. (1993). To guess or not to guess: A decision-theoretic view of formula scoring. *Journal of Educational Measurement*, *30*(4), 277–291. doi: 10.1111/j.1745-3984.1993.tb00427.x

Budescu, D. V. (1979) *Differential weighting of multiple-choice items*. Educational Testing Service Princeton.

Burton, R. F. (2001). Quantifying the effects of chance in multiple choice and true/false tests: Question selection and guessing of answers. *Assessment & Evaluation in Higher Education, 26*(1), 41–50. doi: 10.1080/02602930020022273

Clark, D., & Linn, M. C. (2003). Designing for knowledge integration: The impact of instructional time. *Journal of the Learning Sciences*, *12*, 451–493. doi: 10.1207/S15327809JLS1204_1

Choppin, B. H. (1988). Correction for guessing. In J. P. Keeves (Ed.), *Educational research, methodology, and measurement: An international handbook* (pp. 384–386). Pergamon Press.

DiBattista, D. & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *The Canadian Journal for the Scholarship of Teaching and Learning*, 2(2). doi: 10.5206/cjsotl-rcacea.2011.2.4

Donlon, T.F. & Fitzpatrick, A.R. (1978) *The statistical structure of multiple choice items*. In *Proceedings of the Annual Meeting of the Northeastern Educational Research Association, Oct. 1978*, Ellenville, New York.

Echternacht, G. (1976) Reliability and validity of item option weighting schemes. *Educational and Pyschological Measurement*, *36*, 301-309. doi: 10.1177/001316447603600208

Gözen, G. (2006). Kısa cevaplı ve çoktan seçmeli maddelerin "0-1" ve ağırlıklı puanlama yöntemleri ile puanlanmasının testin psikometrik özellikleri açısından incelenmesi [Analysis of short–answered and multiplechoice items via "1-0" and weighted scoring methods according to pyschometric characteristics of tests]. *Educational Science & Practice*, *5*(9), 35-52

Frary, R. (1989) Partial credit scoring methods for multiple choice test. *Applied Measurement in Education*, *2*(1), 79-96. doi: 10.1207/s15324818ame0201_5

Hendrickson, G., (1971) *The effect of differential option Weighting on multiple choice objective test items*. Report Number 93, The John Hopkins University.

Heubert, J. P., & Hauser, P. M. (1999). *High-stakes testing for tracking, promotion, and graduation*. National Academy Press.

Kubinger, K. D., Holocher-Ertl, S., Reif, M., Hohensinn, C., & Frebort, M. (2010). On minimizing guessing effects on multiple-choice items: Superiority of a two solutions and three distractors item format to a one solution and five distractors item format. *International Journal of Selection and Assessment*, *18*(1), 111–115. doi: 10.1111/j.1468-2389.2010.00493.x

Prihoda, T. J., Pinckard, R. N., McMahan, C. A., & Jones, A. C. (2006). Correcting for guessing increases validity in multiple-choice examinations in an oral and maxillofacial pathology course. *Journal of Dental Education, 70*(4), 378-386. doi: 10.1002/j.0022-0337.2006.70.4.tb04092.x

Jaradat, D. & Tollefson, N. (1997) The impact of alternative scoring procedure for multiple choice items on test reliability, validity and grading. *Educational and Pyschlogical Measurement*, *48*, 627-635. doi: 10.1177/0013164488483006

Mavis, B. E., Cole, B. L., & Hoppe, R. B. (2001). A survey of student assessment in U.S. medical schools: The balance of breadth versus fidelity. *Teaching and Learning in Medicine*, *13*, 74-79. doi: 10.1207/S15328015TLM1302_1

McDougall, D. (1997). College faculty's use of objective tests: State-of-the-practice versus state-of-the-art. *Journal of Research and Development in Education*, *30*, 183–93.

Merwin, J. (1959) Rational and mathematical relationships of six scoring procedures applicable to three-choice items. *Journal of Educational Psychology*, *50*(4). doi: 10.1037/h0045073

Özdemir, D. (2003). Çoktan seçmeli testleri puanlama yöntemlerine bir bakış [An overview of methods for scoring multiple choice tests]. *Eğitim Araştırmaları Dergisi*, *4*(12),121-122

Özdemir, D. (2004) Çoktan seçmeli testlerin klasik test teorisi ve örtük özellikler teorisine göre hesaplanan psikometrik özelliklerinin iki kategorili ve ağırlıklandırılmış puanlaması yönünden karşılaştırılması [A comparison of psychometric characteristics of multiple choice tests based on the binarys and weighted scoring in respect to classical test and latent trait theory]. *Hacettepe University Journal of Education*, *26*, 117-123

Palmer, E.J. & Dewitt,P.G. (2007) Assessment of higher order cognitive skills in undergraduate education: modified assey or multiple choice questions? *BMC Medical Education*, *20*, 129-158. doi: 10.1186/1472-6920-7-49

Ramsay, J.O. (1968) A scoring system for multiple choice test items. *The British Journal of Mathematical and Statistical Psychology*, *41*, 249-262. doi: 10.1111/j.2044-8317.1968.tb00413.x

Reilly R.R., & Jackson,R. (1972). Effects of empirical option weighting on reliability and validity of GRE. *Journal of Educational Measurement*, *10*(3), 185-193. doi: 10.1111/j.1745-3984.1973.tb00796.x

Roediger, H. L., III, & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1155–1159. doi: 10.1037/0278-7393.31.5.1155

Rowley, G.L., & Traub, R.e (1977) Formula scoring, number–right scoring, and test-taking strategy. *Journal of Educational Measurement*, *14*, 15-22.

Sax, G. (1989) *Principle of educational and pscyhological measurement and evaluation*. Wadsworth.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*, 4–14. doi: 10.3102/0013189X029007004

Kurz. T. B. (1999). *A review of scoring algorithms for multiple choice tests.* EDRS Publications, Report NO: ED 428 076

Walsh, C.M. & Seldomridge, L.A. (2006). Critical thinking: Back to square two. *Nursing Education*, *45*, 212-219. doi: 10.3928/01484834-20060601-05

Weitzman, R.A: (1970) Ideal multiple choice items. *Journal of The American Statistical Association*, *65*(329), 71-89. doi: 10.1080/01621459.1970.10481063

Wilson, M., & Wang, W. C. (1995). Complex composites: Issues that arise in combining different modes of assessment. *Applied Psychological Measurement*, *19*, 51–71. doi: 10.1177/014662169501900107

Yurdugül, H. (2010) Farklı madde puanlama yöntemlerinin ve test puanlama yöntemlerinin karşılaştırılması [Different item scoring methods and different test scoring comparison of methods]. *Journal of Measurement and Evaluation in Education and Psychology*, *1*(1) 1-8.