2025, VOL 15, NO 2, 1254-1266
DOI: 10.48146/odusobiad.1599977
Received Date: 11.12.2024
Acceptence Date: 27.03.2025

Research Article *Open Access*

Effect of random item ordering in multiple choice tests on the academic achievement of online learners

Necati Taşkın[1], Bülent Kandemir[1]

[1] Ordu University, Distance Education Application and Research Center, Ordu, Turkiye.

ABSTRACT

This study examined the effect of randomly ordered multiple-choice test forms on students' academic achievement. This study was carried out with a true experimental design. All students participating in the study received their training through online learning. The study group for the research consisted of 2932 freshman university students studying at different faculties in a state university in the fall semester of the 2023-2024 academic year. A 20-item multiple-choice test was used to measure the students' academic achievement. Four different test forms were generated by randomly ordering the items with medium difficulty levels. One-way ANOVA was used to test whether there was a significant difference between the mean scores for students using different test forms. Additionally, test statistics for different test forms were examined. The findings showed no significant difference between the students' mean scores with different test forms (1) and that the score distributions were balanced (2). As a result of this study, it was found that different test forms created through random ordering do not affect students' academic achievements. This study suggests that educators, test developers, and administrators can create different test forms through randomization as a cheating prevention method in multiple-choice tests. Although the test forms had statistically similar results and did not differ in terms of student achievement, the low reliability coefficients of the tests raise the question of whether random item ordering is appropriate for online learners.

KEYWORDS

Assessment, online learning, multiple-choice tests, random item ordering, academic achievement.

Çoktan seçmeli testlerde maddelerin rastgele sıralanmasının çevrimiçi öğrenenlerin akademik başarıları üzerindeki etkisi

ÖZET

Bu çalışmada, rastgele sıralanmış çoktan seçmeli test formlarının öğrencilerin akademik başarılarına etkisi incelenmiştir. Çalışma gerçek deneyselde gerçekleştirilmiştir. Çalışmaya katılan tüm öğrenciler eğitimlerini çevrimiçi öğrenme yoluyla almıştır. Araştırmanın çalışma grubunu, 2023-2024 eğitim-öğretim yılı güz yarıyılında bir devlet üniversitesinin farklı fakültelerinde öğrenim gören 2932 birinci sınıf üniversite öğrencisi oluşturmaktadır. Öğrencinin akademik başarısını ölçmek için 20 soruluk çoktan seçmeli test kullanılmıştır. Orta zorluk derecesine sahip maddeler rastgele sıralanarak dört farklı test formu oluşturulmuştur. Farklı test formlarını alan öğrencilerin ortalama puanları arasında anlamlı bir fark olup olmadığını test etmek için One-way ANOVA kullanılmıştır. Ayrıca farklı test formlarına ait test istatistikleri incelenmiştir. Bulgular, öğrencilerin farklı test formlarındaki ortalama puanları arasında anlamlı bir fark olmadığını (1) ve puan dağılımlarının dengeli olduğunu (2) göstermiştir. Bu çalışmanın sonucunda, rastgele madde sıralamasının öğrencilerin akademik başarılarını etkilemediği ortaya çıkmıştır. Bu çalışma, eğitimcilere, test geliştiricilere ve yöneticilere, çoktan seçmeli testlerde güvenli bir değerlendirme sağlamak için rastgele maddde sıralaması yoluyla farklı test formları oluşturmalarını önermektedir. Test formları istatistiksel açıdan benzer sonuçlar gösterse ve öğrenci başarısı açısından farklılık yaratmasa da, testlerin düşük güvenilirlik katsayıları, rastgele madde sıralamasının çevrimiçi öğrenenler için uygun olup olmadığı sorusunu gündeme getirmektedir.

ANAHTAR KELİMELER

Sorumlu Yazar: necatitaskin@odu.edu.tr

Değerlendirme, çevrimiçi öğrenme, çoktan seçmeli test, rastgele madde sıralaması, akademik başarı

Introduction

Assessment, one of the most difficult and time-consuming processes in education, is carried out to obtain information about students and improve the teaching process (Butler, 2018). In this process, measurement tools are administered as a summative assessment to measure the post-training learning outcomes of the students (Biesta, 2009). Students' knowledge and skills are measured and scored using these tools. These scores indicate students' academic achievement and how much they benefited from the training (Good, 1973). Multiple-choice tests are the most commonly used measurement tool to determine academic achievement (Schuwirth & Van der Vleuten, 2004; Smith, 2020). Multiple choice tests consist of a question and items including one correct option and distracting options (Tamir, 1991). Tests offer the opportunity to measure a wide curriculum because they contain several questions. All learning objectives can be covered by using items that address different cognitive levels (Lowe, 1991). Thanks to computer-aided systems, scores are obtained quickly and objectively. The objective measurement and ease of scoring are the most important factors that make multiple-choice tests stand out (Baghaei & Amrahi, 2011; Roediger & Marsh, 2005). For these reasons, multiple-choice tests have been chosen for large-scale national and international evaluations for many years.

Despite significant advantages, multiple-choice tests are vulnerable to cheating (Şad, 2020). Students can view other students' answer sheets and reach the correct answer by whispering or signaling to each other. For this reason, some precautions are taken to minimize the risk of cheating, such as proctoring and sitting in single rows. The most common of these measures is the use of different forms for the test. For this purpose, equivalent items are prepared or different test forms are generated using the same items. Random ordering of items is the most popular method (Davis, 2017). Randomizing the item order is an effective way to create different test forms (Carnegie, 2017). Creating different test forms by ordering the same items allows students to be evaluated more fairly (Sue, 2009). This method significantly prevents students from cheating (Gyamfi, 2022). The use of this method has become more widespread with the emergence of computer programs that automatically distribute the items randomly.

The most important parameter to consider when preparing different test forms is that they should be equivalent (Opara, 2021). The test should provide equal opportunities to all students and different forms should not affect the performance of students (Papenberg et al., 2021). However, Stanley (1961) stated that a difficult test item will reduce student performance as it will affect the responses to the next few items. It was also claimed that students who encounter difficult items in the early stages of the test will experience a decrease in motivation and excessive time consumption (Cronbach, 1970; Leary & Dorans, 1985). For this reason, starting the test with easy questions has become a generally accepted practice. (Hambleton & Traub, 1974; Hodson, 1984; Skinner, 1999). From another perspective, items preceding an item can improve performance by providing students with a set of cues. The opposite situation can also be confusing and cause student performance to decrease (Carlson & Ostrosky, 1992). Canlar and Jackson (1991) considered students who received the form in which related items were consecutive to be lucky students. Randomizing test items has the potential to affect measurement if a student encounters difficult items early (Şad, 2020) or if items follow the flow of the topic (Baldwin & Howard, 1983). Therefore, there is a strong belief among students that randomly ordered tests are more difficult (Pettijohn & Sacco, 2007) and that this affects their test scores (Bard & Weinstein, 2017). Additionally, educators are concerned about whether item ordering methods provide fair assessment for students (Stout & Heck, 1995).

The effort to prevent cheating is required to provide a reliable assessment (Surahman & Wang, 2022). If item ordering affects student performance and/or the equivalence of test forms, the

assessment is threatened by a factor beyond student control. Such an effect is not desired by either educators or students. Different forms must consistently measure the same basic characteristics; in short, they must be equivalent (Borsboom & Molenaar 2015). As Green (1981) pointed out, achieving this equivalence requires tests to have not only similar score distributions but also similar statistical values. However, in studies investigating the effect of item order on students' academic achievement, the focus has generally been on student performance and the structure of the test has not been taken into account (Aamodt & McShane, 1992; Hauck et al., 2017). Additionally, unlike previous studies, the students in this study were online learners. In this context, this study examined the effects of four different test forms (A-B-C-D) created with random item ordering on students' academic achievement, by considering the structure of the test forms. With this main purpose, answers were sought to the following research questions.

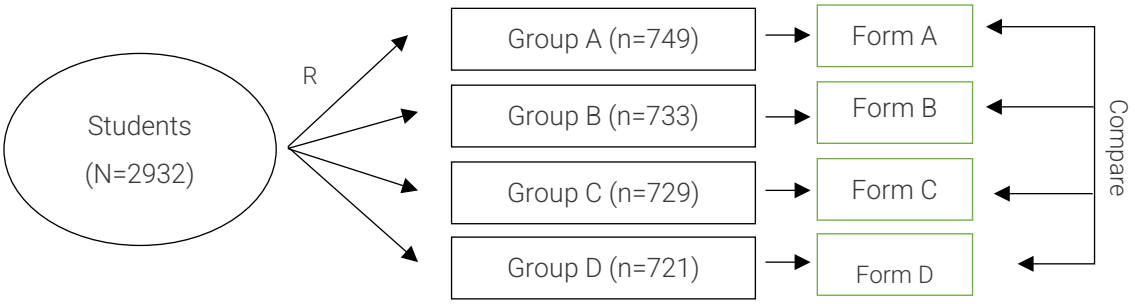
a) What is the distribution of scores on randomly ordered test forms and what is the structure of the test forms?

b) Does student academic achievement differ significantly across different test forms?

Method

Research design

This study was carried out with a true experimental design. Students were randomly assigned to levels of the independent variable in order to minimize the effect of individual differences that may initially exist between the groups and to increase internal validity (Fraenkel & Wallen, 2006). In this way, the impact of possible differences in the students' prior knowledge levels about the subject of the course were also minimized. After the training, the academic achievement test was applied to the groups. All students participating in the study received their training through online learning. The trainings were carried out by the same instructor and using the same course materials. Students' academic achievements were measured using different forms (A, B, C, and D) of the same multiple-choice test, created through random item ordering. In this way, the average scores in the groups were compared and possible differences arising from the order of items in the multiple-choice tests were investigated.



R: Random assignment; Group A, B, C, D: Groups formed by random assignment; Form A, B, C, D: Multiple choice test forms created through random ordering

Figure 1 Research design

Study group

The study group for the research consisted of 2932 freshman university students studying at different faculties in a state university in the fall semester of the 2023-2024 academic year. The study was carried out with Ataturk's Principles and Revolution History course, which is compulsory for all students. The distribution of students into groups is given in Table 1.

Table 1 Distribution of students

Faculty/College	Form A		Form B		Form C		Form D		Total	
	N	%	N	%	N	%	N	%	N	%
Agriculture	23	26.44	22	25.29	20	22.99	22	25.29	87	2.97
Arts and sciences	78	24.3	85	26.48	79	24.61	79	24.61	321	10.95
Dentistry	27	25.23	26	24.3	27	25.23	27	25.23	107	3.65
Economics and administrative	47	24.48	46	23.96	49	25.52	50	26.04	192	6.55
Education	84	25.3	80	24.1	83	25	85	25.6	332	11.32
Fine arts	18	23.38	22	28.57	19	24.68	18	23.38	77	2.63
Health science	35	26.52	33	25	32	24.24	32	24.24	132	4.5
Marine sciences	18	22.5	23	28.75	22	27.5	17	21.25	80	2.73
Music and performing arts	21	27.27	18	23.38	18	23.38	20	25.97	77	2.63
Social sciences	83	27.39	72	23.76	74	24.42	74	24.42	303	10.33
Sports sciences	33	31.13	25	23.58	28	26.42	20	18.87	106	3.62
Technical sciences	80	25.72	79	25.4	75	24.12	77	24.76	311	10.61
Theology	22	27.5	20	25	16	20	22	27.5	80	2.73
Tourism	13	27.66	14	29.79	10	21.28	10	21.28	47	1.6
Vocational schools	167	24.56	168	24.71	177	26.03	168	24.71	680	23.19
Total	749	25.55	733	25	729	24.86	721	24.59	2932	100

Implementation process

The Ataturk's Principles and Revolution History course lasted 8 weeks. Courses conducted via online learning were taught synchronously (live) for 2 hours per week by the same instructor. Courses were recorded and made available to students who did not attend the live course. Students were able to access course recordings wherever and whenever they wanted. In addition, the slides and documents were shared with the students every week via the learning management system (MOODLE). At the end of the training, a multiple-choice test was administered to determine the student's academic achievements.

Tests were administered face-to-face in a traditional paper-and-pencil testing format. Students were randomly assigned to classes within each faculty. Each student was seated in a single row and the tests were conducted under the supervision of instructors. Four different multiple-choice test forms were distributed to the students sequentially. Students were given 20 minutes for the test consisting of 20 questions. Tests were held simultaneously (2.00 PM) in all faculties of the university.

Data collection tools

To measure the students' academic achievement, a 20-item multiple-choice test was prepared by the course instructor according to the learning objectives of the course (YÖKA1, 2018). The distribution of items across topics is given in Table 2.

Table 2 Distribution of test items

Week	Topic	Number of Questions	Question Number
1	Concepts like revolution, reform, republic etc.	2	1, 2
2	The structure of the Ottoman Empire	3	3, 4, 8
3	Constitutional developments in the Ottoman Empire	3	5, 6, 7
4	Political parties and intellectual movements	3	9, 10, 11
5	Trablusgarp (Italo-Turkish) and Balkan Wars, World War I, Armistice of Mudros	4	12, 13, 14, 17
6	Partition Plans for the Ottoman Empire		
7	Life of Mustafa Kemal Pasha and the situation in Anatolia	2	16, 18
8	Amasya, Erzurum, Sivas, and Other National Congresses	3	15, 19, 20

Since the test covers the beginning topics on the course, it focused on the learning goal of "understanding the historical foundations of Ataturk's principles". Items were prepared at the 'knowledge' and 'comprehension' levels of Bloom's taxonomy. This choice is due to the course being at an introductory level and the students having limited prior knowledge of the subject. Test items were selected from the question pool categorized by topic, and opinions were obtained from other faculty members who were experts in the relevant field to ensure content validity. In addition, feedback was received from an academic who is an expert in the field of

measurement and evaluation to evaluate the suitability of the items in terms of measurement tools. The distribution of the items according to Bloom's taxonomy is given in Table 3.

Table 3 Distribution of items according to Bloom taxonomy

Cognitive Domain	Knowledge	Comprehension
Question	1, 2, 5, 7, 8, 9, 10, 12, 13, 14, 15, 17, 18, 19, 20	3, 4, 6, 11, 16

A reliability coefficient was not calculated directly because the test items had not been used together in a single form before. However, the reliability of the test was indirectly ensured by evaluating the discrimination and difficulty indices obtained from previous test uses of each item. The test had medium difficulty level (Başol, 2018), which is considered the ideal difficulty index (pj=0.40-0.60). The test items had discrimination power of over 0.30. Each test item had four options, one of which was correct and the other three were distractors. Each correct answer was worth 5 points, and a minimum of 0 and a maximum of 100 points were received from the test. Four different test forms were created by changing the order of the items through random ordering. The order of the items in the test forms is given in Table 4.

Table 4 The order of the items in the test forms

Form A	Form B	Form C	Form D
Q 1	Q 12	Q 5	Q 8
Q 2	Q 13	Q 6	Q 13
Q 3	Q 17	Q 7	Q 19
Q 4	Q 19	Q 14	Q 2
Q 5	Q 20	Q 15	Q 3
Q 6	Q 14	Q 16	Q 16
Q 7	Q 15	Q 17	Q 1
Q 8	Q 18	Q 8	Q 9
Q 9	Q 16	Q 18	Q 17
Q 10	Q 7	Q 1	Q 4
Q 11	Q 8	Q 2	Q 5
Q 12	Q 9	Q 19	Q 14
Q 13	Q 1	Q 11	Q 20
Q 14	Q 2	Q 12	Q 12
Q 15	Q 6	Q 3	Q 6
Q 16	Q 3	Q 13	Q 15
Q 17	Q 4	Q 20	Q 18
Q 18	Q 10	Q 9	Q 10
Q 19	Q 5	Q 10	Q 11
Q 20	Q 11	Q 4	Q 7

Data analysis

One-way ANOVA was used to test whether there were significant differences between the mean scores of students who used different test forms. The compared scores were independent and exhibit normal distribution (Table 5) with skewness and kurtosis values of the scores in the range of ±1 (Hair et al., 2013). The distributions of the variances of the group scores were equal [Levene F(3,2928)=0.434, p=0.729, p>0.05].

Table 5 Skewness and kurtosis values of the scores

Variable	Skewness		Kurtosis	
	Statistic	Std. Error	Statistic	Std. Error
Form A	0.364	0.089	-0.259	0.178
Form B	0.374	0.090	-0.300	0.180
Form C	0.375	0.091	-0.267	0.181
Form D	0.330	0.091	-0.463	0.182

The mean score, standard deviation, test difficulty, and reliability coefficient were calculated separately for each test form. The reliability of the test, which has different item difficulty levels, was calculated using the KR-20 reliability coefficient (Kuder & Richardson, 1937).

Findings

The findings obtained from the analyses are presented under two subheadings according to the research questions.

Findings regarding score distribution and structure of the test forms

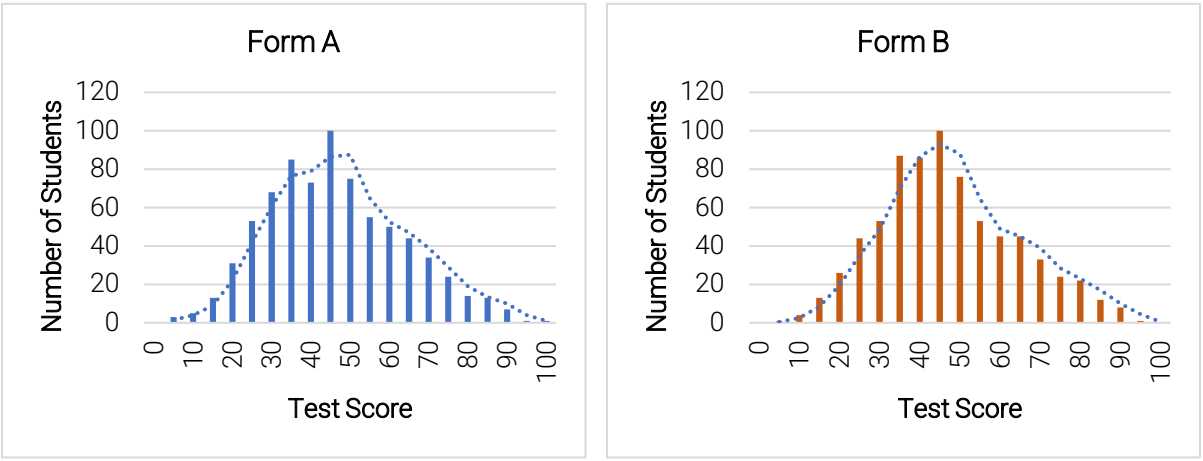
The mean scores and standard deviations for students using different test forms are given in Table 6.

Table 6 Mean score and standard deviations

	N	M	SD	Mode	Median	Range
Form A	749	46.07	17.33	45	45	90
Form B	733	47.04	17.11	45	45	95
Form C	729	46.28	17.13	40	45	90
Form D	721	47.62	17.46	35	45	95

The students' mean scores in different test forms were 46.07 for those using Form A; 47.04 for those using Form B; 46.28 for those using Form C; and 47.62 for those using Form D. Standard deviations were 17.33 for Form A, 17.11 for Form B, 17.13 for Form C and 17.46 for Form D. While the students who took the test with Form D had the highest mean score (M= 47.62), the students who took the test with Form A had the lowest mean score (M= 46.07). There appears to be a difference of 1.55 points between the highest mean and the lowest mean.

The most frequently repeated value (mode) in Form A (n=100) and Form B (n=100) was 45. This value was 40 in Form C (n=91) and 35 in Form D (n=88). When the scores are ranked from low to high, the median value for all forms was 45. In skewed distributions, the arithmetic mean and median move away from each other (Kaplan & Saccuzzo, 2001). At this point, the distance between the arithmetic mean and median values between the forms is close. The range between the highest and lowest scores for the forms are also close to each other. So, the scores for each form are heterogeneous; in other words, they distinguish between those who know and those who do not know the topic (Kaplan & Saccuzzo, 2001). The test score distributions of the students are shown in Figure 2.



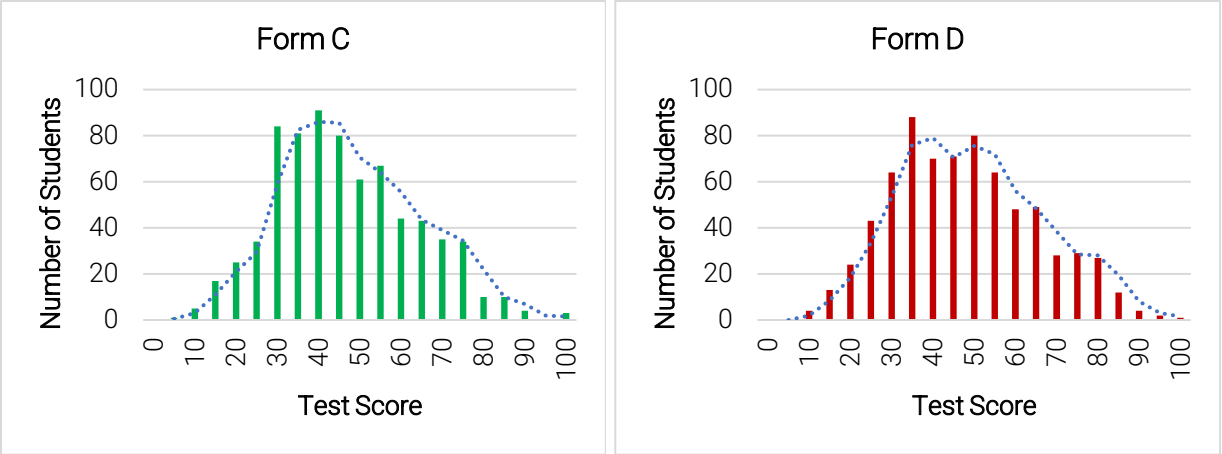


Figure 2 Score distributions

The distributions of student test scores appears to be similar in Figure 2. Score distributions are skewed to the right in all forms. The positive skewness values (Form A=0.364; Form B=0.374; Form C=0.375; Form D=0.330) also show this (see Table 5). These values show that the students' scores are not high. Statistics for the test forms are shown in Table 7.

Table 7 Statistics for test forms

Test Form	$\bar{X}$	Sx2	Sx	rj(KR-20)	p	SHX
Form A	9.24	11.86	3.44	0.65	0.46	2.05
Form B	9.41	11.68	3.42	0.64	0.47	2.05
Form C	9.26	11.80	3.44	0.64	0.46	2.05
Form D	9.54	12.15	3.49	0.66	0.48	2.05

$\bar{X}$: arithmetic mean, Sx2: variance, Sx: standard deviation, rj: reliability, p: average difficulty, SHX: standard error

The reliability coefficient of KR-20 was found to be 0.65 for Form A, 0.64 for Form B, 0.64 for Form C and 0.66 for Form D. While Başol (2018) stated that the KR-20 reliability coefficient exceeding the threshold value of 0.70 indicates that the test reliability is high, Kalaycı (2008) considers values between 0.60 and 0.80 to be acceptable. In short, although these reliability coefficients do not indicate that the different test forms had high internal consistency, they are similar. The standard error value is also the same (SHX=2.05) for all forms of the test. The forms are affected by similar external factors and are consistent in terms of elements that are not related to the measured characteristics (Thissen, 2017). The average difficulty indexes for the test forms were 0.46, 0.47, 0.46 and 0.48. All forms exhibited a medium difficulty level, with an ideal value between 0.60 and 0.40 (Başol, 2018). In addition, test statistics such as the reliability, average difficulty, and standard error values of the test forms are similar. From here, the test forms have statistically close values and are somewhat equivalent.

Findings regarding of academic achievement

The ANOVA test results, which determine whether different test forms created a significant difference in students' test scores, are given in Table 8.

Table 8 ANOVA results of test scores

Form	N	M	SD	df	F	p
Form A	749	46.07	17.33	3-2928	1.255*	0.288
Form B	733	47.04	17.11			
Form C	729	46.28	17.13			
Form D	721	47.62	17.46			

Note. * not significant at $p < 0.05$

The analysis results show that different forms of the multiple-choice test did not create a significant difference in students' test scores [F (3-2928) = 1.255, p>.05]. Although the mean

scores for students taking Form D (M = 47.62) were higher than the mean scores for students taking Form A (M = 46.07), this difference is not significant. According to this finding, different test forms created through random item ordering did not have a significant effect on students' academic achievement.

Discussion

In this study, the effect of different test forms created with random ordering on students' academic achievement was examined. In addition, statistical changes between different test forms were examined. The findings showed that the distribution of scores for students using different test forms was balanced (1) and there was no significant difference between their mean scores (2). In addition, different test forms created with random order had statistically close values.

There are limited studies in the literature comparing different test forms created through random ordering. Schimit and Sheirer (1977) stated that there was no significant difference in student performance between three randomly ordered test forms. Gyamfi et al. (2023) compared five randomly ordered test forms and found that the order of multiple-choice test items did not affect students' performance. Peek (1994) found that randomizing the order of test items did not have a significant effect on students' scores and recommended that educators use randomization to order test items. These findings are consistent with this study's finding that there was no significant difference between students' mean scores on different test forms. However, Vander Schee (2009) stated that random ordering creates a disadvantage for successful students. Similarly, Stout and Heck (1995) argued that students who take the random version of the test are at a disadvantage. In the literature, ordering test items according to the topic flow (Russell et al. 2003; Opara, & Uwah, 2017) and order of increasing item difficulty provides an advantage for students (Baffoe et al., 2024; Weinstein & Roediger, 2012). This raises the question of whether random ordering negatively affects student performance.

Ordering test items according to the topic flow improves student performance compared to random ordering (Canlar & Jackson, 1991; Gruber, 1987). This performance increase is explained by the cognitive learning process (Balch, 1989; Baldwin & Howard, 1983; Norman, 1954). Cronbach (1950) emphasized that the order of topic flow is important in encoding and remembering information in memory. New information is encoded into long-term memory by connecting with previous information (Ertmer & Newby, 2013). According to schema theory, semantic networks connect information in the mind (Frederiksen et al., 1999). Organizing and structuring information in the mind makes retrieving it easier (Sweller, 2011; Kirschner, 2002). In this context, while topic-ordered tests may improve students' performance, randomly ordered tests may cause confusion and decreased performance (Carlson & Ostrosky, 1992). Considering the theoretical framework, topic-ordering tests may positively affect student performance.

It was also observed that ordering test items according to increasing item difficulty positively affects students' performance compared to random ordering (Hodson, 1984; Plake, 1982). Students who encounter difficult items at the beginning of the test have lower performance (Doğan Gül & Çokluk Bökeoğlu, 2018; Paretta & Chadwick, 1975; Vander Schee, 2013). Having difficult questions at the beginning of the test increases anxiety and reduces performance (Zeidner, 1998; McKeachie et al., 1955). For this reason, ordering the items from easy to difficult is thought to increase students' attention and motivation (Linn & Gronlund, 1995; Skinner, 1999). Random ordering of test items may increase the risk due to the possibility of students encountering difficult questions too early (Şad, 2020). Accordingly, the random ordering of the items may decrease the achievement of the students.

While difficult question items at the beginning of the test may negatively affect students' emotional and psychological states, starting with easy questions may increase their

motivation. It should also be taken into consideration that topic-order tests may increase students' performance. However, how the ordering affects the psychometric properties of the test is also crucial. It should not be forgotten that the item order will change the psychometric properties of the test as well as the test scores of the students (Hodson, 1984). In this study, the test reliability, test difficulty levels, and standard error values for the different forms created by random order were similar. Although the forms are equivalent, accurate measurement is only possible with measurement instruments that have valid psychometric properties (Cook & Beckman 2006). Although this study found that the forms created with random order had no effect on academic achievement and that the test forms were equivalent to each other, it should not be forgotten that the reliability coefficient of the test forms was low. Perhaps different ordering methods may increase the psychometric properties of the test, such as reliability, along with student performance.

Online learning pedagogy is different from face-to-face education, and assessment methods should be appropriate for the learning process (Gikandi et al., 2011). Siddiqui et al. (2024) stated that students perceive online learning differently than face-to-face learning and that these differences should be reflected in assessment methods. Therefore, the assessment process must be consistent with the structured and sequential content in the online learning process (Howard & Scott, 2017). Since the content in online learning environments is presented in a certain structure and logical order, failure to maintain this structure during the assessment phase may disadvantage students. In other words, the random order of the items may have made it difficult for students to respond appropriately to the structured learning process, leading to a decrease in their performance. This indicates that test forms created according to the topic or with increasingly difficult item order may increase students' performance. However, based on the findings of this study, different test forms created through random ordering provide fair evaluation of students.

Conclusion and suggestions

This study showed that educators' concerns that some students are disadvantages by randomly ordering test forms are unfounded. Based on the findings obtained in this study, different test forms created through random ordering do not affect the academic achievements of the students, and the test forms are equivalent. This study suggests that educators, test developers, and administrators can create different test forms through the randomization of items as a cheating prevention method in multiple-choice tests. In this context, instructors should be encouraged to use different randomly ordered test forms to reduce students' attempts to cheat on tests. However, in this study, the test forms were evaluated based on their average difficulty levels. Focusing only on the average difficulty level may lead to misleading conclusions about the effects of item order. Therefore, it is recommended that in future studies, the difficulty index be addressed at the item level and the analyses be deepened to consider the individual difficulty of each item on different test forms.

Although this study found that randomly ordered multiple-choice test forms did not have a significant effect on students' academic achievement, if random ordering creates a disadvantage for students compared to other ordering methods, this is an important situation to consider. Additionally, studies should be performed to determine which of these ordering methods provides a more accurate measurement. In this context, it is recommended to conduct experimental studies investigating the effects of different forms ordered according to topic flow, difficulty, and random order on students' academic achievements and test statistics in future studies.

Limitations

In paper-and-pencil tests, students are not restricted to answering questions in a particular order and are free to skip forward/backward within the test to look for items they know or find easy. In this study, the order in which students answered the test items was not controlled and

it was assumed that the students answered the test according to the item order on the test form.

Another limitation of this study is that the test forms were evaluated at the average difficulty level. The individual difficulty levels of the items may have differed between test forms. In addition, the items used focused only on the knowledge and comprehension levels of Bloom's taxonomy. As the cognitive level of the items increases, the effect of random order on students' success may differ.

Since this study was conducted only within a specific course, it may be difficult to generalize the findings to different fields. Considering that different course content and test types can have different effects on students, future studies in different fields may help to understand this situation better. In addition, since only multiple-choice test and random ordering were used in the study, the effect of order on other question types and the use of different ordering methods could not be examined. Studies covering different exam formats and question types will address this deficiency.

Author contribution rates

1st Author: 60%, 2nd Author: 40% contributions to the study.

Conflict of interest declaration

Our article entitled "The effect of random item ordering in multiple choice tests on the academic achievement of online learners" has no financial conflict of interest with any institution, organization or person. There is also no conflict of interest between the authors.

References

Aamodt, M. G., & McShane, T. (1992). A meta-analytic investigation of the effect of various test item characteristics on test scores and test completion times. *Public Personnel Management, 21*(2), 151-160. https://doi.org/10.1177/009102609202100203

Baffoe, J., Asamoah, D., Shahrill, M., Latif, S. N. A., Asamoah Gyimah, K., & Anane, E. (2024, April). Does the sequence of items influence secondary school students' performance in mathematics and science?. In AIP Conference Proceedings (Vol. 3052, No. 1). AIP Publishing. https://doi.org/10.1063/5.0202870

Baghaei, P., & Amrahi, N. (2011). The effects of the number of options on the psychometric characteristics of multiple choice items. *Psychological Test and Assessment Modeling, 53*(2), 192-211.

Balch, W. R. (1989). Item order affects performance on multiplechoice exams. *Teaching of Psychology, 16*(2), 75–77. https://doi.org/10.1207/s15328023top1602_9

Baldwin, B. A., & Howard, T. P. (1983). Intertopical sequencing of examination questions: An evaluation. *Journal of Accounting Education, 1*(1), 89–95. https://doi.org/10.1016/0748-5751(83)90010-6

Bard, G., & Weinstein, Y. (2017). The effect of question order on evaluations of test performance: Can the bias dissolve? *Quarterly Journal of Experimental Psychology, 70*(10), 2130-2140. https://doi.org/10.1080/17470218.2016.1225108

Başol, G. (2018). *Measurement and evaluation in education*. Pegem Akademi. https://doi.org/10.14527/9786053645887

Biesta, G. (2009). Good education in an age of measurement: On the need to reconnect with the question of purpose in education. *Educational Assessment, Evaluation and Accountability, 21*, 33-46. https://doi.org/10.1007/s11092-008-9064-9

Borsboom, D., & Molenaar, D. (2015). Psychometrics. In James D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences* (Second Edition, pp. 418-422). Elsevier Ltd. https://doi.org/10.1016/B978-0-08-097086-8.43079-5

Butler, A. C. (2018). Multiple-choice testing in education: Are the best practices for assessment also good for learning?. *Journal of Applied Research in Memory and Cognition, 7*(3), 323-331. https://doi.org/10.1016/j.jarmac.2018.07.002

Canlar, M., & Jackson, W. K. (1991). Alternative test question sequencing in introductory financial accounting. *Journal of Education for Business, 67*(2), 116-119. https://doi.org/10.1080/08832323.1991.10117529

Carlson, J. L., & Ostrosky, A. L. (1992). Item sequence and student performance on multiple-choice exams: Further evidence. *The Journal of Economic Education, 23*(3), 232–235. https://doi.org/10.1080/00220485.1992.10844757

Carnegie, J. A. (2017). Does correct answer distribution influence student choices when writing multiple choice examinations? *Canadian Journal for the Scholarship of Teaching and Learning, 8*(1), 11. http://ir.lib.uwo.ca/cjsotl_rcacea/vol8/iss1/1

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine, 119*(2), 166-e7. https://doi.org/10.1016/j.amjmed.2005.10.036

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement, 10*(1), 3-31. https://doi.org/10.1177/001316445001000l0

Davis, D. B. (2017). Exam question sequencing effects and context cues. *Teaching of Psychology, 44*(3), 263-267. https://doi.org/10.1177/009862831771275

Doğan Gül, Ç., & Çokluk Bökeoğlu, Ö. (2018). The comparison of academic success of students with low and high anxiety levels in tests varying in item difficulty. *Inonu University Journal of the Faculty of Education, 19*(3), 252-265. https://doi.org/10.17679/inuefd.341477

Ertmer, P. A., & Newby, T. J. (2013). Behaviorism, cognitivism, constructivism: Comparing critical features from an instructional design perspective. *Performance Improvement Quarterly, 26*(2), 43-71. https://doi.org/10.1002/piq.21143

Fraenkel, J. R., & Wallen, N. E. (2012). *How to design and evaluate research in education* (7th ed.). McGraw-Hill.

Frederiksen, J. R., White, B. Y., & Gutwill, J. (1999). Dynamic mental models in learning science: The importance of constructing derivational linkages among models. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching, 36*(7), 806-836. https://doi.org/10.1002/(SICI)1098-2736(199909)36:7<806::AID-TEA5>3.0.CO;2-2

Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers and Education, 57*(4), 2333–2351. https://doi.org/10.1016/j.compedu.2011.06.004

Good, V. C. (1973). *Dictionary of education*. N.Y. : McGraw Hill Book Company

Gruber, R. A. (1987). *Sequencing exam questions relative to topic presentation*. Journal of Accounting Education, 5, 77–86. https://doi.org/10.1016/0748-5751(87)90039-X

Green, B. F. (1981). A primer of testing. *American Psychologist, 36*(10), 1001-1011. https://doi.org/10.1037/0003-066X.36.10.1001

Gyamfi, A. (2022). Controlling examination malpractice in Senior High Schools in Ghana through performance-based assessment. *Journal of Advances in Education and Philosophy, 6*(3), 203-211. https://doi.org/10.36348/jaep.2022.v06i04.002

Gyamfi, A., Acquaye, R., & Adjei, C. (2023). Multiple-Choice Items should be sequenced in order of difficulty with the easiest ones placed first. Does it really affect performance? *Research Square*. https://doi.org/10.21203/rs.3.rs-2882983/v1

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2013). *Multivariate data analysis*. Pearson Education Limited.

Hambleton, R. K., & Traub, R. E. (1974). The effects of item order on test performance and stress. *The Journal of Experimental Education, 43*(1), 40-46. https://doi.org/10.1080/00220973.1974.10806302

Hauck, K. B., Mingo, M. A., & Williams, R. L. (2017). A review of relationships between item sequence and performance on multiple-choice exams. *Scholarship of Teaching and Learning in Psychology, 3*(1), 58–75. https://doi.org/10.1037/stl0000077

Hodson, D. (1984). Some effects of changes in question structure and sequence on performance in a multiple choice chemistry test. *Research in Science & Technological Education, 2*(2), 177–185. https://doi.org/10.1080/0263514840020209

Howard, J. M., & Scott, A. (2017). Any time, any place, flexible pace: Technology-enhanced language learning in a teacher education programme. *Australian Journal of Teacher Education (Online), 42*(6), 51-68. https://doi.org/10.14221/ajte.2017v42n6.4

Kalaycı, Ş. (2008). *Spss uygulamalı çok değişkenli istatistik teknikleri*. Ankara: Asil Yayın Dağıtım.

Kaplan, R. M., & Saccuzzo, D. P. (2001). *Psychological testing: Principles, applications, and issues* (5th ed.). Wadsworth/Thomson Learning.

Kirschner, P. A. (2002). Cognitive load theory: Implications of cognitive load theory on the design of learning. *Learning and Instruction, 12*(1), 1-10. https://doi.org/10.1016/S0959-4752(01)00014-7

Kuder, G. F. & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151–160. https://doi.org/10.1007/BF02288391

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research, 55*(3), 387-413. https://doi.org/10.3102/00346543055003387

Linn, R. L., & Gronlund, N. E. (1995). *Measuring and assessment in teaching* (7th ed.). Ohio: Prentice Hall.

Lowe, D. (1991). Set a multiple choice question (MCQ) examination. *British Medical Journal*, 302, 780-782. https://doi.org/10.1136/bmj.302.6779.780

McKeachie, W. J., Pollie, D., & Speisman, J. (1955). Relieving anxiety in classroom examinations. *The Journal of Abnormal and Social Psychology, 50*(1), 93–98. https://doi.org/10.1037/h0046560

Norman, R. D. (1954). The effects of a forward retention set on an objective achievement test presented forwards or backwards. *Journal of Educational & Psychological Measurement, 14*, 487–498. https://doi.org/10.1177/001316445401400305

Opara, I. M. (2021). *Test construction and measurement, concepys and applications*. Reliable Publishers.

Opara, I. M., & Uwah, I. V. (2017). Effect of test item arrangement on performance in mathematics among junior secondary school students in obio/akpor local government area of rivers state Nigeria. *British Journal of Education, 5*(8), 1-9. https://eajournals.org/bje/vol-5-issue-8-july-2017-special-issue/

Papenberg, M., Diedenhofen, B., & Musch, J. (2021). An experimental validation of sequential multiple-choice tests. *The Journal of Experimental Education, 89*(2), 402–421. https://doi.org/10.1080/00220973.2019.1671299

Plake, B. S., Ansorge, C. J., Parker, C. S., & Lowry, S. R. (1982). Effects of item arrangement, knowledge of arrangement test anxiety and sex on test performance. *Journal of Educational Measurement, 19*(1), 49–57. http://www.jstor.org/stable/1434918

Paretta, R. L., & Chadwick, L. W. (1975). The sequencing of examination questions and its effects on student performance. *The Accounting Review, 50*(3), 595-601. https://www.jstor.org/stable/245020

Peek, G. S. (1994). Using test-bank software for randomized test-item sequencing in managerial accounting. *Journal of Education for Business, 70*(2), 77–81. https://doi.org/10.1080/08832323.1994.10117728

Pettijohn,Terry F.,,II, & Sacco, M. F. (2007). Multiple-choice exam question order influences on student performance, completion time, and perceptions. *Journal of Instructional Psychology, 34*(3), 142-149. https://www.proquest.com/scholarly-journals/multiple-choice-exam-question-order-influences-on/docview/213904129/se-2

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*(5), 1155–1159. https://doi.org/10.1037/0278-7393.31.5.1155

Russell, M., Fischer, M. J., Fischer, C. M., & Premo, K. (2003). Exam question sequencing effects on marketing and management sciences student performance. *Journal for Advancement of Marketing Education, 3*, 1–10. https://www.asbbs.org/files/marketing.pdf#page=168

Schimit, J. C. & Sheirer, C. J. (1977). The effect of item order on objective tests. *Teaching of Psychology, 4*(3), 144-153. https://doi.org/10.1207/s15328023top0403_11

Schuwirth L. W. T. & van der Vleuten C. P. M. (2004). Different written assessment methods: What can be said about their strengths and weaknesses? *Medical Education, 38*(9), 974-979. https://doi.org/10.1111/j.1365-2929.2004.01916.x

Siddiqui, A. A., Zain Ul Abideen, M., Fatima, S., Talal Khan, M., Gillani, S. W., Alrefai, Z. A., Waqar Hussain, M., & Rathore, H. A. (2024). Students' perception of online versus face-to-face learning: What do the healthcare teachers have to know? *Cureus, 16*(2), e54217. https://doi.org/10.7759/cureus.54217

Skinner, N. F. (1999). When the going get tough, the tough get going: Effects of item difficulty on multiple-choice test performance. *North American Journal of Psychology, 7*(1), 79-82. https://files.eric.ed.gov/fulltext/ED449388.pdf#page=83

Smith, S. M. (1979). Remembering in and out of context. *Journal of Experimental Psychology: Human Learning and Memory, 5*, 460–471. https://doi.org/10.1037/0278-7393.5.5.460

Stanley, J. C. (1961) Studying status vs. manipulating variables. *Educational and Psychological Measurement, 21*(4), 793-795. https://doi.org/10.1177/001316446102100

Stout, D. E., & Heck, J. L. (1995). Empirical findings regarding student exam performance and question sequencing: The case of the cumulative final. *Journal of Financial Education, 21*, 29-35. https://www.jstor.org/stable/41948181

Sue, D. L. (2009). The effect of scrambling test questions on student performance in a small class setting. *Journal for Economic Educators, 9*(1), 32-41. https://libjournals.mtsu.edu/index.php/jfee/article/view/1454

Surahman, E., & Wang, T. H. (2022). Academic dishonesty and trustworthy assessment in online learning: A systematic literature review. *Journal of Computer Assisted Learning, 38*(6), 1535-1553. https://doi.org/10.1111/jcal.12708

Sweller, J. (2011). Cognitive load theory. In *Psychology of Learning and Motivation* (Vol. 55, pp. 37-76). Academic Press. https://doi.org/10.1016/B978-0-12-387691-1.00002-8

Şad, S. N. (2020). Does difficulty-based item order matter in multiple-choice exams?(Empirical evidence from university students). *Studies in Educational Evaluation, 64*, 100812. https://doi.org/10.1016/j.stueduc.2019.100812

Tamir, P. (1991). Multiple choice items: How to gain the most out of them. *Biochemical Education, 19*(4), 188–192. https://doi.org/10.1016/0307-4412(91)90094-O

Thissen, D. (2017). Reliability and measurement precision. In H. Wainer (Ed.). *Computerized and adaptive testing: A primer* (pp. 161-185). Lawrence Erlbaum.

Vander Schee, B. A. (2009) Test item order, academic achievement and student performance on principles of marketing examinations. *Journal for Advancement of Marketing Education, 14*(1), 23-30. https://www.proquest.com/openview/508fdf8b2223b6d77c9bc85f634dd0fc/1?pq-origsite=gscholar&cbl=5256660

Vander Schee, B. A. (2013). Test item order, level of difficulty, and student performance in marketing education. *Journal of Education for Business, 88*(1), 36-42. https://doi.org/10.1080/08832323.2011.633581

Weinstein, Y., & Roediger, H. L. (2012). The effect of question order on evaluations of test performance: How does the bias evolve?. *Memory & Cognition, 40*, 727-735. https://doi.org/10.3758/s13421-012-0187-3

YÖKA1 (2018). Principles of Ataturk and History of Revolution I, Ordu Universty Course Catalog/Information Package. Retrieved June 6, 2024, from https://bologna.odu.edu.tr/DerecеProgramlari/Ders/0/237/43481/41348/1?lang=en-US

Zeidner, M. (1998). *Test anxiety-the state of art*. USA: Plenum Press.