

Article

Exploring Question Order Effects in Multiple-Choice Assessments: Evidence from Undergraduate Education Courses

Abdulqader Alyasin *, Murielle El Hajj , Hiba Harb and Ramzi Nasser *

College of Education and Arts, Lusail University, Lusail P.O. Box 9717, Qatar; mnahas@lu.edu.qa (M.E.H.); hharb@lu.edu.qa (H.H.)

* Correspondence: aalyasin@lu.edu.qa (A.A.); rnasser@lu.edu.qa (R.N.)

Abstract

This research examines the impact of multiple-choice question (MCQ) sequencing—forward (instructional order) versus randomized—on undergraduate student performance in education courses. Grounded in cognitive perspectives on memory organization and retrieval, the study investigates whether question order influences test outcomes. Using a 2×2 mixed repeated-measures design across four courses at an Arab Gulf university, data were collected from 212 students who completed two MCQ-based assessments administered under alternating sequencing conditions. Repeated-measures Analysis of Variance (ANOVA) and Analysis of Covariance (ANCOVA) were employed to analyze performance differences while controlling for testing phase and academic achievement (GPA). Results indicated no statistically significant main effects of question order or testing phase, nor any significant interaction between sequencing and testing phase. GPA did not moderate the relationship between sequencing format and performance. However, descriptive trends suggested slightly higher performance under forward sequencing conditions, pointing to the possibility of small or context-dependent sequencing influences. Overall, the study findings indicate that undergraduate students perform similarly across forward and randomized MCQ formats under typical classroom conditions. By integrating within- and between-subject analyses and controlling for individual academic achievement, the study strengthens the methodological evidence base on MCQ sequencing. The findings provide cautious support for the use of randomized sequencing as a fairness-oriented assessment strategy, highlighting the importance of test design features and contextual factors. Future research should examine sequencing effects in high-stakes settings, assessments targeting higher-order cognitive skills, and diverse disciplinary contexts.

Keywords: MCQ sequencing; academic integrity; assessment fairness; question order; cognitive psychology; repeated-measures design; undergraduate education



Academic Editor: Jack F. Eichler

Received: 27 November 2025

Revised: 25 May 2026

Accepted: 6 June 2026

Published: 26 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Student overcrowding in classroom testing contexts often leads educators to design multiple versions of the same assessment. One common strategy involves administering identical multiple-choice question (MCQ) examinations that differ only in question order, a practice widely used to reduce academic dishonesty and improve test security (Geiger & Simons, 1994; Haasnoot et al., 2025; Pettijohn & Sacco, 2007). Recent scholarship characterizes academic dishonesty as both a moral and practical issue, given its demonstrated implications for students' future professional conduct (Eshet et al., 2023). As assessment en-

vironments expand, institutions increasingly adopt strategies such as multiple test versions and question randomization to deter dishonest behavior while maintaining fairness.

Two principal sequencing formats are typically used: a forward version that presents questions in alignment with the instructional sequence, and a randomized version that disrupts this structure and may separate conceptually related items. To prevent cheating, instructors commonly create multiple test forms by mixing up the order of MCQs across exam versions (Pettijohn & Sacco, 2007), often by randomly reordering items from an original forward-sequenced version, such that the forward format serves as the base from which randomized alternatives are derived (Haasnoot et al., 2025). The forward version reflects the order in which students encountered course content (Kagundu & Ross, 2015), offering a logical progression consistent with instructional delivery. Although rearranging exam questions has become a widely adopted approach to mitigate cheating in higher education, its cognitive implications remain insufficiently explored, and educators often administer MCQ versions without considering how such changes may influence cognitive processing or performance (Carlson & Ostrosky, 1992; Pettijohn & Sacco, 2007).

From a cognitive psychology perspective, chaining models suggest that information is encoded via associations between successive items, whereby the retrieval of one item cues the activation of the next (Lashley, 1951; Lewandowsky & Murdock, 1989). Test-takers may rely on contextual cues from preceding items (Woodford & Bancroft, 2005), and research on cued-recall practice further indicates that students frequently rely on cue-driven retrieval strategies (Badali et al., 2023). Chunking (grouping adjacent items) can further facilitate sequential retrieval from long-term memory (Cowan & Chen, 2009; Orbán et al., 2008; Perlman et al., 2010). These mechanisms support the hypothesis that instructional order consistency in testing may facilitate retrieval and reduce cognitive load.

The primary aim of this study is to determine the conditions under which maintaining consistency with instructional order influences student performance, and when randomized sequencing has a cognitively neutral effect, thereby informing evidence-based assessment design that balances academic integrity with pedagogical equity. The present study investigates whether the sequencing of MCQ items (either in a forward or randomized order) significantly influences student performance across multiple courses at the College of Education and Arts, while controlling for variables such as testing phase and cumulative grade point average (GPA).

1.1. Literature Review

1.1.1. Theoretical Foundation

Research grounded in cognitive psychology's chaining models suggests several mechanisms through which the content and order of test items may influence performance on subsequent questions. When questions are thematically related, they can provide contextual cues and priming effects that facilitate memory retrieval (Haasnoot et al., 2025). Order association further suggests that students often organize their learning and preparation according to the instructional sequence, making them more likely to perform well when examinations reflect that structure (Sue, 2009). Associative models of serial recall indicate that items encoded in sequence form structured links, whereby the retrieval of one element increases the likelihood of recalling adjacent or contextually related information, and that memory performance is shaped by the organizational structure established during learning (Lewandowsky & Murdock, 1989; Slamecka, 1985). Students may therefore benefit from increased retrieval cues when test items are presented in the same sequence as instructional delivery during the course, suggesting that forward sequencing may reduce cognitive load and enhance retrieval accuracy. Building on these theoretical assumptions, the present study proposes that:

Hypothesis 1. *Student performance on multiple-choice assessments may vary as a function of question sequence, such that students receiving forward-sequenced and randomized item orders will perform differently on MCQ assessments.*

1.1.2. MCQ Sequencing and Student Performance

Beyond content creation, the structure and sequencing of MCQ assessments have important implications for educational outcomes. Researchers have therefore extensively examined the impact of question order on student performance. This includes investigations into randomized question order, sequencing questions from easiest to hardest (or vice versa), grouping questions by topic, and maintaining the instructional sequence of the course (Gohmann & Spector, 1989; Gray, 2004; Taub & Bell, 1975).

Regarding difficulty-based ordering and related sequencing effects, early research indicated that organizing exam questions from easiest to hardest can lead to improved student performance (Leary & Dorans, 1985). A large-scale field experiment showed that students who started with easier items completed more questions and performed significantly better (Anaya et al., 2022), while analyses of PISA (Programme for International Student Assessment) data revealed that harder-first formats increased question abandonment and lowered scores (Balart & Oosterveen, 2019). Beyond performance, question order also shapes metacognitive processes: students beginning with easier items consistently overestimated their performance, while those starting with harder items underestimated it, an anchoring bias unaffected by repeated exposure (Bard & Weinstein, 2017). Applying Capaldi's (1994) Sequential Theory, R. M. Miller and Andrade (2020) found that item sequencing influenced persistence rather than performance in technology-based and online learning environments. The use of control questions as an agility mechanism for test-takers, for example, to establish a "physics mindset," i.e., problem-solving techniques, at the start of an exam, has also been proposed (Gray et al., 2002). Although this line of research focuses primarily on difficulty-based sequencing, it highlights the broader impact of item order on cognitive engagement and provides important insights for examining instructional order consistency.

With respect to instructional order consistency, which is the specific variable examined in the present study, the importance of question sequencing has been demonstrated across diverse assessment contexts. In a study of second-year medical students, LaRochelle et al. (2016) found that coherence scores, quantifying the logical sequencing of diagnostic questions, were positively correlated with both checklist scores and global performance ratings, with the association strengthening at higher levels of student performance. In MCQ contexts, early studies in business and economics reported that forward-ordered examinations were associated with higher scores than randomized versions (Taub & Bell, 1975). No performance differences were observed when MCQs were organized under topical headings, although forward formats again showed advantages over reversed sequences (B. R. Dunn, 1979). Subsequent research confirmed variability in findings depending on subject area, sample size, and test design (Marso, 1970; Pettijohn & Sacco, 2007; Şad, 2020), and similar inconsistencies across disciplines suggested that contextual and disciplinary factors may moderate the influence of question order on performance (Perlini et al., 1998; Schmitt & Scheirer, 1977; J. A. Dunn, 1976; Tal et al., 2008).

Several studies have concluded that question order does not significantly affect student performance (Bresnock et al., 1989; Davis, 2017; Sue, 2009). Gohmann and Spector (1989) recommended incorporating measures of academic ability, such as GPA, to account for individual differences. Sue (2009) found no significant order-related bias in Macroeconomics and Microeconomics examinations, attributing this to the small class sizes and closer instructor–student interaction that may reduce reliance on order-based recall strate-

gies. Similarly, Davis (2017) reported that reordering examination questions across two psychology courses did not significantly influence student performance. Norman (1954) was among the first to explore this issue, and although no significant differences across forward, reversed, and random formats were found (Hughes et al., 1974), a later study reported marginally better performance for forward-ordered versions (Balch, 1989).

In contrast, Haasnoot et al. (2025) found that students who received scrambled versions of a microeconomics test were at a disadvantage, as correct responses were more likely when an MCQ followed a conceptually related question from the same or a previous lecture. The authors interpreted these results as evidence that related questions provide contextual cues, promote order association, and facilitate memory priming, thereby reducing cognitive effort and improving recall under forward sequencing conditions. Randomization has also been recommended in large classes, as it has been associated with improvements in student grades and reductions in cheating (McLeod et al., 2003), and as test developers commonly designate two-thirds of response options as distractors (Haladyna & Downing, 1993), the sequencing of items may influence how students allocate cognitive resources across questions.

Given these mixed findings, the present study hypothesizes that:

Hypothesis 2. *Student performance may vary across testing phases due to learning progression and increased familiarity with course content, independent of question sequencing.*

Hypothesis 3. *An interaction effect between testing phase and question sequencing is anticipated, such that the influence of forward versus randomized sequencing may differ across testing phases.*

1.1.3. Academic Integrity and MCQ Test Fairness

Recent debates have focused on the cognitive validity, fairness, and design features of MCQs, with particular attention to the sequencing of questions and response options (Rodriguez & Albano, 2017; Das et al., 2021). Several studies called for a balance between academic integrity and test fairness to prevent creating unfair advantages when utilizing different MCQs sequencing (Becker et al., 2022; Gray et al., 2002; Haasnoot et al., 2025). Prior research indicates that students' engagement in academic dishonesty is often facilitated by mechanisms of moral disengagement, whereby responsibility for unethical behavior is displaced onto external factors, including evaluations perceived as unfair or poorly structured (Eshet et al., 2023; McIntire et al., 2024). Academic dishonesty is often rationalized as a response to institutional pressures, emphasizing the importance of transparent and equitable assessment practices.

Under time-constrained conditions, students' performance was significantly influenced by item ordering, as test-takers managed time differently depending on the structure and organization of questions (Becker et al., 2022), and although multiple test forms may reduce cheating, they may introduce inequity when speediness becomes a factor. Haasnoot et al. (2025) recommended maintaining a similar proportion of forward-sequenced questions across all test versions, identifying instructional order consistency as a key condition for balancing academic integrity with assessment fairness. Contrary to these concerns, Sue (2009) and Davis (2017) suggested that scrambled MCQ versions could be safely used without biasing results, and that scrambling questions appeared fair and did not produce significant performance differences.

Hypothesis 4. *In light of these considerations, academic achievement (GPA) will moderate the relationship between question sequence and performance, such that students with different achievement levels may respond differently to forward and randomized MCQ formats.*

1.1.4. Gaps

Despite extensive research, significant gaps remain, particularly within higher education. Much of the existing literature focuses on either secondary education contexts or narrowly defined subject areas such as business education or economics, limiting the generalizability of findings. While some studies suggest that sequential ordering facilitates cognitive recall and improves student performance (e.g., Balch, 1989; Haasnoot et al., 2025; Kagundu & Ross, 2015), others argue that randomization does not significantly influence outcomes (Davis, 2017; McLeod et al., 2003; Sue, 2009). Methodological constraints, including a lack of control for individual differences such as GPA, limited sample diversity, and the absence of repeated-measures designs, further complicate interpretation. Crucially, studies reporting non-significant differences are often underemphasized, leaving a gap in understanding whether MCQ sequencing truly matters for student performance or whether cognitive adaptability accounts for null effects. This is particularly significant in undergraduate education, where assessment design carries both pedagogical and ethical implications.

The study addresses this gap by empirically examining the effects of MCQ sequencing on student performance using a 2×2 mixed repeated measures design, while controlling for GPA. Conducted across diverse content areas and academic levels, this research analyzes both within- and between-subject effects of forward and randomized sequencing across multiple testing phases using ANCOVA. To address this gap and achieve the study's objectives, the following research questions are examined:

- (1) Does the sequence of MCQs (forward vs. random) affect student performance?
- (2) Are there significant differences in performance across two testing phases, and does this vary by question sequence?
- (3) Is there an interaction effect between test phase and question sequence on student performance?
- (4) Does GPA moderate the relationship between question sequence and student performance?

2. Materials and Methods

2.1. Sample

Over the course of one academic term, four different courses were taught by four different instructors. The study sample comprised $N = 212$ undergraduate students enrolled in four Education and English Language courses offered at a Gulf Arab University. These courses included Child Psychology, Teaching Methods, Sustainable Development, and Foundations of Education. Each course was taught by a different instructor to ensure instructional variability, and the student population was predominantly Arab. No demographic data beyond course enrollment was collected. The student GPA was collected through the Banner system following participant consent.

2.2. Procedure

Two tests (quizzes) were developed for each of the four courses highlighted earlier. Each test covered different material and was administered at a different time. Each quiz consisted of 10 multiple-choice questions created by the instructor to assess students' understanding of the course content. The instructors ensured the validity of the questions by carefully reviewing them for clarity and relevance. Each assessment was designed to evaluate students' understanding in alignment with Bloom's taxonomy.

For each testing phase, the quizzes were offered in two formats: one in which the questions were scrambled (referred to as the "random" format) and another in which the questions were presented in the instructional sequence used in the course (referred to as "forward"). The forward order was determined based on the instructor's planned

teaching sequence. Although both test formats included identical questions, they differed in the arrangement of items. In the second testing phase, students were assessed on newly introduced course material. A crossover design was implemented where students who received the first quiz in the forward format completed the second quiz in the random format, while those who initially received the random format completed the second quiz in the forward format.

To ensure proper randomization, the sequence of the 10 forward-order questions was randomized using an open access number generator (Calculator Soup; <https://www.calculatorsoup.com/calculators/statistics/random-number-generator.php>, accessed on 28 September 2025), which generated unique numbers within the specified range without repetition. To assign students to the forward or random groups, a random list was generated for each class by entering half the total number of enrolled students into the number generator. Students were numbered sequentially based on their course registration order. For example, in the Child Psychology class, which included 38 students, numbers from 1 to 38 were entered into the generator, and 19 numbers were randomly selected. Students corresponding to these numbers were assigned the forward quiz format, while the remaining students received the random (scrambled) format. In the second testing phase, the sequencing conditions were switched between the two groups.

During each quiz session, both versions were administered simultaneously to control for time-related variables. A pre-prepared allocation list ensured that the student received a forward or a randomly formatted quiz and was accurately implemented in the second quiz, with item order alternated accordingly. Students were seated freely in a large examination room for each class, with sufficient distance maintained between them. They were also informed that multiple versions of the test were being used to help reduce opportunities for cheating. This procedure followed a mixed repeated-measures analysis of variance (Geisser, 1961; Richardson, 2018), which allows control of order effects. In practice, this meant that half of the students received the forward-ordered quiz first and the randomized version second, while the other half experienced the reverse sequence.

Each quiz was administered as a 10 min speed–power test designed to assess students' ability to process and respond to questions efficiently. The focus was on cognitive processing rather than completion of extended tasks. Research suggests that students who complete tests more quickly often perform slightly better, although differences in outcomes are typically minimal (Lovett et al., 2017). The standardized 10 min duration across all four courses provided a consistent timeframe considered sufficient for evaluating students' cognitive understanding of the material, allowing responses to be influenced primarily by item difficulty and individual response times.

2.3. Data Analysis

Data analysis followed a quantitative approach. Both descriptive and inferential statistics were calculated and presented in tabular form. The first phase involved summarizing sociodemographic characteristics, including nationality, gender, and major, using means and standard deviations. The second phase involved paired-samples *t*-tests with independent groups and unequal variances (Xu et al., 2017). These tests compared performance across question order formats (forward and random) during the first and second testing periods. In the third stage, a mixed repeated-measures analysis of variance (ANOVA) was conducted to examine the main and interaction effects of question order across two testing occasions. The mixed design increases statistical sensitivity and power by accounting for individual variability, thereby facilitating the detection of differences that might not be observable in independent-group comparisons. Question order (forward vs. random) functioned as the between-subjects factor, while testing time (first vs. second assessment)

served as the within-subjects (repeated-measures) factor, where participants' scores across the two testing phases were compared, and individuals effectively served as their own controls (Beyer, 2021).

The mixed analysis assessed whether student performance changed over time and whether the sequencing of questions influenced assessment outcomes. This was followed by repeated-measures ANOVA and repeated-measures analysis of covariance (ANCOVA). The repeated-measures ANOVA included two independent variables: question order (forward vs. random), treated as a between-subjects factor, and testing time, treated as a within-subjects factor (Cardinal & Aitken, 2006; Roberts & Russo, 1999). This design enabled the comparison of sequencing effects and testing-period differences across groups.

In the final phase, ANCOVA was conducted with GPA included as a covariate. Including GPA allowed the analysis to control for variance in performance outcomes attributable to academic achievement, thereby isolating the effect of assessment order. Prior to conducting ANCOVA, mean differences between assignment groups were examined in terms of GPA and overall performance across the two testing phases. Because ANCOVA reduces error variance, it enhances the sensitivity of statistical tests and increases the likelihood of detecting group differences after controlling for extraneous factors (Field, 2013; G. A. Miller & Chapman, 2001). This final analysis also explored the potential influence of extraneous factors not directly measured in the study, but that may have contributed to performance variance.

3. Results

3.1. Sample Demographics

The sample comprised 212 students in education programs at Aab-Gulf University. The mean age was 24.78 years ($SD = 4.08$; $n = 172$). The sample was predominantly female, with only five male participants. This is common in education programs in the Arab context and highlights a significant gender imbalance. The majority of the participants were from the English Language Teaching program ($n = 132$), while the remaining were from the Counseling and Mental Health Program ($n = 40$). Although the study was conducted in Qatar, the majority of participants were non-Qatari students from various Arab countries, reflecting the regionally diverse nature of the student population (see Table 1 for further details).

Table 1. Study Demographics.

Nationality	<i>n</i>	%	Gender	<i>n</i>	%	Program	<i>n</i>	%
Qatari	23	10.8	Female	167	78.8	English Language Teaching	132	62.3
Non-Qatari	149	70.3	Male	5	2.4	Counseling Mental Health	40	18.9
Missing	40	18.9	Missing	40	18.9	Missing	40	18.9
Total	212	100.0	Total	212	100.0	Total	212	100.0

3.2. Data Normality

The data underwent a normality test using a Shapiro–Wilk test for scores from the two tests conducted during the first and second rounds, both for the forward and random conditions. The results of the Shapiro–Wilk test showed $W = 0.90$, $p = 0.00$, and $W = 0.93$, $p = 0.00$, respectively, indicating that the distribution was significantly deviating from normality. A nonparametric analysis was conducted using Friedman's test to compare the differences between the two testing periods. The results revealed no significant differences across the two periods, $\chi^2(df = 1) = 0.01$, $p = 0.94$. Thus, this finding should be supported by the non-significant difference between the time points of taking the test, as detailed below. It is noteworthy that repeated-measures ANOVA is comparatively resilient to

normality violations, particularly in within-subjects designs; therefore, the results can still be considered reliable and robust, and the conclusions drawn from the analysis are unlikely to be substantially affected by these assumption violations.

3.3. Comparisons Between Forward and Random Groups: Testing for Hypotheses 1 and 2

Table 2 presents the mean scores and the results of *t*-tests comparing the two test versions administered in forward and random formats across all courses. No statistically significant difference was found between the forward and random test formats, $t(df = 101) = 1.50$, $d = 0.12$, $p = 0.14$. Also, during the second testing phase, no significant difference was found between the random and forward formats, $t(df = 105) = -0.78$, $d = -0.08$, $p = 0.44$, observed power = 0.32 (see Table 2, first two rows). Overall, the results indicated no significant performance differences across courses, as students performed similarly in both assessment formats regardless of question order (see Table 2 for each course comparing forward and random, and second run random versus forward). There were near-significant differences at a *p*-value of 0.01. The forward format also had higher average scores than the random format in the Sustainable Development courses. This means that a structured progression of content may support students' comprehension and performance. However, the effect size associated with it ($\eta^2 = 0.20$) indicates a relatively small magnitude of difference. This implies that despite the forward format advantage, the practical significance of this difference is still limited. Thus, while the trend appears to be towards the forward format, the results should be treated with some caution, as the observed differences were not large enough to support strong conclusions about the effects of instructional sequencing aligned with the quiz format. The statistical power of 0.08 was considered low, and the effect sizes observed were small, indicating limited sensitivity to detect meaningful differences. Conventionally, a statistical power of 0.80 (80%) is regarded as adequate. Accordingly, Hypothesis 1 was not supported. Similarly, no statistically significant differences were observed between testing phases across all courses (see Table 3), irrespective of whether participants completed the assessments in forward or random formats. Therefore, Hypothesis 2 was also not supported.

Table 2. Means and Standard Deviations for Forward and Random Conditions Across Courses.

Source	N	Forward M (SD)	Random M (SD)	Mean Diff.	<i>t</i>	<i>df</i>	<i>p</i>	Effect Size	Observed Power
Overall	102	Forward 77.69 (20.06)	Random 74.04 (19.44)	3.65	1.50	101	0.14	0.15	0.32
Overall	106	Random 74.71 (20.64)	Forward 76.23 (18.98)	−1.52	−0.78	105	0.44	−0.08	0.12
Child Psychology	19	Forward 77.89 (18.13)	Random 65.26 (24.12)	12.63	2.29	18	0.04	0.53	0.58
Child Psychology	19	Random 71.58 (21.15)	Forward 67.89 (16.19)	3.68	0.85	18	0.41	0.20	0.13
Sustainable Development	43	Forward 77.67 (21.25)	Random 73.78 (16.72)	3.90	1.191	42	0.24	0.18	0.21
Sustainable Development	43	Random 68.81 (68.81)	Forward 76.67 (76.67)	−7.86	−2.37	41	0.02	−0.36	0.64
Teaching Methods	22	Forward 79.55 (23.40)	Random 90.00 (13.1)	−10.5	−1.93	21	0.07	−0.41	0.45
Teaching Methods	21	Random 90.95 (13.38)	Forward 93.3 (12.38)	−2.38	−0.90	20	0.38	−0.20	0.13
Foundations of Education	18	Forward 75.25 (15.53)	Random 64.44 (14.64)	10.81	1.81	17	0.09	0.43	0.40
Foundations of Education	24	Random 73.30 (16.19)	Forward 67.08 (16.74)	6.22	1.47	23	0.156	0.30	0.29

Table 3. Within-Subject Effects.

Source		Mean Square	<i>F</i>	<i>df</i>	Sig.	Partial Eta Squared
Within First and Second Period (whether Forward or Random first and vice versa)	Sphericity Assumed	288.75	0.98	1	0.32	0.002
Between-Forward and Random	Sphericity Assumed	11.18	0.19	1	0.89	0.00
Type of Questions (Forward or Random) $\times$ Period	Sphericity Assumed	661.81	2.24	1	0.14	0.01

3.4. Main and Interaction Effects: Testing Hypothesis 3

Table 3 reports that the repeated-measures ANOVA was used to analyze the effects of question order and testing time. Assumption testing showed that the requirements were satisfied: Box's M test was not significant ($M = 6.10$, $F = 2.00$, $p = 0.11$), supporting the homogeneity of covariance matrices. In addition, Levene's tests were non-significant at both measurement points ($F(1, 206) = 0.62$, $p = 0.43$; $F(1, 206) = 0.82$, $p = 0.37$), confirming the homogeneity of variance. There were no significant main effects of testing period or question order, $F(1, 206) = 0.98$, $p = 0.32$, and no significant interaction between question order and timing of assessment, $F(1, 206) = 2.24$, $p = 0.14$. These results do not support the interaction between order and time hypothesized in Hypothesis 3.

3.5. ANCOVA Results Using Grade Point Average

To examine differences in cumulative GPA for the forward and random sequencing groups, an independent-samples *t*-test was used. Assumption testing was carried out, and Levene's test was non-significant, $F(1, 210) = 0.15$, $p = 0.70$, indicating that homogeneity of variance was met. Results showed that there was no significant difference between the GPA of those who completed the random format first ($M = 2.63$, $SD = 0.83$) and those who completed the forward format ($M = 2.68$, $SD = 0.81$), $t(210) = -0.50$, $p = 0.62$, $d = -0.07$, showing a minimal effect. Additional analyses using GPA as a covariate revealed no significant interaction with the assessment phase, $F(1, 209) = 0.82$, $p = 0.37$. However, the GPA demonstrated a significant main effect, $F(1, 209) = 33.23$, $p < 0.001$, $\eta^2 = 0.137$, but the effect size was modest. In sum, these findings do not support Hypothesis 4.

4. Discussion

4.1. Synthesis

In response to Research Question 1, which explored whether the sequence of MCQs (forward vs. random) affects student performance, the findings indicate that no statistically significant main effect of question order was detected under the conditions of this study (Rejecting Hypothesis 1). However, descriptive inspection of the data revealed a consistent directional tendency toward higher performance in the forward-order condition across most comparisons, student groups, and course contexts. Although this tendency did not reach statistical significance within the present modelling framework, it suggests the possibility of small or context-dependent sequencing effects that warrant cautious interpretation.

Overall, the results align with previous studies (B. R. Dunn, 1979; McLeod et al., 2003; Sue, 2009), which reported negligible differences in performance across sequencing formats. Importantly, the present study extends this body of work by examining sequencing effects under stricter methodological controls, including a mixed repeated-measures design and the inclusion of multiple course contexts. These findings therefore suggest a general robustness of student performance across different MCQ sequencing conditions. At the same time, descriptive inspection of the data revealed a consistent tendency toward higher

mean scores under forward sequencing in several course contexts, particularly in Child Psychology, Sustainable Development, and Foundations of Education (see Table 2 above). Although these differences did not reach statistical significance, this directional pattern may tentatively suggest subtle cognitive mechanisms related to thematic progression, primacy-related retrieval cues, or increased cognitive effort when encountering less structured item orders. However, these interpretations remain speculative and should be treated cautiously. It is also possible that characteristics of the assessments themselves reduced the sensitivity to detect very small sequencing effects. These include relatively short test length, moderate item difficulty, and the use of instructor-developed MCQs that were not formally piloted.

Research Question 2 examined whether student performance differed across two testing periods and whether such differences varied by question sequence. The repeated-measures ANOVA did not reveal any main effect of time or a significant interaction between sequencing and assessment timing (Rejecting Hypothesis 2). These findings suggest that performance remained relatively stable across administrations, regardless of whether students received forward or randomized versions first. This pattern is consistent with earlier research reporting similar outcome stability when sequencing conditions were counterbalanced (Gohmann & Spector, 1989; Kagundu & Ross, 2015).

Research Question 3 explored the interaction between test timing and question sequence. Although no statistically significant interaction was detected (Rejecting Hypothesis 3), some course-specific variations emerged, particularly in Child Psychology and Teaching Methods. These patterns resonate with prior suggestions that sequencing under time constraints may introduce fairness concerns in certain contexts (Becker et al., 2022; Gray et al., 2002). Although the current findings do not strongly support such effects, they emphasize the significance of contextual factors, such as course design, item characteristics, and student familiarity with the content, in influencing potential sequencing effects.

Research Question 4 examined whether GPA moderated the relationship between question sequence and performance. The ANCOVA results indicated that prior academic achievement did not significantly interact with sequencing conditions (Rejecting Hypothesis 4). This contributes to the literature by addressing long-standing calls for better control of individual differences (Gohmann & Spector, 1989) and suggests that randomized MCQ formats are unlikely to disadvantage students across achievement levels.

4.2. Limitations and Future Directions

Several limitations should be considered when interpreting these findings. First, the study relied primarily on overall response accuracy, which represents a relatively coarse indicator of potential item-order effects. Sequencing influences may operate at the level of cognitive processing rather than final correctness, affecting response time, perceived difficulty, or confidence. Future research should therefore utilize more fine-grained measures, including response-time analysis, confidence ratings, or process-tracing methods such as eye-tracking.

Second, the study may have limited sensitivity to detect small or item-specific sequencing effects. Analyses were conducted at the aggregated test level, and the binary scoring approach did not capture partial knowledge or strategic elimination of distractors. Previous research suggests that common scoring methods, such as number-right scoring, may obscure distinctions between full understanding and guessing (Bereby-Meyer et al., 2002; Bradbard et al., 2004; Choppin, 1988; Frary, 1980).

Although no statistically significant sequencing effect was identified, descriptive forward-order tendencies suggest that subtle effects may not have been fully detectable. The relatively short assessments, moderate item difficulty, and use of instructor-developed MCQs that were not formally piloted may have reduced score variability and overall test

sensitivity. Future studies should therefore employ longer assessments, validated items, and experimental manipulation of sequencing structures.

Finally, the study findings are based on specific course contexts and assessment formats, which may limit generalizability to other disciplines or high-stakes settings. Emerging technological approaches, including AI-supported item generation and adaptive testing systems, offer promising avenues to systematically examine sequencing effects and enhance assessment validity (Kıyak et al., 2024; Shwe et al., 2024; Yigit & Amasyali, 2023).

5. Conclusions

This study contributes to the literature on MCQ sequencing by examining the effects of question order in undergraduate assessment contexts using a mixed repeated-measures design and controlling for prior academic achievement. Within the scope of accuracy-based outcomes, no statistically significant sequencing effect was detected at the test level. These findings are broadly consistent with earlier research suggesting minimal performance differences between forward and randomized MCQ formats (McLeod et al., 2003; Sue, 2009), while extending the evidence base through stronger methodological controls and the inclusion of multiple course contexts. At the same time, descriptive trends indicating slightly higher performance in forward-order conditions suggest that small or context-dependent sequencing influences cannot be entirely ruled out. The findings, therefore, provide cautious support for the use of randomized MCQ sequencing as a fairness-oriented assessment practice under typical classroom conditions and highlight the importance of test design features and contextual factors in shaping potential sequencing effects. Future research should build on these results by employing item-level analyses, longer and psychometrically validated assessments, and more fine-grained measures of student performance and cognitive processing. Such work will help clarify the conditions under which sequencing may influence assessment outcomes and further inform evidence-based test design in higher education.

Author Contributions: Conceptualization, A.A. and R.N.; methodology, A.A. and R.N.; formal analysis, R.N.; data curation, A.A., R.N., M.E.H. and H.H.; writing—original draft preparation, A.A., R.N., M.E.H. and H.H.; writing—review and editing, A.A., R.N., M.E.H. and H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Lusail University, Qatar.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Lusail University (protocol code LU/IRB/25-25/01 and date 25 September 2025).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are not publicly available due to privacy and ethical restrictions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Anaya, L., Iriberry, N., Rey-Biel, P., & Zamarro, G. (2022). Understanding performance in test taking: The role of question difficulty order. *Economics of Education Review*, 90, 102293. [CrossRef]
- Badali, S., Rawson, K. A., & Dunlosky, J. (2023). How do students regulate their use of multiple-choice practice tests? *Educational Psychology Review*, 35, 43. [CrossRef] [PubMed]
- Balart, P., & Oosterveen, M. (2019). Females show more sustained performance during test-taking than males. *Nature Communications*, 10, 3798. [CrossRef] [PubMed]
- Balch, W. R. (1989). Item order affects performance on multiple-choice exams. *Teaching of Psychology*, 16(2), 75–77. [CrossRef] [PubMed]

- 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(2), 362–371. [CrossRef] [PubMed]
- Becker, B., van Rijn, P., Molenaar, D., & Debeer, D. (2022). Item order and speededness: Implications for test fairness in higher educational high-stakes testing. *Assessment & Evaluation in Higher Education*, 47(7), 1030–1042. [CrossRef]
- Bereby-Meyer, Y., Meyer, J., & Flascher, O. M. (2002). Prospect theory analysis of guessing in multiple-choice tests. *Journal of Behavioral Decision Making*, 15(4), 313–327. [CrossRef]
- Beyer, A. (2021). Introduction to *t*-tests. In *Introduction to statistics for psychology*. Maricopa Open Digital Press. Available online: <https://open.maricopa.edu/psy230mm/chapter/chapter-11-introduction-to-t-tests/> (accessed on 6 October 2025).
- Bradbard, D. A., Parker, D. F., & Stone, G. L. (2004). An alternate multiple-choice scoring procedure in a macroeconomics course. *Decision Sciences Journal of Innovative Education*, 2, 11–26. [CrossRef]
- Bresnock, A. E., Graves, P. E., & White, N. (1989). Multiple-choice testing: Question and response position. *The Journal of Economic Education*, 20(3), 239–245. [CrossRef]
- Capaldi, E. J. (1994). The sequential view: From rapidly fading stimulus traces to the organization of memory and the abstract concept of number. *Psychonomic Bulletin & Review*, 1(2), 156–181. [CrossRef] [PubMed]
- Cardinal, R. N., & Aitken, M. R. (2006). *ANOVA for the behavioral sciences researcher*. Lawrence Erlbaum Associates.
- 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. [CrossRef]
- Choppin, B. H. (1988). *The validity of multiple-choice tests*. OECD/CERI.
- Cowan, N., & Chen, Z. (2009). How chunks form in long-term memory and affect short-term memory limits. In M. Page, & A. Thorn (Eds.), *Interactions between short-term and long-term memory in the verbal domain* (pp. 86–107). Psychology Press.
- Das, B., Majumder, M., Phadikar, S., & Sekh, A. A. (2021). Automatic question generation and answer assessment: A survey. *Research and Practice in Technology Enhanced Learning*, 16, 5. [CrossRef]
- Davis, D. B. (2017). Exam question sequencing effects and context cues. *Teaching of Psychology*, 44(4), 356–360. [CrossRef]
- Dunn, B. R. (1979). Effects of specific cueing on multiple-choice test performance. *Psychological Reports*, 44(3), 1051–1054. [CrossRef]
- Dunn, J. A. (1976). The effect of study habits, personality, and order of presentation on success in an open-book objective examination. *Teaching of Psychology*, 3(1), 33–34. [CrossRef] [PubMed]
- Eshet, Y., Dickman, N., & Ben Zion, Y. (2023). Academic integrity in the HyFlex learning environment. *Heliyon*, 9, e13301. [CrossRef] [PubMed]
- Field, A. P. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). Sage.
- Frary, R. B. (1980). The effect of misinformation, partial information, and guessing on expected multiple-choice test item scores. *Applied Psychological Measurement*. [CrossRef]
- Geiger, M. A., & Simons, K. A. (1994). Intertopical sequencing of multiple-choice questions: Effect on exam performance and testing time. *Journal of Education for Business*, 70(2), 87–90. [CrossRef]
- Geisser, S. (1961). The Latin square as a repeated measurements design. In *Proceedings of the fourth Berkeley symposium* (Vol. 4, pp. 241–250). University of California Press.
- Gohmann, S. F., & Spector, L. C. (1989). Test scrambling and student performance. *The Journal of Economic Education*, 20(3), 235–238. [CrossRef]
- Gray, K. (2004). *The effect of question order on student responses to multiple-choice physics questions* [Doctoral dissertation, Kansas State University].
- Gray, K., Rebello, N., & Zollman, D. (2002, August 7–8). *The effect of question order on responses to multiple-choice questions*. 2002 Physics Education Research Conference, Boise, ID, USA.
- Haasnoot, C. W., van Krevel, C., & de Vaal, A. (2025). Ordered or scrambled: How the forward sequencing of multiple-choice questions affects test item scores. *Assessment & Evaluation in Higher Education*, 51, 247–262. [CrossRef]
- Haladyna, T., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*, 53(4), 999–1010. [CrossRef]
- Hughes, G. E., Prytula, R. E., & Schnelle, J. F. (1974). Relationship between a forward learning set and three orders of multiple-choice examinations. *Psychological Reports*, 34(3), 727–733. [CrossRef]
- Kagundu, P., & Ross, G. (2015). The impact of question order on multiple-choice exams on student performance in an unconventional introductory economics course. *Journal for Economic Educators*, 15, 19–36.
- Kiyak, Y. S., Coşkun, Ö., Budakoğlu, İ. İ., & Canan, U. (2024). ChatGPT for generating multiple-choice questions: Evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam. *European Journal of Clinical Pharmacology*, 80(5), 729–735. [CrossRef] [PubMed]
- LaRochelle, J., Durning, S. J., Boulet, J. R., van der Vleuten, C., van Merriënboer, J., & Donkers, J. (2016). Beyond standard checklist assessment: Question sequence may impact student performance. *Perspectives on Medical Education*, 5(2), 95–102. [CrossRef] [PubMed]

- Lashley, K. (1951). The problem of serial order in behavior. In L. Jeffress (Ed.), *Cerebral mechanisms in behavior* (pp. 112–136). Wiley.
- 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. [\[CrossRef\]](#)
- Lewandowsky, S., & Murdock, B. B., Jr. (1989). Memory for serial order. *Psychological Review*, 96(1), 25–57. [\[CrossRef\]](#)
- Lovett, B. J., Lewandowski, L. J., & Potts, H. E. (2017). Test-taking speed: Predictors and implications. *Journal of Psychoeducational Assessment*, 35(4), 351–360. [\[CrossRef\]](#)
- Marso, R. N. (1970). Test item arrangement, testing time, and performance. *Journal of Educational Measurement*, 7(2), 113–118. [\[CrossRef\]](#)
- McIntire, A., Calvert, I., & Ashcraft, J. (2024). Pressure to plagiarize and the choice to cheat: Toward a pragmatic reframing of the ethics of academic integrity. *Education Sciences*, 14(3), 244. [\[CrossRef\]](#)
- McLeod, I., Zhang, Y., & Yu, H. (2003). Multiple-choice randomization. *Journal of Statistics Education*, 11(1–7). [\[CrossRef\]](#)
- Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology*, 110(1), 40–48. [\[CrossRef\]](#) [\[PubMed\]](#)
- Miller, R. M., & Andrade, M. S. (2020). The effects of test question order on task persistence. *Research & Practice in Assessment*, 15, 44–51.
- Norman, R. D. (1954). The effects of a forward retention set on an objective achievement test presented forwards or backwards. *Educational and Psychological Measurement*, 14(3), 487–498. [\[CrossRef\]](#)
- Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences of the United States of America*, 105(7), 2745–2750. [\[CrossRef\]](#) [\[PubMed\]](#)
- 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*, 39(4), 299–307. [\[CrossRef\]](#)
- Perlman, A., Pothos, E. M., Edwards, D. J., & Tzelgov, J. (2010). Task-relevant chunking in sequence learning. *Journal of Experimental Psychology: Human Perception and Performance*, 36(3), 649–661. [\[CrossRef\]](#) [\[PubMed\]](#)
- 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, 142–149.
- Richardson, J. T. (2018). The use of Latin-square designs in educational and psychological research. *Educational Research Review*, 24, 84–97. [\[CrossRef\]](#)
- Roberts, M. J., & Russo, R. (1999). *A student's guide to analysis of variance*. Routledge.
- Rodriguez, M. C., & Albano, A. D. (2017). *The college instructor's guide to writing test items: Measuring student learning*. Routledge.
- Schmitt, J. C., & Scheirer, C. J. (1977). The effect of item order on objective tests. *Teaching of Psychology*, 4(3), 144–145. [\[CrossRef\]](#) [\[PubMed\]](#)
- Shwe, L. L., Matayong, S., & Witosurapot, S. (2024). The unified difficulty ranking mechanism for automatic multiple-choice question generation in digital storytelling domain. *Education and Information Technologies*, 29, 20317–20350. [\[CrossRef\]](#)
- Slamecka, N. J. (1985). Ebbinghaus: Some associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11(3), 414–435. [\[CrossRef\]](#)
- Sue, D. L. (2009). The effect of scrambling test questions on student performance in a small class setting. *Journal for Economic Educators*, 9, 32–41.
- 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. [\[CrossRef\]](#)
- Tal, I. R., Akers, K. G., & Hodge, G. K. (2008). Effect of paper color and question order on exam performance. *Teaching of Psychology*, 35(1), 26–28. [\[CrossRef\]](#)
- Taub, A. J., & Bell, E. B. (1975). A bias in scores on multiple-form exams. *Journal of Economic Education*, 7(1), 58–59. [\[CrossRef\]](#)
- Woodford, K., & Bancroft, P. (2005). Multiple-choice questions not considered harmful. In *Proceedings of the 7th australasian conference on computing education* (pp. 109–116). ACM.
- Xu, M., Fralick, D., Zheng, J. Z., Wang, B., & Feng, C. (2017). The differences and similarities between two-sample *t*-test and paired *t*-test. *Shanghai Archives of Psychiatry*, 29(3), 184–188. [\[CrossRef\]](#) [\[PubMed\]](#)
- Yigit, G., & Amasyali, M. F. (2023). Enhancing multiple-choice question answering through sequential fine-tuning and curriculum learning strategies. *Knowledge and Information Systems*, 65(12), 5025–5042. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.