
# Unpacking the Impact of Item Difficulty: Traditional Testing in Online Learning

**Necati Taşkın**
Ordu University, Turkiye

www.ijte.net

**To cite this article:**

Taskin, N. (2025). Unpacking the impact of item difficulty: Traditional testing in online learning. *International Journal of Technology in Education (IJTE), 8*(4), 998-1021. https://doi.org/10.46328/ijte.1210

IJTE International Journal of Technology in Education
3035, Vol. 8, No. 4, 998-1021 https://doi.org/10.46328/ijte.1210

# Unpacking the Impact of Item Difficulty: Traditional Testing in Online Learning

**Necati Taşkın**

## Article Info

***Article History***

Received:
7 December 2024
Accepted:
11 August 2025

***Keywords***

Online learning
Paper-pencil tests
Test parameters
Item order
Academic achievement
Student perceptions

## Abstract

This study examines the effect of item order (random, increasingly difficult, and decreasingly difficult) on student performance, test parameters, and student perceptions in multiple-choice tests administered in a paper-and-pencil format after online learning. In the research conducted using an explanatory sequential mixed methods design, quantitative data were first analyzed and then qualitative data were collected to examine these findings in depth. 2131 freshman university students participated in the quantitative part of the study and 312 students participated in the qualitative part. After 14 weeks of online foreign language education, tests with different item orders were applied to measure the academic achievement of the students. The findings revealed that item order did not significantly affect academic achievement of students. Item order was found to not affect test parameters such as test difficulty and reliability. The most striking finding the difficulty level of the items changes depending on the item ordering. Findings regarding student perceptions show that encountering difficult questions at the beginning of the test reduces motivation, increases anxiety, and creates a negative perception of the assessment process. Additionally, students emphasize that the assessment process should be compatible with the pedagogical structure of online learning. These findings indicate that students' perceptions as well as test statistics should be taken into account in test design. In this context, conducting the assessment of education provided through online learning in an online environment can eliminate many discussions regarding the item order. Based on these findings, it is recommended that future research be expanded to include different courses, item order method, and individual student differences.

## Introduction

Assessment is one of the three basic elements of education. Assessment undertakes the role of guiding and developing the other two basic elements of education, learning and teaching (Novak et al., 2005). It also determines how effective the teaching process is and whether students are achieving their learning objectives (Harlen et al., 1992). If the assessment is done to guide and improve the teaching process, it is called formative, and if it is done to determine the student's academic achievement, it is called summative (Reynolds et al., 2006). This study focuses on summative assessment. The high risk of academic dishonesty in online assessment raises

concerns about the reliability of assessment processes (Zanetti & Butera, 2025). Although technological solutions such as learning analytics, artificial intelligence, LMS-based solutions, and webcam proctoring are used to ensure security in online assessment (Hilliger et al., 2022), privacy concerns, data protection issues, and high costs hinder this (Nigam et al., 2021). Therefore, summative assessment often requires moving from online environments to face-to-face traditional environments (Xiong & Suen, 2018). This situation brings with it some handicaps. Online learning has a different pedagogical approach where the contents are presented in a structured and logical order. Since students build knowledge step by step in this process, it is expected that the assessment methods will also be suitable for this structure (Howard & Scott, 2017). However, traditional face-to-face tests applied after the online learning process may not match this pedagogical structure. Especially in cases where the questions are randomly ordered, it may be difficult for students to produce answers that are appropriate for their learning styles (Siddiqui et al., 2024). Unfortunately, this situation is often ignored and impossibilities make face-to-face tests mandatory for summative assessment.

A summative assessment is done and scored to see whether the students have achieved the learning objectives. These scores play a critical role in understanding the academic achievements of the student. These scores can be created by evaluating in-class activities, oral presentations, and written documents, or they can be revealed by measurement tools applied at the end of the education (Walsh & Betz, 1995). Undoubtedly, the most popular measurement tool multiple-choice tests (Butler, 2018).

Multiple-choice tests provide an opportunity to objectively measure student achievement effectively and reliably (Lowe, 1991). Scoring of the tests is easy and fast, and they can be evaluated without error with the help of a computer. For this reason, they are widely preferred in assessments involving large groups of students (Anaya et al., 2022). Despite these advantages, multiple-choice tests can lead to the creation of false information through guesswork and chance (Roediger & Marsh, 2005). In addition, since they facilitate communication and interaction with other students, they are vulnerable to cheating (Taşkın, 2024). In large groups, especially in situations where seating space is limited, multiple-choice tests are developed in more than one form to reduce the risk of cheating (Davis, 2017). To develop more than one form of the test, alternative forms are produced by changing the item order (Carnegie, 2017). Although creating different test forms greatly reduces cheating attempts, it raises concerns about whether it affects students' test performance (Gyamfi et al., 2023). There is a widespread and strong belief among students that the item order affects their performance (Bard & Weinstein, 2017).

Bachman (1990) states that the measurement method should not affect students' performance and should not interfere with the structure being measured, whereas the item order has the potential to affect the measurement. In particular, it is suggested that the difficulty level of a previous item may influence the student's response to subsequent items. Cronbach (1970) stated that an incorrect answer to a difficult item can negatively affect students' motivation, while Carlson & Ostrosky (1992) stated that a difficult item encountered early on will cause a decrease in students' test performance. In this context, students feel more successful in tests ordered from easy to difficult and evaluate the test as less challenging (Chen, 2012). On the other hand, Pettijohn & Sacco (2007) found that randomly ordered tests were perceived as more difficult by students. This perception created by the item order on students is an important finding in itself (Weinstein & Roediger, 2012). These findings suggest that test order may

play an important role in both student perceptions and test performance. According to the Cognitive-Attentional Model, anxiety, negative thoughts, and distraction negatively affect human cognitive resources (Naveh-Benjamin et al., 1991). During the test, when students have to deal with negative thoughts instead of focusing on the questions, their cognitive resources are depleted and their performance decreases. Therefore, it is assumed that arranging the test items in order of increasing item difficulty will increase students' motivation and improve their test performance (Akugri, 2023). In fact, over time, it has become a generally accepted approach to start test with easy items (Skinner, 1999).

With random item ordering, students are likely to encounter a difficult item early in the test (Şad, 2020). This is likely to disadvantage some students, leading to unfair outcomes. However, different test formats should give equal chances to all students and no external factors should affect student performance (Gyamfi & Yeboah, 2022). The most important parameter to consider when preparing more than one test form is that the forms created must be equivalent (Papenberg et al., 2021). Findings in the literature on the effects of random order and increasing difficulty order test forms on student performance are contradictory. Studies are showing that students performed better in the increasing difficulty order test (Baffoe et al., 2024) or the random order test (Abdullahi & Akwashiki, 2020), as well as studies showing that no effect was observed (Opara & Ogbanu, 2023). The weakness of these studies is that the effect of item ordering is examined based only on student performance. According to Classical Test Theory, the parameters of the test change depending on the performance of the group (Crocker & Algina, 1986). In particular, changing the item order may affect basic parameters of the test such as reliability, validity and discrimination (Kaplan & Saccuzzo, 2001). The difference in performance between groups or the way the test is administered will affect the test parameters. On the other hand, if the groups are similar to each other and the test contents are the same, the test parameters are expected to remain consistent (Fan, 1998). Therefore, item order is likely to affect the accuracy of measurement by affecting test parameters as well as students' test performance. For a fair and reliable evaluation, test parameters as well as student performance must be taken into account (Anastasi & Urbina, 1997). In this context, the effect of item order (random order, decreasing difficulty order and increasing difficulty order) on students' academic achievement was examined in this study, while test parameters were also taken into account. The quantitative findings obtained were supported by qualitative findings revealing students' perceptions of item order. In this direction, the following research questions were sought:

1. What is the distribution of students' test scores across different test forms?
2. Is there a significant difference in academic achievement among students who take different forms of the test?
3. What are the test parameters in different forms of the test?
4. Is there a relationship between the difficulty indexes of the items in different forms of the test?
5. What are the students' perceptions of the item order?

## Method

### Research Design

This research was conducted using mixed-methods and an explanatory sequential design was used (Figure 1). This design involves first collecting and analyzing quantitative data, and then collecting qualitative data to

examine these findings in more depth (Creswell, 2014).

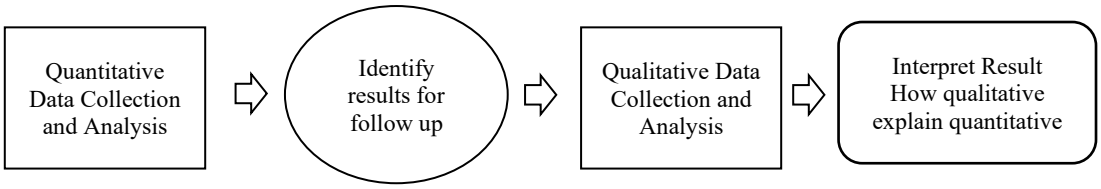


Figure 1. Explanatory Sequential Design

In the quantitative phase of the research, true experimental design was used. In this design, the independent variable is the item ordering methods (random order, increasing difficulty order, decreasing difficulty order) and the dependent variable is the students' academic achievement. The study consisted of three experimental groups (Group A, Group B, Group C) formed through random assignment. All groups took a foreign language (English) course online for 14 weeks. At the end of the training, multiple-choice tests were applied. Quantitative data were subjected to statistical analysis to examine the effect of the independent variable between the groups, and the parameters of the test forms (reliability coefficient, test difficulty, etc.) were also examined.

In the quantitative phase of the research, a semi-structured interview form was applied to understand the students' perceptions regarding the item order. The data obtained from this form was evaluated using the content analysis method (Drisko & Maschi, 2016).

**Participants**

The participants of the study are freshman university students studying at various faculties of a state university in the fall semester of the 2023-2024 academic year. Participants were informed before the study that the data would be analyzed at the group level, would be used for scientific purposes, and their personal information would be kept confidential. A 2131 freshman university students participated in the quantitative part of the study. 312 students who volunteered their opinions participated in the qualitative phase of the study. Data were collected randomly and anonymously from the participants. The distribution of the participants into groups is presented in Table 1.

Table 1. Distribution of Participants

| Faculty/ Vocational School | Group A | | Group B | | Group C | | Total | |
|---|---|---|---|---|---|---|---|---|
| | N | % | N | % | N | % | N | % |
| Faculty of Agriculture | 20 | 2.81 | 19 | 2.73 | 22 | 3,05 | 61 | 2,86 |
| Faculty of Arts and Sciences | 70 | 9.83 | 64 | 9.18 | 61 | 8,45 | 195 | 9,15 |
| Faculty of Dentistry | 18 | 2.53 | 18 | 2.58 | 16 | 2,22 | 52 | 2,44 |
| Faculty of Economics and Administrative Sciences | 48 | 6.74 | 50 | 7.18 | 51 | 7,06 | 149 | 6,99 |
| Faculty of Education | 73 | 10.25 | 77 | 11.05 | 78 | 10,8 | 228 | 10,7 |
| Faculty of Fine Arts | 17 | 2.39 | 15 | 2.15 | 15 | 2,08 | 47 | 2,21 |

| Faculty/ Vocational School | Group A | | Group B | | Group C | | Total | |
|---|---|---|---|---|---|---|---|---|
| | N | % | N | % | N | % | N | % |
| Faculty of Health Sciences | 22 | 3.09 | 22 | 3.16 | 22 | 3,05 | 66 | 3,1 |
| Faculty of Marine Sciences | 11 | 1.54 | 9 | 1.29 | 11 | 1,53 | 31 | 1,45 |
| Faculty of Medicine | 7 | 0.98 | 6 | 0.86 | 8 | 1,11 | 21 | 0,99 |
| Faculty of Music and Performing Arts | 19 | 2.67 | 15 | 2.15 | 18 | 2,49 | 52 | 2,44 |
| Faculty of Sports Sciences | 25 | 3.51 | 18 | 2.58 | 18 | 2,49 | 61 | 2,86 |
| Faculty of Theology | 25 | 3.51 | 24 | 3.44 | 24 | 3,32 | 73 | 3,43 |
| Social Sciences | 69 | 9.69 | 75 | 10.76 | 78 | 10,8 | 222 | 10,42 |
| Technical Sciences | 101 | 14.19 | 92 | 13.2 | 96 | 13,3 | 289 | 13,56 |
| Akkuş Vocational School | 9 | 1.27 | 11 | 1.58 | 8 | 1,11 | 28 | 1,31 |
| Fatsa Vocational School | 47 | 6.6 | 50 | 7.18 | 53 | 7,34 | 150 | 7,04 |
| Ikizce Vocational School | 38 | 5.34 | 36 | 5.16 | 39 | 5,4 | 113 | 5,3 |
| Mesudiye Vocational School | 9 | 1.27 | 9 | 1.29 | 10 | 1,39 | 28 | 1,31 |
| Ulubey Vocational School | 30 | 4.21 | 34 | 4.88 | 37 | 5,12 | 101 | 4,74 |
| Unye Vocational School | 54 | 7.58 | 53 | 7.6 | 57 | 7,89 | 164 | 7,7 |
| Total | 712 | 100 | 697 | 100 | 722 | 100 | 2131 | 100 |

**Application**

The study was conducted within the scope of a foreign language course that freshman university students. After 14 weeks of online training, a 20-question multiple-choice test was administered to students to measure their achievement. This test was conducted simultaneously in all faculty of the university, with a paper-and-pencil method, and under the supervision of instructors.

Three different forms consisting of the same questions were prepared for the test: random order, increasingly difficult order, and decreasingly difficult order. Students were randomly placed in classes and the distribution of the test forms was random. Each student sat at individual desks and 25 minutes were given to answer the test questions. The supervisors ensured the security of the assessment and made sure that the students were acting by the rules.

The "*Student Opinion Form*" was used to collect students' perceptions on the item order of test. This form was presented online to students participating in the test and data was collected on a voluntary basis.

**Data Collection Tool**

The academic achievement of the students was measured through a multiple choice test. The test consists of 20 items and was prepared in line with the learning objectives. The test items mainly focus on the students' ability to understand texts (O3), but also include the objectives of creating dialogues (O1) and learning new words (O2) (see Table 2).

Table 2. Learning Objectives

| Number | Learning Objectives | Number of Items | Items |
|---|---|---|---|
| O1 | Ability to create dialogues on basic topics | 4 | 6, 7, 15, 17 |
| O2 | Learning new vocabulary | 6 | 1, 3, 9, 10, 13, 16 |
| O3 | Understanding sentences, paragraphs, and texts with increasing vocabulary knowledge | 10 | 2, 4, 5, 8, 9, 11, 12, 14, 18, 19 |

Based on the cognitive levels of Bloom Taxonomy, 65% of the test items were structured at the Remembering domain and 35% at the Understanding domain. In this way, it was aimed to measure the students' foreign language skills in a balanced way (see Table 3).

Table 3. Distribution of Items According to Bloom Taxonomy

| Cognitive Domain | Number of Items | Items |
|---|---|---|
| Remembering | 13 | 1, 2, 3, 4, 5, 7, 8, 10, 13, 14, 16, 18, 20 |
| Understanding | 7 | 6, 9, 11, 12, 15, 17, 19 |

In order to ensure the content validity of the test, the distribution of test items according to the topics was arranged by taking into account the weight in the course content (see Table 4).

Table 4. Distribution of Items by Topic

| Topic | Explain | Number of Items | Items |
|---|---|---|---|
| Greetings and Introductions | Introducing yourself, subject pronouns, and possessive adjectives. | 1 | 17 |
| Exchanging Personal Information | Discussing age, country, nationality, job etc. Writing sentences to introduce others. | 2 | 4, 13 |
| Grammar | The simple present verb "to be" | 6 | 5, 8, 12, 18, 19, 20 |
| Dialogue Practice | Asking and answering about birthdays, telephone numbers etc. Creating dialogues to exchange personal information. | 2 | 6, 7 |
| Nouns | Singular and plural nouns, demonstratives ("this/that," "these/those") | 1 | 9 |
| Describing People and Things | "Have got/has got", adjectives for descriptions. | 1 | 1 |
| Possessives | Using possessive "'s" and family-related nouns. | 2 | 3, 16 |
| Describing Physical Appearances and Personalities | Adjectives for people. | 1 | 10 |
| Describing Places | "There is/There are" and adjectives for places. | 2 | 2, 15 |
| Parts of a House and | Vocabulary for parts of a house and furniture. | 1 | 14 |

| Topic | Explain | Number of Items | Items |
|---|---|---|---|
| Furniture<br>Food and Drinks | Vocabulary for food and drinks, quantifiers, uncountable and countable nouns | 1 | 11 |

A sample item from multiple choice test is given in Table 5.

Table 5. Sample of Items

| Item Number | Taxonomy | Learning Objective | Topic | Item |
|---|---|---|---|---|
| 14 | Remembering | O3 | Parts of a House and Furniture | A : Where is your car?<br>B : It's in the ..... .<br>A) restaurant<br>B) garage<br>C) balcony<br>D) home |

The qualitative data of the study were collected with an "*Student Opinion Form*". This form consist of open-ended questions. The data collection process was based entirely on voluntary participation, and 312 students filled out the form online. The open-ended questions in the form are given in Table 6.

Table 6. Open-ended Questions

| Question |
|---|
| Please state your opinions on whether the changing order of the items in the test creates an inequality. |
| Please state your opinions on whether the changing order of the items in the test affects your test performance. |

The form was edited based on the opinions of three field experts experienced in measurement and evaluation and a pilot study conducted on a group (n=24) of university students.

**Data Analysis**

The data analysis process was carried out to gain a detailed understanding of the relationship between different test forms and their effects on students' academic achievement. It was observed that the skewness and kurtosis values of the students' scores on three different test forms were between 0 and ±1 (see Table 7). These data show that the scores are quite close to a normal distribution (Tabachnick & Fidell, 2019). It was observed that the variances between the groups were homogeneous (Levene $F(3,3)=3.333$, $p=.333$, $p<3.33$). The scores were compared using single-factor analysis of variance (ANOVA) and it was examined whether the differences in scores between the students' test forms were significant.

Table 7. Skewness and Kurtosis Values

| Variable | Skewness | | Kurtosis | |
|---|---|---|---|---|
| | Statistic | Std. Error | Statistic | Std. Error |
| Form A | 0.274 | 0.092 | -0.466 | 0.183 |
| Form B | 0.298 | 0.093 | -0.459 | 0.185 |
| Form C | 0.417 | 0.091 | -0.139 | 0.182 |

The score distributions of the test forms were visualized through frequency tables and graphs. In addition, information about the parameters of the tests was obtained by calculating values such as the average number of correct answers (M), standard deviations (SD), variances ($s^2$), KR-20 test reliability ($\alpha$), test difficulties (p) and standard error of measurement (SEM). To understand whether the order of the test items worked as expected, the Spearman-Brown correlation coefficient was examined. Wilcoxon Signed Ranks Test was conducted to see whether there was a statistically significant difference between the difficulty indexes of each item in different forms. Finally, student opinions were examined through content analysis and frequency distribution. In the coding process, first the data was approached from a holistic perspective, then detailed coding was done. In the last stage, the patterns between the codes were determined using the inductive analysis technique, similar codes were combined to form themes. Themes that emerged from student comments were determined and these themes were expressed with frequency values and visualized with graphs.

## Results

### Results regarding Distribution of Students' Test Scores

Random order (Form A), increasing difficulty order (Form B), and decreasing difficulty order (Form C) forms of the test were applied to three different groups (Group A, Group B, and Group C). The frequency distribution of students' test scores in different test forms is given in Table 8.

Table 8. Frequency Distribution Table

| Score | Form A | | | | Form B | | | | Form C | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | f | cf | rf | crf | f | cf | rf | crf | f | cf | rf | crf |
| 100 | 10 | 712 | 0.014 | 1 | 5 | 698 | 0.007 | 1 | 9 | 721 | 0.012 | 1 |
| 95 | 14 | 702 | 0.02 | 0.988 | 14 | 693 | 0.02 | 0.994 | 22 | 712 | 0.03 | 0.985 |
| 90 | 27 | 688 | 0.038 | 0.968 | 6 | 679 | 0.009 | 0.974 | 12 | 690 | 0.017 | 0.955 |
| 85 | 20 | 661 | 0.028 | 0.93 | 30 | 673 | 0.043 | 0.965 | 20 | 678 | 0.028 | 0.938 |
| 80 | 30 | 641 | 0.042 | 0.902 | 32 | 643 | 0.046 | 0.922 | 35 | 658 | 0.048 | 0.91 |
| 75 | 42 | 611 | 0.059 | 0.86 | 45 | 611 | 0.065 | 0.876 | 24 | 623 | 0.033 | 0.862 |
| 70 | 44 | 569 | 0.062 | 0.801 | 41 | 566 | 0.059 | 0.811 | 46 | 599 | 0.064 | 0.829 |
| 65 | 60 | 525 | 0.084 | 0.739 | 53 | 525 | 0.076 | 0.752 | 48 | 553 | 0.066 | 0.765 |
| 60 | 66 | 465 | 0.093 | 0.655 | 69 | 472 | 0.099 | 0.676 | 69 | 505 | 0.096 | 0.699 |
| 55 | 78 | 399 | 0.11 | 0.562 | 66 | 403 | 0.095 | 0.577 | 84 | 436 | 0.116 | 0.603 |
| 50 | 74 | 321 | 0.104 | 0.452 | 83 | 337 | 0.119 | 0.482 | 87 | 352 | 0.12 | 0.487 |

| Score | Form A | | | | Form B | | | | Form C | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | f | cf | rf | crf | f | cf | rf | crf | f | cf | rf | crf |
| 45 | 56 | 247 | 0.079 | 0.348 | 72 | 254 | 0.103 | 0.363 | 86 | 265 | 0.119 | 0.367 |
| 40 | 82 | 191 | 0.115 | 0.269 | 65 | 182 | 0.093 | 0.26 | 74 | 179 | 0.102 | 0.248 |
| 35 | 44 | 109 | 0.062 | 0.154 | 61 | 117 | 0.088 | 0.167 | 49 | 105 | 0.068 | 0.146 |
| 30 | 32 | 65 | 0.045 | 0.092 | 33 | 56 | 0.047 | 0.079 | 29 | 56 | 0.04 | 0.078 |
| 25 | 19 | 33 | 0.027 | 0.047 | 13 | 23 | 0.019 | 0.032 | 14 | 27 | 0.019 | 0.038 |
| 20 | 9 | 14 | 0.013 | 0.02 | 4 | 10 | 0.006 | 0.013 | 8 | 13 | 0.011 | 0.019 |
| 15 | 4 | 5 | 0.006 | 0.007 | 4 | 6 | 0.006 | 0.007 | 4 | 5 | 0.006 | 0.008 |
| 10 | 1 | 1 | 0.001 | 0.001 | 1 | 2 | 0.001 | 0.001 | 1 | 1 | 0.001 | 0.002 |
| 5 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0.001 | 0.001 |

f: frequency, cf: cumulative frequency, rf: relative frequency, crf: cumulative relative frequency

56% of the students in group A (cf=399), 58% of the students in group B (cf=403) and 60% of the students in group C (cf=436) received scores below 60. Considering that the condition for success in the course is to have a score of 60 or above, it is seen that students who took the random order form have a higher success rate. That is, 44% of the students who solved the randomly ordered form, 42% of the students who solved the increasing difficulty ordered form, and 40% of the students who solved the decreasing difficulty ordered form met the success criterion. The histogram and boxplot graphs of the score distributions are given in Figure 2, 3 and 4.

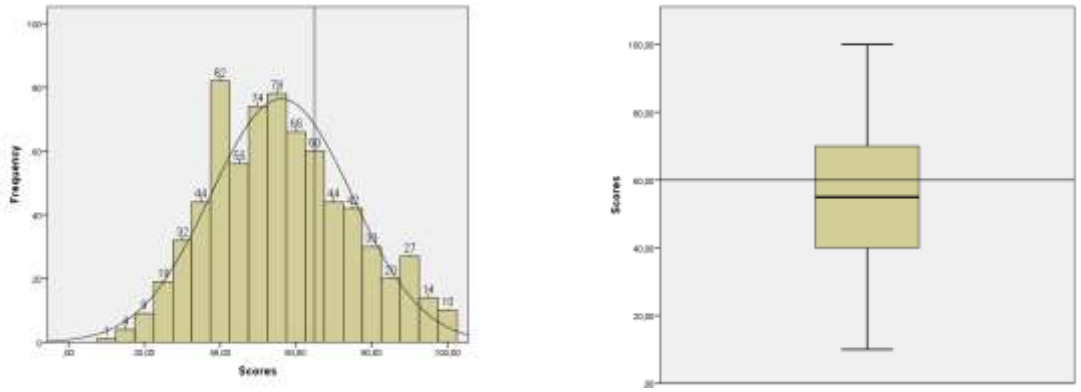


Figure 2. Histogram and Boxplot Chart for Group A

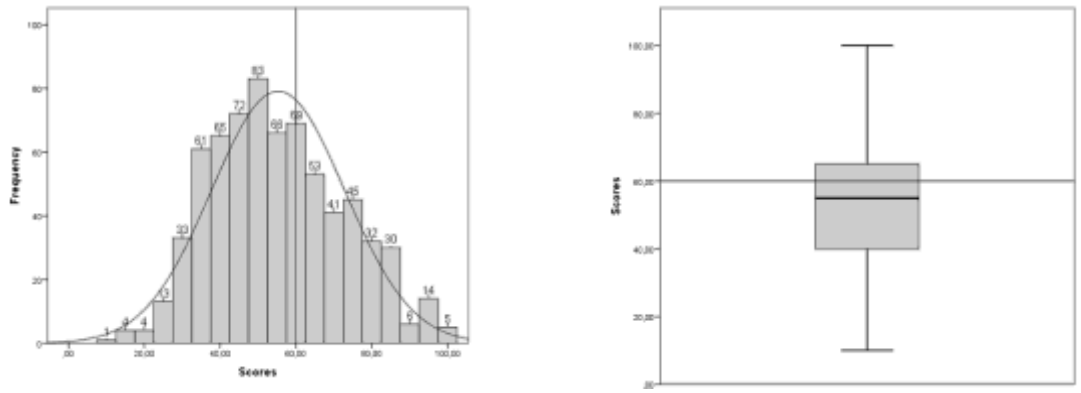


Figure 3. Histogram and Boxplot Chart for Group B

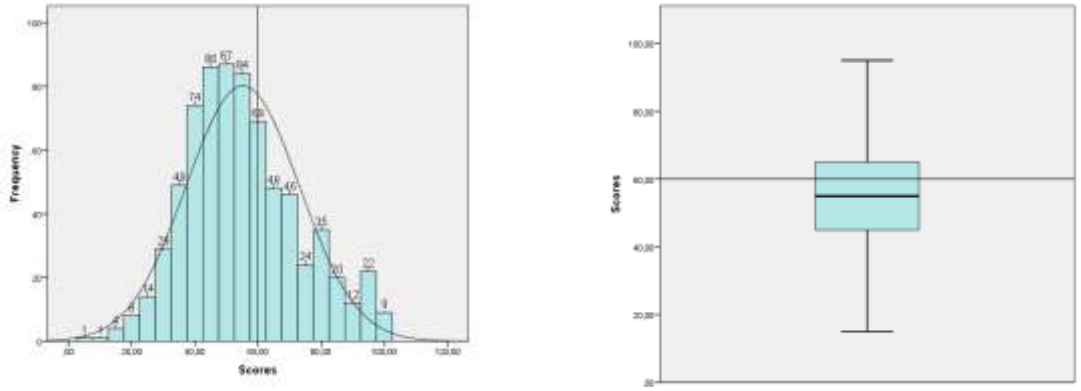


Figure 4. Histogram and Boxplot Chart for Group C

Although the scores were seen to be skewed to the right (positive) in all groups, the skewness value of the scores being less than 0.5 indicates that the distribution is almost symmetrical (Gravetter et al., 2017). According to the skewness values, the distribution of Group C is more skewed to the right, but the kurtosis value is closer to the normal distribution (See Table 2). Although the score distributions are seen to be similar between the groups, statistical analyses on the mean scores will provide a better understanding of these findings.

**Results regarding Academic Achievement**

Findings regarding academic achievement were obtained by statistically examining the mean scores of the groups and the differences between these means. The mean scores of the students are shown in Table 9.

Table 9. Mean Scores and Standard Deviations

| Group | N | M | SD |
|---|---|---|---|
| A | 712 | 56.13 | 18.6 |
| B | 697 | 55.28 | 17.58 |
| C | 722 | 55.15 | 17.97 |

The mean score of the students in Group A (M = 56.13) who solved the random order form was higher than the mean score of the students in Group B (M = 55.28) and Group C (M = 55.15). Standard deviations (SD) vary. These values show that there is a difference between the mean scores of the groups. One-way ANOVA was performed to test whether this difference was significant (see Table 10).

Table 10. ANOVA Results

| Source | Sum of Squares | df | Mean Square | F | Sig. (p) |
|---|---|---|---|---|---|
| Between Groups | 403.282 | 2 | 201.641 | 0.618 | 0.539 |
| Within Groups | 693.918.538 | 2128 | 326.090 | | |
| Total | 694.321.821 | 2130 | | | |

According to the ANOVA result, the difference between the groups was not statistically significant ($F_{(2, 2128)}$ = 0.618, p = 0.539). This indicates that the mean score differences observed between the groups may have occurred

by chance (Field, 2018). The effect size is quite small and the ordering method explains 0.058% of the total variance (η2=0.00058) (Cohen, 1998). These findings show that the test form has no significant effect on the students' test performance. In other words, the order of the items in the test did not statistically change the students' academic achievements. The test parameters will contribute to the interpretation of the findings regarding academic achievement.

**Results regarding Test Parameters**

Three test forms (Form A, Form B and Form C) arranged according to different item orders were evaluated by comparing the test parameters. The average number of correct answers (M), standard deviation (SD), variances (s2), KR-20 test reliability (α), test difficulties (p), and standard error of measurement (SEM) were considered. Statistical values of the test forms are given in Table 11.

Table 11. Test Statistics

| Form | M | SD | $s^2$ | $\alpha$ | p | SEM |
|---|---|---|---|---|---|---|
| Form A | 11.23 | 3.73 | 13.82 | 0.75 | 0.56 | 1.87 |
| Form B | 11.06 | 3.5 | 12.34 | 0.71 | 0.55 | 1.89 |
| Form C | 11.03 | 3.63 | 12.9 | 0.73 | 0.55 | 1.89 |

M: Mean, SD: Standard deviation of test, s2: variance, $\alpha$: KR-20 test reliability, p: test difficulty, SEM: Standard error of measurement

Although the average correct number in Form A is higher than in the other forms the average correct numbers in the three forms are close to each other. The variance values show a wider distribution in Form A. This can be interpreted as the random order form leading to greater score diversity among students (DeMars, 2010). Although the differences between the reliability coefficients of the forms are small, it is seen that the reliability coefficient in all three forms is above the acceptable value of 0.70. (Kline, 2015). The fficulty levels of tests are medium ($p_{FormA}$=0.56, $p_{FormB}$=0.55, $p_{FormC}$=0.55). These values show that different item orders do not affect the averages difficulty of the tests. The similarity of SEM values shows that the measurement precision of all tests is at the same. These findings can be interpreted as the item order in the tests having no considerable effect on the test parameters.

**Results regarding Item Difficulty Indexes**

The test items were arranged in increasing and decreasing items difficulty based on previous administrations. The item difficulty indexes obtained after the administration are given in Table 12.

Table 12. Item Difficulty Indexes

| Order of Items in the Test | Form A | | Form B | | Form C | |
|---|---|---|---|---|---|---|
| | Item | pj | Item | pj | Item | pj |
| 1 | Item 1 | 0.54 | item 2 | 0.89 | item 14 | 0.17 |

| Order of Items in the Test | Form A | | Form B | | Form C | |
|---|---|---|---|---|---|---|
| | Item | pj | Item | pj | Item | pj |
| 2 | Item 2 | 0.33 | item 17 | 0.83 | item 15 | 0.42 |
| 3 | Item 3 | 0.76 | item 12 | 0.98 | item 9 | 0.38 |
| 4 | Item 4 | 0.45 | item 6 | 0.54 | item 18 | 0.31 |
| 5 | Item 5 | 0.32 | item 4 | 0.48 | item 8 | 0.37 |
| 6 | Item 6 | 0.39 | item 7 | 0.77 | item 10 | 0.39 |
| 7 | Item 7 | 0.52 | item 9 | 0.87 | item 1 | 0.29 |
| 8 | Item 8 | 0.33 | item 1 | 0.67 | item 19 | 0.42 |
| 9 | Item 9 | 0.46 | item 19 | 0.54 | item 7 | 0.64 |
| 10 | Item 10 | 0.59 | item 15 | 0.59 | item 2 | 0.46 |
| 11 | Item 11 | 0.56 | item 3 | 0.51 | item 16 | 0.57 |
| 12 | Item 12 | 0.52 | item 16 | 0.66 | item 6 | 0.56 |
| 13 | Item 13 | 0.86 | item 11 | 0.42 | item 3 | 0.67 |
| 14 | Item 14 | 0.91 | item 20 | 0.3 | item 17 | 0.87 |
| 15 | Item 15 | 0.17 | item 10 | 0.4 | item 4 | 0.75 |
| 16 | Item 16 | 0.97 | item 14 | 0.36 | item 12 | 0.5 |
| 17 | Item 17 | 0.62 | item 5 | 0.32 | item 13 | 0.51 |
| 18 | Item 18 | 0.67 | item 18 | 0.37 | item 20 | 0.95 |
| 19 | Item 19 | 0.39 | item 13 | 0.39 | item 5 | 0.87 |
| 20 | Item 20 | 0.84 | item 8 | 0.18 | item 11 | 0.91 |

To see whether this ordering fulfilled its function, Spearman's rank correlation coefficient between the item ordering and difficulty index (pj) was examined (see Table 13).

Table 13. Relationship between Item Order and Difficulty Indexes

| | | Form A (pj) | Form B (pj) | Form C (pj) |
|---|---|---|---|---|
| Item | r | 0.394 | -0.860* | 0.859* |
| order | p | 0.085 | 0.000 | 0.000 |

* Correlation is significant at the 0.01 level.

In the order of increasing difficulty, a strong and negatively significant relationship was found between the item order and the difficulty indexes ($r=-0.860$, $p<0.01$). As the item number progresses (item 1→item 20), the difficulty index (pj) of the items decreases, that is, the items are perceived as increasingly more difficult. While students encountered easy items at the beginning of the test, they encountered more difficult items as the test progressed. In the order of decreasing difficulty, a strong and positive significant relationship was found between the item order and the difficulty index ($r=0.859$, $p<0.01$). As the item number progresses, the difficulty indexe (pj) of the items increases, that is, the items are perceived as increasingly more easy. In the random order form, no statistically significant relationship was found between the order of the item and the difficulty index ($r=0.396$; $p>0.05$). This finding shows that the arrangement made in line with the purpose of the study reflects the truth and

increases the reliability of the findings.

The average difficulty level of the test forms was found to be the same (see Table 11). The Wilcoxon Signed Ranks Test results also show no statistically significant difference between the difficulty indexes of each item in different forms [Form A-B (z=-0.141, p=0.89), Form A-C (z=-0.093, p=0.93) and Form B-C (z=-0.131; p=0.90)]. Similarity in average test difficulty can be misleading, so the effect of order should be examined in detail on an item-by-item basis (Lane et al., 2015). Although there is no statistical difference between the item difficulty index, the change in the item difficulty index also changes the item difficulty level (see Figure 5).

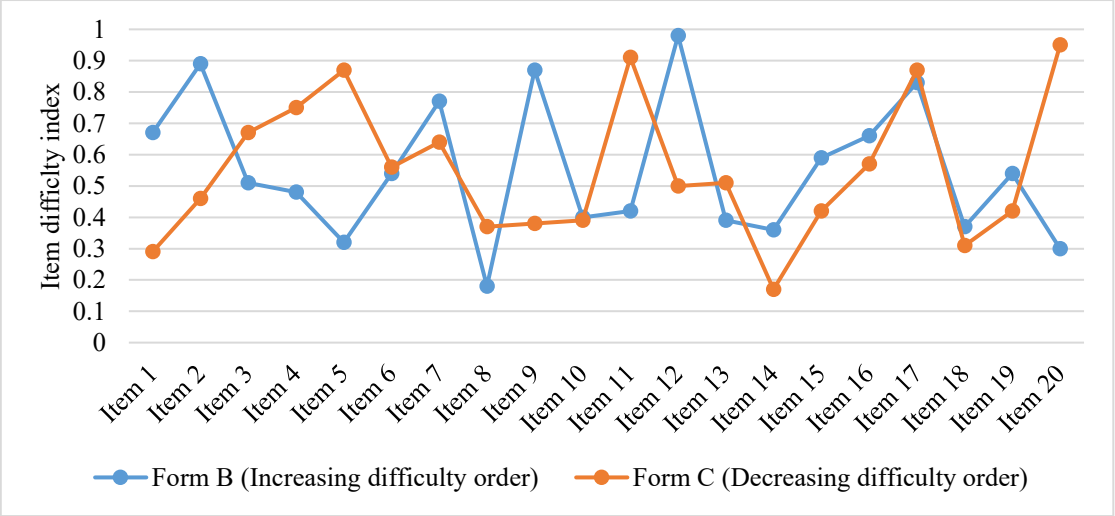


Figure 5. Changes in Item Difficulty Index

The fact that item difficulty levels change depending on the order indicates that student's responses to items may be related to the other items. For example, the difficulty index (pj) of Item 1 was 0.67 (medium) in Form B and 0.29 (difficult) in Form C. Form B is ordered from easy to difficult, Item 1 is in the 8th place, and the questions before it consist of easier items. Form C is ordered from difficult to easy, Item 1 is in the 7th place here and the items surrounding it consist of more difficult questions. Item 5 was perceived as a difficult item by the students because it was placed at the end (row 17) in Form B (easy to difficult). However, since it was placed near the end (row 18) in Form C (difficult to easy), it was perceived as an easier question this time.

These differences indicate that the order has a significant effect on the individual item's difficulty. However, since this change occurred differently across all items, there was a trade-off between the overall test difficulty averages. These findings emphasize the importance of average difficulty in test form as well as individual item difficulty. The fact that the rate of correct answers to the question changes depending on the item order is an important finding in itself. In this case, it can be said that the item difficulty levels change depending on the item order. Students' perceptions of the item order will enable a more detailed interpretation of these findings.

**Results of Students' Perceptions**

Students' perceptions were expressed under two subheadings: whether the item order created inequality and

whether it affected their test performance.

*Item Order and Inequality*

Student opinions regarding the perception of inequality consisted of 3 themes and 114 codes. These themes are "*Fairness and equality*", "*Motivation*", and "*Misperceptions*". The hierarchical code-subcode model is given in Figure 6.

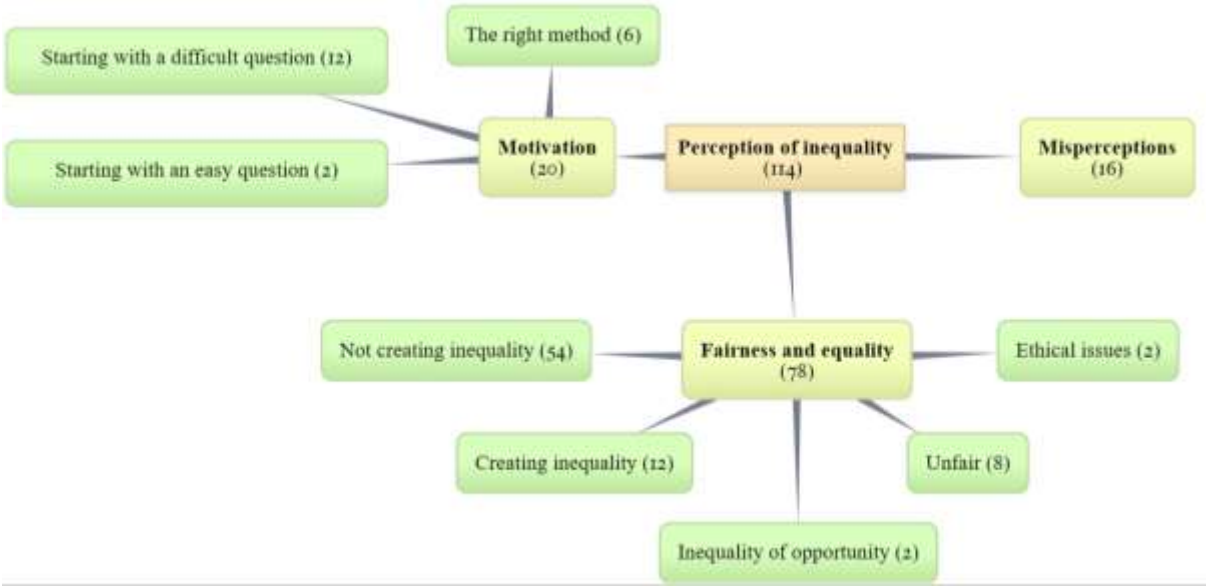


Figure 6. Hierarchical Code-subcode Model for Inequality Perception

Among the student opinions on the perception of inequality, the theme of "*Fairness and equality*" stands out (f=78). Under this theme, "*Not creating inequality*" (f=54) is the most frequently mentioned category. Students stated that the order of the items was fair with statements such as "*Everyone encountering questions of the same difficulty ensures an equal exam*." and "I *don't think it creates inequality because everyone is asked the same questions.*" Under this theme, students emphasized that the item order creates inequality (f=12) with statements such as "*Changing the question order according to groups creates inequality because each group may face a different level of difficulty*." These statements show that the students think that there is a difference in difficulty levels in the groups. Under the category of "*Unfair*" (f=8), students expressed that the item order was unfair with statements such as "*The fact that it is easy for some groups and difficult for others triggers injustice*." There are also opinions that this method is unethical (f=2) and creates inequality of opportunity in education (f=2).

Under the theme of motivation (f=20), students frequently stated that difficult questions encountered at the beginning of the test decreased their motivation (12).

> "*Students who start with difficult questions experience a disadvantage in terms of time management and motivation*"
>
> "*When faced with a difficult question at first, motivation decreases*"
>
> "*When a student sees a difficult question for the first time, their mind gets stuck on that question and the*

*possibility of doing the next easy question decreases*"
"*If the first question is difficult, motivation and desire decrease*"

There are also students who report that groups that start with easy questions are better motivated (f=2) and that although it reduces their motivation, it prevents cheating (f=6).

Under the theme of misperceptions (f=16), it was observed that students had a perception that the items changed in the test, not the order of the items. Expressions such as "*The fact that the questions are easy for some groups and difficult for others triggers injustice*" and "*Easy questions go to one group and difficult questions go to another group, and in my group, I always get parts that I haven't studied*" show the misperceptions of the students.

*Item Order and Test Performance*

Student opinions regarding performance perception were created with a total of 3 themes and 138 codings. Themes are "*Perception of performance*", "*Item order*", and "*Teaching method*". The hierarchical code-subcode model is given in Figure 7.

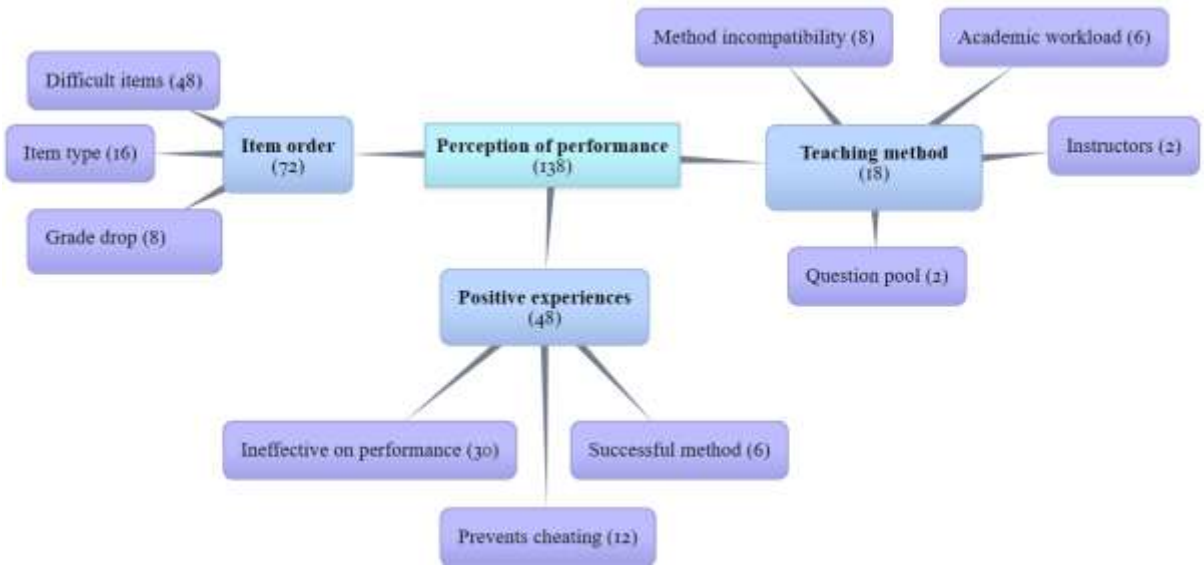


Figure 7. Hierarchical Code-subcode Model for Perception of Performance

Item order (f=72) is one of the prominent themes. In the difficult items category (f=48), students stated that the difficult questions encountered at the beginning of the test negatively affected their motivation. Statements such as "*I cannot focus on other questions, which causes failure*" and "*if the first question is difficult, my motivation decreases*" indicate the negative impact of difficult items on students' motivation. Statements such as "*Encountering difficult questions at the beginning of a test can reduce motivation and self-confidence*" and "*I feel anxious when I encounter difficult questions*" reveal the anxiety that difficult questions create in students. Similarly, some students stated that encountering easy items at the beginning of the test contributed positively to their test performance. "*Starting with easier questions positively affects exam performance by building confidence*" and, "*I think going from easy to difficult questions makes students feel more comfortable*

*psychologically*" are the opinions that show the positive effect of easy items on performance. The statement "*If a difficult question comes at the beginning, time is not enough and it will cause time loss*" shows that difficult items at the beginning affect the time management and cause students to make mistakes under pressure. In the item type category (f=16), students state that the question order negatively affects their test strategies. The statement "*Random question order renders the exam strategies developed by students ineffective*" drew attention to the uncertainty created by the question order. In addition, the statement "*It would be more orderly if the vocabulary questions were written one under the other and the fill-in-the-blank questions were written one under the other*" shows the desire to present the questions in a systematic manner. Under this theme, students also stated that item ordering caused a grade drop (f=8).

Regarding the theme of positive experiences (f=48), students stated that the item order ineffective in their performance (f=30). Some students emphasized that the application did not affect their performance by saying, "*It does not affect me. I think there was an equal environment in every way."* Moreover, expressions such as "*It prevents cheating*" and "*...it does not affect my performance, it prevents cheating*" show that it creates a fair assessment environment to prevent cheating (f=12). Some students stated that the test was successful in terms of item order and that the inclusion of the topics covered in the lessons in the exam provided a fair assessment process (f=6). Positive comments such as "*It was a very nice fluid, I can say it was very good in terms of item order*", "*The questions were good*" and "*I think it was very orderly*" reflect the views that the exam order provided the expected effect.

Under the theme of teaching method (f=18), in the category of method incompatibility (f=8), students stated that the online learning process should be reflected in the assessment process.

"*It is better to be online and it can reduce the workload*"

"*It seems ridiculous for all students to have a face-to-face test for an online course.*"

"*The questions are difficult because we do not take the course face-to-face.*"

Students expect harmony between how the course is taught and how it is assessed. Students who do not find it fair to assess online courses with face-to-face exams think that online exams will reduce their workload. This shows that students care about the assessment process being appropriate for the learning environment. Under the category of academic workload (f=6), students emphasized that they could not allocate enough time to online courses due to the intensity of the courses and that this situation limited the time allocated for other courses.

"*Our field courses are intensive, if we only spend time on this course; other courses fail, if we spend time on other courses, these courses fail*"

"*There are too many topics and details*"

These statements show that students experience an additional workload and that the intense information makes the preparation process difficult. It was also stated that under this category, it would be more beneficial to focus on important topics that may appear in the exam instead of synchronous (live) course (f=2) and that a sample question pool should be created for evaluation (f=2).

## Discussion

This study investigated the effects of item order (random order, decreasing difficulty order, and increasing difficulty order) in multiple-choice tests on online learner student's academic achievement and test parameters. The findings show that the score distributions in the groups are similar (Q1) and the item order does not affect academic achievements of students (Q2). In addition, it was observed that item ordering according to item difficulty did not have a significant effect on the test parameters (Q3). The most striking finding of the study is that although item order does not affect average test difficulty, it does affect individual item difficulty level (Q4). Also it has been observed that students perceive the beginning of the test with difficult questions negatively (Q5). These findings were discussed under two subheadings in the context of the literature.

### Score Distribution and Academic Achievement

The finding in this study that item ordering according to the difficulty index did not have a significant effect on student's academic achievement is consistent with the findings of some studies in the literature. Cobbinah (2016), Gül and Bökeoğlu (2018), Perlini et al. (1998), and Vander Schee (2013) found that random ordering, increasing difficulty ordering, and decreasing difficulty ordering of items did not have a significant effect on student test performance. Şad (2020) and Weinstein and Roediger (2012) revealed that test forms with increasing and decreasing difficulty item orders did not make a difference on student performance. Although the findings suggest that item order does not affect test performance, the results are not always consistent. There are also studies showing that students' test performance varies according to the order of item difficulty. Baffoe et al., (2024), Margaret & Victor (2017), and Soureshjani (2011) found that increasing difficulty order had a positive and significant effect on students' performance. These inconsistent findings in the literature show that changes in student performance cannot be explained solely by item difficulty order.

Plake et al., (1982) and Weinstein & Roediger, (2012) suggest that the ordering of items from easy to difficult creates an optimistic impression on students, and this perception is a noteworthy finding. The findings in this study regarding students' perceptions of item order also support this situation. It is the "Difficult items" category that stands out under the theme of "Item order", reflecting students' opinions on the effect of item order on their performance. The statements that are especially difficult items at the beginning of the test negatively affect students' motivation and point to the psychological dimension of the test. Students stated that encountering difficult questions at the beginning of the test reduces their motivation, damages their self-confidence, and increases their anxiety about the assessment process. This situation is consistent with the findings of Plake et al. (1982) and Weinstein and Roediger (2012) that the order from easy to difficult creates an optimistic impression in students. In this context, despite the quantitative findings that item order does not affect student test performance, qualitative findings show that perceptual and affective factors related to the test process should not be ignored. Pettijohn and Sacco (2007) state that students' perception of difficulty is affected by the version of the test and this may have an impact on test anxiety.

Studies on students' test perceptions and anxiety show that item order has essential effects not only on cognitive

performance but also on affective factors. Bard and Weinstein (2017) note that students' strong bias toward item order affects their test performance. The Cognitive-Attentional Model (Naveh-Benjamin et al., 1991) shows that negative perceptions affect students' test performance. When students encounter difficult items at the beginning of the test, it affects their negative thoughts and reduces their test motivation (Marsh & Martin, 2011). Akugri (2023) states that the interaction of item order and test anxiety can directly affect student performance and recommends educators to use test versions that will minimize students' test anxiety. Additionally, Chen (2012) demonstrated test anxiety as a factor affecting students' performance by interacting with item order. This finding is consistent with student opinions indicating that students' motivation may decrease and their test anxiety may increase, especially if they encounter difficult questions at the beginning of the test.

In this context, this study also draws attention to affective factors in item ordering. Hauck et al. (2017) state that students' perceptions may provide a better prediction than statistically item difficulty. It would not be correct to look for an effect solely based on item order on students' test performance. This situation shows that students' perceptual and psychological experiences are also important. Therefore, discussions on whether item order has a direct effect on academic achievement should be based not only on quantitative findings but also on taking into account students' perceptions.

**Test Parameters and Item Difficulty Index**

In this study, it was observed that the test parameters were not affected by the item order. Opara and Ogbono (2023) also stated that there was no difference in student performance and test reliability due to the order of the items. On the other hand, Gül and Bökeoğlu (2018) reported that although there was no difference in student test performance, the reliability of the test was affected. Hodson (1984) observed that although the reliability of the test was not affected, student's test performance changed. There are contradictory findings in the literature regarding the effect of ordering according to item difficulty level on test parameters.

According to Classical Test Theory (CTT), when groups are similar and the items remain the same, test parameters are expected not to change (Fan, 1998). Test parameters are expected to vary depending on the performance of the group (Crocker & Algina, 1986). In this context, the findings obtained in the study are as expected according to CTT. In the study, it was observed that different forms did not affect the students' performance and there was no effect on the parameters of the test such as average test difficulty and reliability coefficient. Şad (2020) also found that test parameters move together with students' test performance. On the other hand, Lane et al. (2015) stated that average test difficulty may be misleading and therefore items should be considered individually.

The most striking finding in this study is that while the average difficulty of the test forms does not change, the difficulty level of the item changes depending on the item order. Gül and Bökeoğlu (2018) found that although the average test difficulty was similar, item ordering led to differences in the item difficulty level. The change in the difficulty level of the items shows that the average test difficulty will always not reflect the truth (Livingston, 2011). Students who have the perception that the order of the items does not create inequality do not guess that the difficulty level of the questions may change depending on the order. "Statements such as "*Everyone being*

*presented with questions of the same difficulty ensures a fair exam*" and "*I don't think it creates inequality because as a result, everyone is asked questions of the same difficulty*" demonstrate this situation. Some students think that the order of the items in the test forms creates differences in difficulty, causing injustice and inequality of opportunity. Students' awareness of the effect of item order on difficulty level can reshape their perception of equity in the test process. The perception, especially among students, that placing difficult items at the beginning of the test reduces motivation and performance shows that the order has a psychological dimension. Therefore, the fact that item difficulty varies according to item order may eliminate students' perception that the item order is fair.

Davis (2017) addressed this issue from a different perspective and claimed that the ordering effect would be more pronounced in courses where topics progress consecutively. The finding in this study that student performance was not affected by item ordering can be considered within this framework. Baldwin et al. (1989) also argued that the effect of item order would be reduced in courses where topics were independent of each other. Indeed, Ollennu and Etsey (2015) found that students performed similarly in the English test in the increasingly difficult, decreasingly difficult, and randomly ordered test versions, but in the mathematics test, students who took the increasingly difficult ordered form performed better. This situation points to a context effect originating from the subject area. The topics in this study, which was conducted within the scope of the English course, consist of content that is independent of each other and does not require follow-up of the topic. This may have eliminated the increase in cognitive load based on the subject context. Presenting questions that require cognitive processing first allows students to gradually move on to more complex questions and improves performance by reducing cognitive load (Newman et al. 1988). Margaret and Victor (2017) support this claim by arguing that items should be arranged from simple to complex. In this context, it may be appropriate to create different test forms according to the item order in order to ensure fairness and equivalence in courses where the topics are not related to each other. This situation can also explain why the order effect is observed in courses such as mathematics (Opara & Uwah, 2017), science (Baffoe et al. 2024) and chemistry (Hodson, 1984), but not in courses such as psychology (Perlini et al., 1998), marketing principles (Vander Schee, 2013) and English (Ollennu & Etsey, 2015). It should also be noted that average test difficulty can be misleading. The fact that item difficulty indexes change depending on the order has shown that students' responses to items are related not only to the content but also to the surrounding items. In this context, instead of statistical item difficulty in test design, considering the cognitive hierarchy of items and student perceptions can offer a strategic approach to test designs. Perhaps, as students expressed under the method theme, presenting tests in the same item order in an online environment can be an important step to eliminate the discussions and negative perceptions regarding the effect of item order on test performance.

## Conclusion and Recommendations

This study examined the effects of item order (random order, decreasing difficulty order, and increasing difficulty order) on online learners' academic achievement and test parameters in multiple-choice tests. The findings show that item order does not significantly affect academic achievement, but it has important consequences in terms of psychological factors related to student perceptions. Students stated that the presence of difficult items at the

beginning of the test negatively affected their motivation, increased their anxiety, and created negative perceptions about the test process. In quantitative findings, no difference was observed in the test parameters, but it was determined that item order affected individual item difficulties.

These results reveal that not only statistical item difficulty but also student perceptions should be taken into account in exam design. As a result, considering students' affective experiences in test design will contribute to making the evaluation process more fair and equitable. Especially for students with high anxiety levels, including easy items at the beginning of the test can provide a more positive start to the exam process. Therefore, not only the statistical properties of the test but also students' perceptions and motivations should be taken into account in test design.

Considering the pedagogical structure of online learning and the expectations of students, it is noteworthy that the assessment process is compatible with the teaching method. The fact that the tests of online learners are also online can make the assessment process more efficient. In this way, the assessment method becomes more compatible with the teaching method and an assessment process that supports cognitive integrity can be created. Conducting the assessment in an online environment can eliminate discussions related to item order by providing the same item order to all students.

## Limitations

This study was conducted within a specific course and the effects of item ordering were examined only in the context of that course. Future studies should consider the effects of item ordering across different types of courses to assess the interdisciplinary validity and effectiveness of ordering strategies. In particular, it should be examined how item ordering may have different effects in courses that require cognitive processing.

The study did not comprehensively address students' differences (e.g., cognitive ability, learning strategies, or previous academic achievements). Furthermore, students' anxiety levels and affective factors at the time of the test were not taken into account. However, variables such as test anxiety may directly affect students' performance. Future studies should also consider factors such as test anxiety and motivation when evaluating the effects of item ordering. Only decreasing difficulty, increasing difficulty, and random ordering methods were used in this study. However, different item ordering strategies (e.g., ordering based on topic, ordering based on question type, or adaptive ordering that supports students' cognitive processes) could also be included in the study.

## References

Abdullahi, G. S., Vincent, P., & Akwashiki, A. G. (2020). Effect of changes in item-sequence on students academic achievement in multiple-choice test of mathematical-economics in colleges of education, *World Journal of Innovative Research (WJIR), 9*(4), 69-76. https://www.wjir.org/vol/vol-9issue-4

Akugri, F. S. (2023). Interactive effect of order of items and test anxiety on students' academic performance. *Journal of Advances in Education and Philosophy, 7*(8), 269-276.

https://doi.org/10.36348/jaep.2023.v07i08.004

Anastasi, A., & Urbina, S. (1997). Psychological testing (7th ed.). Prentice Hall/Pearson Education.

Anaya, L., Iriberri, N., Rey-Biel, P., & Zamarro, G. (2022). Understanding performance in test taking: The role of question difficulty order. *Economics of Education Review, 90*, 102293. https://doi.org/10.1016/j.econedurev.2022.102293

Bachman, L. (1990). Fundamental considerations in language testing. London: OUP.

Baffoe, J., Asamoah, D., Shahrill, M., Latif, S. N. A., Asamoah Gyimah, K., & Anane, E. (2024). 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

Baldwin, B. A., Pattison, D. D., & Toolson, R. B. (1989). Intertopical ordering effects: The case of managerial accounting. *Journal of Accounting Education, 7*(1), 83-91. https://doi.org/10.1016/0748-5751(89)90024-9

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

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

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 ınfluence student choices when writing multiple choice examinations?. C*anadian Journal for the Scholarship of Teaching and Learning, 8*(1), 11. https://doi.org/10.5206/cjsotl-rcacea.2017.1.11

Chen H. (2012). The moderating effects of item order arranged by difficulty on the relationship between test anxiety and test performance. *Creative Education, 3*(3), 328-333. https://doi.org/10.4236/ce.2012.33052

Cobbinah, A. (2016). Items' sequencing on difficulty level and students' achievement in mathematics test in Central Region of Ghana. *African Journal of Interdisciplinary Studies, 9*, 55-62. http://publications.uew.edu.gh/2015/sites/default/files/55-62%20COBBINAH.pdf

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Creswell, J. W. (2014). A concise introduction to mixed methods research. Sage Publications.

Crocker, L., & Algina.J (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement, 10*(1), 3-31. https://doi.org/10.1177/001316445001000101

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

DeMars, C. (2010). Item Response Theory. Oxford University Press.

Drisko, J. W., & Maschi, T. (2016). Content analysis. Oxford university press.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their itemperson

statistics. *Educational and Psychological Measurement, 58*(3), 357-38. https://doi.org/10.1177/001316449805800300

Field, A. (2018). Discovering statistics using IBM SPSS statistics (5th ed.). SAGE Publications.

Gravetter, F. J., Wallnau, L. B., Forzano, L. A. B., & Witnauer, J. E. (2021). Essentials of statistics for the behavioral sciences. Cengage Learning.

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

Gyamfi, A. & Yeboah, A. (2022). The changing phase of validity: The past and now. *Global Scientific Journals, 10*(6), 1016-1023. https://www.globalscientificjournal.com/journal_volume10_issue6_June_2022_edition_p4.html

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?. *Preprint from Research Square*. https://doi.org/10.21203/rs.3.rs-2882983/v1

Harlen, W., Gipps, C., Broadfoot, P., & Nuttall, D. (1992). Assessment and the improvement of education. *The curriculum journal, 3*(3), 215-230. http://dx.doi.org/10.1080/0958517920030302

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

Hilliger, I., Ruipérez-Valiente, J. A., Alexandron, G., & Gašević, D. (2022). Trustworthy remote assessments: A typology of pedagogical and technological strategies. *Journal of Computer Assisted Learning, 38*(6), 1507-1520. https://doi.org/10.1111/jcal.12755

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

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

Kline, R. B. (2015). Principles and practice of structural equation modeling (4th ed.). Guilford Press.

Lane, S., Raymond, M. R., & Haladyna, T. M. (2015). Handbook of test development (2th ed.). Routledge. https://doi.org/10.4324/9780203102961

Livingston, S. A. (2011). Item analysis. In S. M. Downing & T. M. Haladyna (Eds.), Handbook of test development (pp. 421-441). Lawrence Erlbaum Associates, Inc

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

Margaret, O. I. & Victor, U. I. (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 8*(5), 1-9. https://eajournals.org/bje/vol-5-issue-8-july-2017-special-issue/

Marsh, H. W., & Martin, A. J. (2011). Academic self-concept and academic achievement: Relations and causal ordering. *British Journal of Educational Psychology, 81*(1), 59-77. https://doi.org/10.1348/000709910X503501

Naveh-Benjamin, M. (1991). A comparison of training programs intended for different types of test-anxious students: Further support for an information-processing model. *Journal of Educational Psychology, 83*(1), 134. https://doi.org/10.1037/0022-0663.83.1.134

Newman, D. L., Kundert, D. K., Lane Jr., D. S., & Bull, K. S. (1988). Effect of varying item order on multiple-choice test scores: Importance of statistical and cognitive difficulty. *Applied Measurement in Education, 1*(1), 89–97. https://doi.org/10.1207/s15324818ame0101_8

Nigam, A., Pasricha, R., Singh, T., & Churi, P. (2021). A systematic review on AI-based proctoring systems: Past, present and future. *Education and Information Technologies, 26*(5), 6421-6445. https://doi.org/10.1007/s10639-021-10597-x

Novak, J. D., Mintzes, J. I., & Wandersee, J. H. (2005). Learning, teaching, and assessment: A human constructivist perspective. In Assessing Science Understanding (pp. 1-13). Academic Press. https://doi.org/10.1016/B978-012498365-6/50003-2

Ollennu, S. N. N., & Etsey, Y. K. A. (2015). The Impact of Item Position in Multiple-Choice Test on Student Performance at the Basic Education Certificate Examination (BECE) Level. *Universal Journal of Educational Research, 3*(10), 718-723. http://doi.org/10.13189/ujer.2015.031009

Opara, I. M., & Ogbanu, G. I. (2023). Effect of item order on the reliability of mathematics test among secondary school students in rivers state. *Journal of Advances in Education and Philosophy, 7*(11), 460-466. https://doi.org/10.36348/jaep.2023.v07i11.003

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

Perlini, A. H., Lind, D. L., & Zumbo, B. D. (1998). Context effects on examinations: The effects of time, item order and item difficulty. *Canadian Psychology / Psychologie Canadienne, 39*(4), 299–307. https://doi.org/10.1037/h0086821

Pettijohn, T. 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. http://www.tpettijohn.net/academic/research.htm

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*, 49-57. https://doi.org/10.1111/j.1745-3984.1982.tb00114.x

Reynolds, C. R., Livingston, R. B., & Willson, V. (2006). Measurement and assessment in education. Allyn & Bacon/Pearson Education.

Roediger III, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(5), 1155. http://dx.doi.org/10.1037/0278-7393.31.5.1155

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

Soureshjani, K. H. (2011). Item sequence on test performance: Easy items first?. *Language Testing in Asia, 1*(3), 46. https://doi.org/10.1186/2229-0443-1-3-46

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

Tabachnick, B. G., & Fidell, L. S. (2013). Using multivariate statistics (6th ed.), Boston: Allyn and Bacon

Taşkın, N. (2024). Cheating and prevention strategies in online assessment. In Teaching and Assessment in the Era of Education 5.0 (pp. 151-172). IGI Global. https://doi.org/10.4018/979-8-3693-3045-6.ch009

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

Walsh, W. B., & Betz, N. E. (1995). Tests and assessment (3rd ed.). Prentice-Hall, Inc.

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

Xiong, Y., & Suen, H. K. (2018). Assessment approaches in massive open online courses: Possibilities, challenges and future directions. *International Review of Education, 64*(2), 241–263. https://doi.org/10.1007/s11159-018-9710-5

Zanetti, C., & Butera, F. (2025). Detecting collective cheating culture in academic contexts. *Educational Psychology*, 1–20. https://doi.org/10.1080/01443410.2025.2449976

## Author Information

**Necati Taşkın**

https://orcid.org/0000-0001-8519-6185

Ordu University

Distance Education Application and Research Center

Türkiye

Contact e-mail: *necatitaskin@odu.edu.tr*