

The Effect of Response Order on Item Difficulty, Item Discrimination, Distractor Functioning, and Guessing: A Comparison of Listening, Reading, and Grammar Items

Leila Amini, Alireza Ahmadi

Department of Foreign Languages and Linguistics, School of Literature and Humanities, Shiraz University, Shiraz, Iran, Email: leila.amini2010@gmail.com

Professor, Department of Foreign Languages and Linguistics, School of Literature and Humanities, Shiraz University, Shiraz, Iran, Email: arahmadi@shirazu.ac.ir

Abstract

As a means of curbing academic dishonesty in high-stakes multiple-choice examinations, various versions of a given test are generated by manipulating the order of responses, thereby impeding students' ability to cheat. The current quantitative study employed a between-subjects experimental design to examine the impact of response order on item difficulty, item discrimination, distractor functioning, and guessing across different language skills and components (listening comprehension, grammar, and reading comprehension). An adapted retired version of a TOEFL-PBT test was used to generate four different versions of the test by changing the response order. Participants were 72 EFL learners who voluntarily took part in the study. They randomly received one of the four test versions. The results indicated significant effects of response order on item difficulty (for 10% of the items), mal-functioning and non-functioning options. However, the effect was not significant for item discrimination, well-functioning options, and guessing. Some traces, though non-significant, were also found for the primacy effect. Finally, the effect of response order did not significantly change across language skills. The results imply that the scores coming from the same test with different response orders are not necessarily comparable, thereby compromising the validity of the decisions made and the test consequences.

Keywords: distractor functioning, guessing, item difficulty, item discrimination, response order

1. Introduction

Multiple-choice (MC) test is one of the most successful, long-lasting and effective forms of measurement (Gierl et al., 2017) which is widely used in educational assessments (Butler, 2018; Gierl et al., 2017; Hagenmüller, 2020). Many high-stakes and large-scale tests benefit from this format (Baghaei & Amrahi, 2011; Hagenmüller, 2020; Lions, Monsalve et al., 2022). For instance, many important international tests such as IELTS, TOEFL, and Duolingo employ this format along with the other formats to measure test takers' proficiency. As for a high-stakes test like the Iranian National University Entrance Exam which is aimed at screening candidates for admission to university, the whole test is conducted in MC format. This test format is so popular and considered an economical form of educational measurement (Gierl et al., 2017) as a result of many advantages it offers such as being efficient to administer and easy to score objectively. Moreover, it can cover a wide range of content in a single event (Butler, 2018; Gierl et al., 2017). Besides, it can directly assess many aspects of knowledge and skills in various disciplines and content areas; for example, it can be used to measure the ability to interpret data, draw inferences, make judgment, understand concepts and principles, complete statements, and apply information (Gierl et al., 2017).

However, MC format has also been criticized for failing to measure complex mental abilities like writing and mathematical problem-solving abilities (Haladyna et al., 2002), being limited to the recognition of factual knowledge, exposing examinees to incorrect information (Butler, 2018), and being sensitive to guessing, pattern guessing, cheating, and the number of response options (Baghaei & Amrahi, 2011; McKenna, 2019). In addition, writing good test items is time-consuming (Clifton & Schriener, 2010) and developing distractors is a formidable task for test designers (Gierl et al., 2017; Haladyna & Rodriguez, 2013). In Haladyna and Rodriguez's (2013) proposed guidelines for developing MC items, "writing the options" has the highest number of guidelines indicating the complexity of developing distractors.

Considering the fact that the results of MC tests in high-stakes contexts can be strikingly consequential, either academically or professionally, for test-takers (Haladyna & Downing, 2004), the highest standards should be followed in constructing these tests to prevent the effect of construct-irrelevant variance (Lions, Monsalve et al., 2022). One of the factors that may have unanticipated influences on MC item characteristics and produce construct-irrelevant variance is the response order (e.g., Attali & Bar-Hillel, 2003; Carnegie, 2017; Hangenmuler, 2020; Hohensinn & Baghaei, 2017; Holzknecht et al., 2021; Kiat et al., 2018; Lions, Dartnell et al., 2022). Response order is the order in which response options in an MC item are presented to test-takers which may have a significant effect on their choice selection (Krosnick et al., 1987).

Response order may be rearranged in different versions of an MC test in order to prevent cheating (Hagenmüller, 2020; Hohensinn & Baghaei, 2017; Wang, 2019). Because response options in MC items are labeled A, B, C, and D (or 1, 2, 3, and 4) and items are usually in a fixed order, even if the examinee does not understand the questions, they can simply copy answers from the answer

sheets or computer screens of other test takers sitting near them. Thus, in order to control cheating the examiner might need to produce multiple versions of the same test. This includes developing various items of similar difficulty tapping the same content area. But this could be a very challenging and time-consuming strategy. A more practical and effective strategy is to develop more than one version of the same test by simply changing the response order. This practice makes cheating much harder for test takers because item keys are different across different versions of the same test (Wang, 2019).

While previous studies have explored the effect of response order on different aspects such as guessing (e.g., Bar-Hillel et al., 2005), item difficulty and item discrimination (e.g., Brown, 1997) the results have been conflicting and inconclusive. Different response order effects have been observed: a) correct response location is not a significant source of item difficulty (Marcus, 1963; Sonnleitner et al., 2016), b) response order in an MC test can have significant but unpredictable effects on item difficulty (Cizek, 1994), c) questions with middle correct answers are easier and less discriminating than questions with extreme correct answers, a phenomenon known as middle bias (attraction to central options) or edge aversion (avoidance of first and last options) (Attali & Bar-Hillel, 2003), d) there is a primacy effect with an A answer aversion (Tellinghuisen & Sulikowski, 2008), and e) there is a primacy effect; that is, items with answers in the first position are easier than later positions (Hohensinn & Baghaei, 2017; Holzknecht et al., 2021; Lions, Dartnell et al., 2022). In addition, there is a paucity of studies on the effect of response order on distractor functioning (choice distribution). Furthermore, most of the studies have been basically limited to majors other than second language testing. To date, there have been few studies (Hohensinn & Baghaei, 2017; Holzknecht et al., 2021; Lions, Dartnell et al., 2022; Sonnleitner et al., 2016) examining the effect of response order in MC tests in this field. Moreover, none of the previous studies in L2 assessment have dealt with the effect of response order across different language skills. It is likely that response order effect varies across different language skills (Holzknecht et al., 2021). Therefore, the effect of response order on different aspects of test performance is an important issue requiring further investigation (Lions, Monsalve et al., 2022).

The purpose of the present study is two-fold. First, the study is aimed at evaluating the effect of response order in MC items on item characteristics including item difficulty, item discrimination, distractor functioning and guessing. Second, in response to the need for further research on response order across different skills (Holzknecht et al., 2021), the study intends to examine if response order affects such item characteristics differently across different language skills and components (i.e., listening, reading and grammar). The study seeks the answer to the following research questions:

- 1) What is the effect of response order on item difficulty, item discrimination, distractor functioning and guessing in MC items?
- 2) Does the response order impact upon item characteristics (as specified in Q1) differently across different language skills and components (i.e., listening, reading and grammar)?

2. Literature review

2.1. Order Effects in Everyday Life

Researchers have found that when people are presented with several options from which to select, the order of the options in relation to each other can have an effect on the final choice. Studies have also indicated that positional tendency in choosing between options depends on the type of decision (Holzknecht et al., 2021). For example, Christenfeld (1995) found that people when presented with identical options, often prefer middle options in grocery shopping or in toilet selection. Hence, it is advantageous for them to select from a middle of a grocery shelf because products tend to be replaced more often, as a result fresher but disadvantageous to select central toilets because they tend to be used more often, as a result dirtier and more often out of paper. In the same study, Christenfeld reported that the last options are more favorite when people decide a route on a map or when select through a maze. In another study, Bruin (2005) showed that performers in song contests and skating contests are evaluated more favorably when they present later. In these contests, performers present sequentially and each judge gives each of them a numerical evaluation. The winner is the performer who receives the highest total number of points. It was found that the last few performers receive higher points from judges. In some cases, top and bottom options are more popular than choices in the middle. For instance, when travelers want to book hotels online (Ert & Fleischer, 2016) or when people want to order food from menus in restaurants (Dayan & Bar-Hillel, 2011), they prefer the top and bottom options.

Some studies have found primacy effect, that is, the tendency to choose the first option or object which is presented. Winke and Lim (2015), for instance, found that raters showed an apparent left-to-right bias in the use of rubric categories for scoring students' writing performances. In other words, they focused longer on the rubric presented towards the left, due to a primacy effect. Similarly, Ballard (2017) reported that if raters want to skip the reading of the criterion, the skipping part was likely to be the criterion which was placed on the right rather than the left one. In other words, raters would consider the criteria on the left more important than the right one. These studies indicate that sequence of presentation matters in various decision contexts in general. Such effects can also be expected in MC items, especially when the test taker is not entirely sure about the correct answer to the question.

2.2. Response Order in MC Items

Various studies have investigated the effect of response order on guessing (Attali & Bar-Hillel, 2003) and item characteristics like item difficulty (Attali & Bar-Hillel, 2003; Cizek, 1994; Fagely, 1987; Hagemüller, 2020; Hohensinn & Baghaei, 2017; Holzknecht et al., 2021; Marcus, 1963; Mcnamara & Weitzman, 1945; Tellinghuisen & Sulikowski, 2008) and item discrimination (Attali & Bar-Hillel, 2003). Such studies have produced mixed results.

Some studies have revealed that response order has no significant effect on item difficulty (e.g., Marcus, 1963; Taylor, 2005) test performance (e.g., Golda, 2011; Jessel & Sulins, 1975) and test reliability (e.g., Jessel & Sulins, 1975). On the other hand, some other studies have reported that response order can have a significant effect on item difficulty. Some of these studies (e.g., Cizek, 1994; Shin et al., 2020) while reporting a significant effect failed to show any predictable patterns for this effect; that is, while the effect was present, no clear pattern was found for this effect. However, others reported a clear pattern for this effect. For instance, Attali and Bar-Hillel (2003) found that 80% of undergraduate psychology students had a tendency to choose one of the middle positions—B and C—as the correct response option. Therefore, the authors concluded that questions with middle correct answers were easier, less discriminating and more vulnerable to guessing than questions with extreme correct answers, a phenomenon known as middle bias or edge aversion. Similar findings were reported by Devore et al. (2016) based on a collection of 243084 responses to MC questions by physics students. However, the study by Tellinghuisen and Sulikowski (2008) partially supported this finding as test takers appeared to be avoiding A but not D.

In contrast to the above-mentioned studies, there are some studies (e.g., Bresnock et al., 1989; Fagely, 1987; Hagenmüller, 2020; Hohensinn & Baghaei, 2017; Kiat et al., 2018; Mcnamara & Weitzman, 1945; Schroeder et al., 2012) that have reported primacy effect; that is, items with answers in the first position are easier than later positions. The assumption is that presenting the correct answer first allows processing it before considering any misleading alternative (Bresnock et al., 1989) and would hinder attractive distractors from causing interference in correct answer memory retrieval (Kiat et al., 2018). For instance, Holzkecht et al. (2021) found that four-option MC items with key in fourth positions are significantly more difficult than items with key in first positions. For this study, around 6000 test takers took a 200-item four-option MC listening test and the results showed that answers in the second, third and fourth positions were 0.27, 0.29 and 0.48 logits more difficult than answers in the first position.

Similarly, Lions, Dartnell et al. (2022) conducted a study on a Chilean national high-stakes standardized test, responded by 195,715 examinees. The results showed that although the effect of option position was small, there existed a clear and systematic pattern (i.e., a five-option item was slightly easier when the correct response was A rather than E and when the most attractive distractor was after and far away from the correct response).

2.3. The Research Gap and Rationale for the Present Study

The current study was set to fill in the gap on response order effect in second language testing as explained below.

- 1) Many of the previous studies on response order have been conducted in majors other than second language testing like navy flight (Mcnamara & Weitzman, 1945), psychology (Marcus, 1963), television trivia and learning skills (Fagely, 1987), medicine (Cizek, 1994), scholastic

abilities (Attali & Bar-Hillel, 2003), chemistry (Tellinghuisen & Sulikowski, 2008), and intelligence (Hagenmüller, 2020). The focus of the current study, however, is on English language testing.

- 2) While there have been some recent response order studies (Hohensinn & Baghaei, 2017; Holzknecht et al., 2021; Lions, Dartnell et al., 2022; Sonnleitner et al., 2016) in the field of second language testing, none have dealt with the effect of response order across different language skills. It is likely that response order effect varies across different skills. The contradictory findings reported by the previous studies may also be relevant to the skill tested. The current study fills this gap by comparing the effect of response order across listening, grammar, and reading.
- 3) Most of the previous studies were limited to item difficulty. More insights could be provided if the effect of response order on other item characteristics such as item discrimination, distractor functioning, and guessing is investigated. This is done in the current study.
- 4) Many of the previous studies included only one or two versions of the same test with different response orders (e.g., Hohensinn & Baghaei, 2017; Hagenmüller, 2020; Lions, Dartnell et al., 2022; Marcus, 1963). In this study, however, four versions of the same test are employed.

3. Method

This study is a quantitative study employing a between-subjects experimental design; in this design, participants are randomly assigned to separate groups and then compared based on their performance in relation to a specific factor in order to determine the effect of the specified factor on their performance (Ary et al., 2018). In the current study the participants were randomly assigned to four groups each receiving a different version of a MC test created by changing the response order. The design was aimed at exploring the effect of response order on test performance.

3.1. Participants

The participants of this study included 72 (28 male and 44 female) Teaching English as a Foreign Language (TEFL) and English Language and Literature university undergraduates who ranged in age between 20 to 35 (with the mean age of 23). They were selected from one of the major universities in Iran based on their availability and willingness to participate in the study.

3.2. Instruments

An adapted retired version of TOEFL Paper-Based Test (Longman Complete Course for the TOEFL Test, 2001) was used. Three sections of the test were used for the purpose of this study: listening, reading and grammar. Each section was made up of 20 MC items. A form was added to the beginning of the test to collect demographic information about the participants. To measure

students' level of certainty in answering each item, a confidence scale was added to each item. The scale asked the participants to indicate their level of confidence in answering each item by selecting an appropriate option showing the confidence level. Options included not confident at all, a little confident, moderately confident, and highly confident.

In order to investigate the impact of response order, four versions of the TOEFL test were created. The item order was the same in four versions and the only difference was in the response order or position of options. Version A was the original retired version of the TOEFL test; in Version B, the position of options A and B, and the position of options C and D were changed; in Version C, the position of options A and C, and options B and D were changed; finally, in Version D, the position of options A and D, and options B and C were changed (Table 1). This way, four versions of the same test with the same item order and different response orders were developed.

Table 1
Different Test Versions

	Version A	Version B	Version C	Version D
	A	B	C	D
	B	A	D	C
	C	D	A	B
	D	C	B	A
Response order		A↔B	A↔C	A↔D
rearrangements		C↔D	B↔D	B↔C

3.3. Data Collection Procedure

To collect the data, the participants were randomly assigned to one of the four test versions. They were asked to fill in the demographic survey before the test. Furthermore, they were instructed to show their level of confidence in answering each item by selecting one of the appropriate options. They completed the test in 50 minutes with the researchers being present in the testing session to insure everything would be ok.

3.4. Data Analysis

Data analysis was performed by comparing four versions of the test to investigate whether changing the response order has any effect on item difficulty, item discrimination, distractor functioning, and guessing. In addition, the three sections of the test were compared with each other to explore which section was more significantly affected by changing the response order.

The four item characteristics were operationally defined as follows. Item difficulty was defined as the proportion of wrong answers to the total number of answers to an item (Carr, 2011; Haladyna & Rodriguez, 2013). Item discrimination was defined as the proportion of correct responses in the high vs low group (Carr, 2011; Haladyna & Rodriguez, 2013). One-Way ANOVA was employed to see whether the differences in item difficulty and item discrimination were

statistically significant or not.

Distractor functioning is checked by examining the frequency with which each option is selected by the participants. A distractor functions well when at least than 5% of examinees select it. Non-functioning distractors are options which are selected by < 5% of the examinees. If too many examinees select a distractor (e.g., more examinees select a particular distractor than the correct option), the distractor is said to be malfunctioning (Burud et al., 2019). As for the effect of response order on distractor functioning, analysis was done to check if the correct choice position had any effect on distractor functioning. To measure guessing, as mentioned earlier, the participants were asked to indicate for each item how confident they were while answering that item or whether they had simply guessed at the right answer with no information. There were four options for each item to show the participants' confidence level; namely, not confident at all, a little confident, moderately confident, and highly confident.

Finally, the three test sections were compared with each other based on the above-discussed criteria (item difficulty, item discrimination, distractor functioning, and guessing) through one-way ANOVA to investigate whether the effect of response order varied across language skills and components.

4. Results

4.1. The Effect of Response Order on Item Characteristics

4.1.1. The Effect of Response Order on Item Difficulty

Item difficulty refers to how difficult or challenging a particular item is for participants to answer, and it can be measured by computing the percentage of participants who answer the item incorrectly. The difficulty ranged from 0 to 0.67 with the mean of 0.19 in version A; 0 to 0.89, and a mean of 0.36 in version B; 0 to 0.72, with a mean of 0.20 in version C; and 0 to 0.72, with a mean of 0.23 in version D. Then, through One-Way ANOVA, item difficulty for each item was compared across four versions of the test. The results showed that item difficulty for 54 items did not significantly change based on response order effect. In other words, for 90% of the items, response order had no significant effect on item difficulty. Only, for 10% of the items, (6 out of 60 items), item difficulty changed significantly based on response order. These six items included a listening comprehension item ($F = 4.329$, $p = 0.007$, effect size = 0.19), three grammar items ($F = 6.597$, $p = 0.001$, effect size = .29; $F = 3.390$, $p = 0.023$, effect size = 0.15; $F = 2.784$, $p = 0.047$, effect size = 0.12), and two reading comprehension items ($F = 6.810$, $p = 0.000$, effect size = 0.30; $F = 2.873$, $p = 0.043$, effect size = 0.34). The depicted for all these six items showed large effect sizes based on Cohen's (1988) recommended values (0.01 = small effect, 0.06 = medium effect and 0.14 = large effect).

Then, in order to check for any possible patterns for the effect of correct answer position on

item difficulty, all items with option A as the correct answer were put in one group and similarly all the items with the correct answer being one of the options B, C, or D were respectively put in groups 2 to 4. Then, the mean of item difficulty was compared across these four groups. The results are depicted in Tables 2 and 3.

Table 2
Item Difficulty Based on the Correct Answer Position

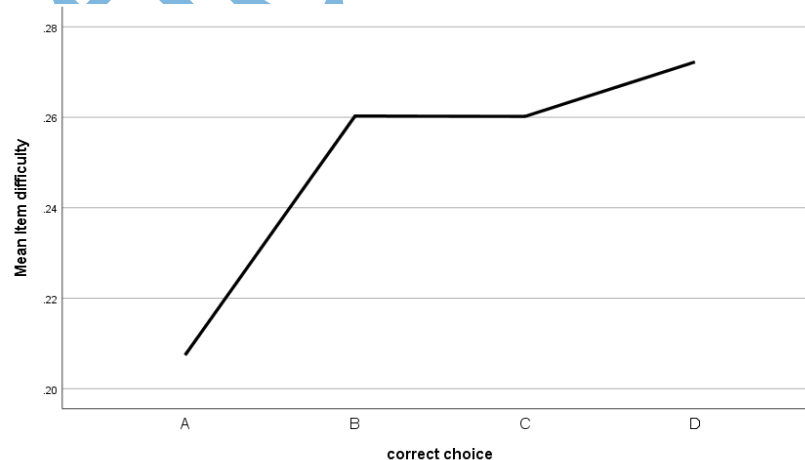
Descriptives					Lower Bound	Upper Bound		
A	60	.2074	.17975	.02321	.1610	.2538	.00	.83
B	60	.2603	.18423	.02378	.2127	.3079	.00	.72
C	60	.2602	.19501	.02518	.2098	.3106	.00	.89
D	60	.2722	.17766	.02294	.2263	.3181	.00	.72
Total	240	.2500	.18484	.01193	.2265	.2735	.00	.89

Table 3
ANOVA Results for Item Difficulty Across Groups

	Sum of Squares	DF	Mean Square	F	Sig.
Between Groups	.151	3	.050	1.483	.220
Within Groups	8.015	236	.034		
Total	8.166	239			

Table 2 indicates a pattern that on average, items with option A as the correct answer are the easiest ones and items with option D as the correct answer are the most difficult ones. This pattern can also be observed in Figure 1. However, the results of one-way ANOVA (Table 3) reveal that the depicted differences are not statistically significant.

Figure 1
Mean Item Difficulty Across Four Groups



4.1.2. The Effect of Response Order on Item Discrimination

Item discrimination refers to the ability of a test item to differentiate between high-

performing and low-performing test takers. The item discrimination ranged from -0.1 to 0.8, with the mean of 0.20 in version A; 0 to 0.6 and a mean of 0.28 in version B; -0.2 to 0.7 with a mean of 0.21 in version C, and finally, -0.1 to 0.6 with a mean of 0.24 in version D.

Then, the means of item discrimination were compared among four groups based on the correct choice position. All items with option A as the correct answer were put in one group and similarly all the items with the correct answer being one of the options B, C, or D were respectively put in groups 2 to 4. The results are shown in Table 4 and Table 5.

Table 4

Item Discrimination Based on the Correct Answer Position

	N	Mean	Std. Deviation	Std. Error	Lower Bound		Minimum	Maximum
					Upper Bound			
A	60	.2083	.18160	.02345	.1614	.2552	-.20	.70
B	60	.2317	.17514	.02261	.1864	.2769	-.20	.70
C	60	.2217	.15414	.01990	.1818	.2615	-.10	.60
D	60	.2883	.18784	.02425	.2398	.3369	-.10	.80
Total	240	.2375	.17670	.01141	.2150	.2600	-.20	.80

Table 5

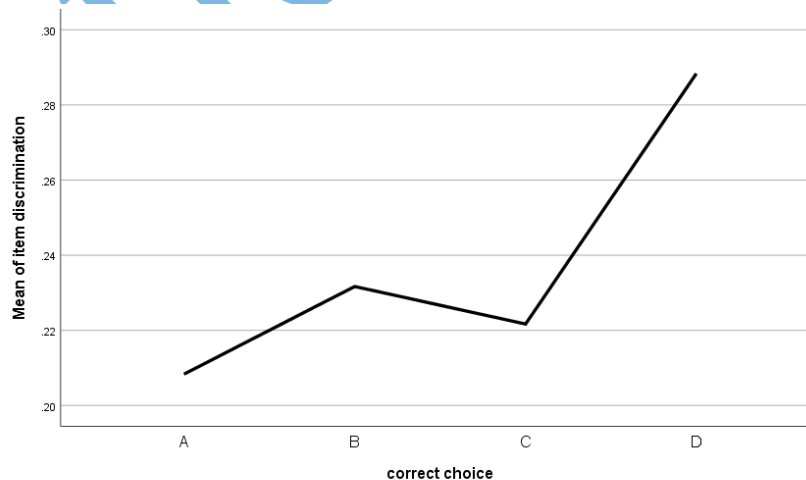
ANOVA Results for the Item Discrimination Across Groups

	Sum of Squares	DF	Mean Square	F	Sig.
Between Groups	.223	3	.074	2.425	.066
Within Groups	7.239	236	.031		
Total	7.463	239			

Table 4 and figure 2 depict that items with option A as the correct answer are the least and items with option D as the correct answer are the most discriminating items and items with option B as the correct answer are more discriminating than items with option C as the correct option. However, Table 5 reveals that the depicted differences are not statistically significant.

Figure 2

Mean Item Discrimination Across Groups



4.1.3. The Effect of Response Order on Distractor Functioning

Distractor functioning refers to the frequency with which options are selected by test takers. In order to assess the impact of response order on distractor functioning, first the distractor functioning for each option of each item was determined. Then, the results were compared across the four versions. It was found that changing response order had a negligible effect on the frequency of well-functioning and non-functioning options. Regarding mal-functioning options, Version B had 4.5% (11 out of 240) mal-functioning options, Version C had 0.41% (1 out of 240) mal-functioning option, and Version D had 1.25 % (3 out of 240) mal-functioning options. No mal-functioning option was observed in Version A as the base version of the test.

In order to check if the position of the correct choice had any effect on distractor functioning, analysis was done by comparing the frequency of well-functioning, mal-functioning and non-functioning options across the four groups based on the correct choice position. The results are shown in Table 6 and Figure 3.

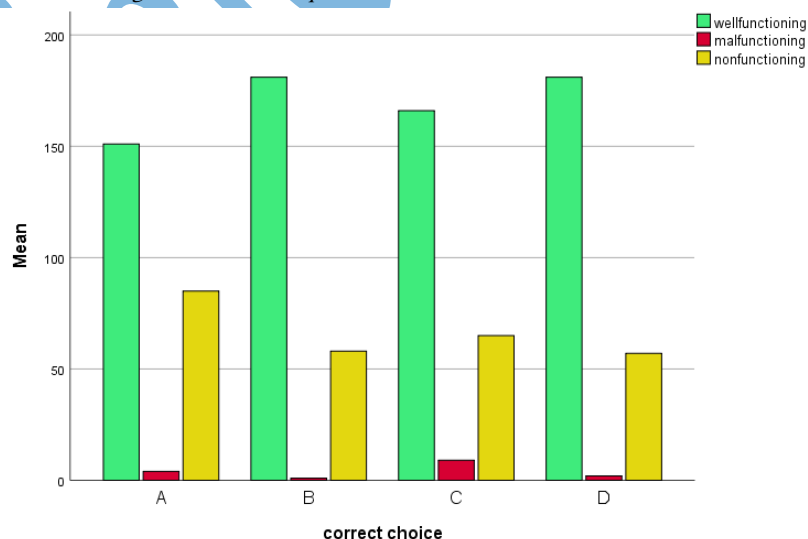
Table 6

Distractor Functioning Based on the Correct Answer Position

Correct Choice	Well-functioning options	Mal-functioning options	Non-functioning options
A	151(62.91%)	4(1.6%)	85(35.41%)
B	181(75.41%)	1(0.41%)	58(24.16%)
C	166(69.16%)	9(3.75%)	65(27.08%)
D	181(75.41%)	2(0.83%)	57(23.75%)

Figure 3

Distractor Functioning Across the Groups



Chi-square test was used to check for the significance of differences depicted in Table 6 and Figure 3. The findings indicated non-significant differences for well-functioning options (Chi-square=3.408, df=3, p= 0.333), significant differences for mal-functioning options (Chi-

square=30.528, $df=3$, $p= 0.000$), and barely significant differences for non-functioning options (Chi-square=7.649, $df=3$, $p= 0.054$).

4.1.4. The Effect of Response Order on Confidence Level (Guessing)

Wild guessing in MC items refers to selecting an answer without having the required knowledge about the correct answer. The results for the level of confidence (guessing) in answering items are depicted in Tables 7 and 8. Table 7 and also Figure 4 depict a pattern in which participants are more confident when the correct answer is listed first (option A) compared to when it is listed last (option D). But the results of ANOVA (Table 8) indicate that the differences were not significant.

Table 7

Confidence Level Based on the Correct Answer Position

	N	Mean	Std. Deviation	Std. Error	Lower Bound		Upper Bound	Minimum	Maximum
A	60	3.3056	.45646	.05893	3.1876	3.4235		1.94	3.94
B	60	3.2074	.45536	.05879	3.0898	3.3250		2.17	3.94
C	60	3.2343	.45779	.05910	3.1160	3.3525		2.11	3.94
D	60	3.2000	.45205	.05836	3.0832	3.3168		2.17	3.94
Total	240	3.2368	.45448	.02934	3.1790	3.2946		1.94	3.94

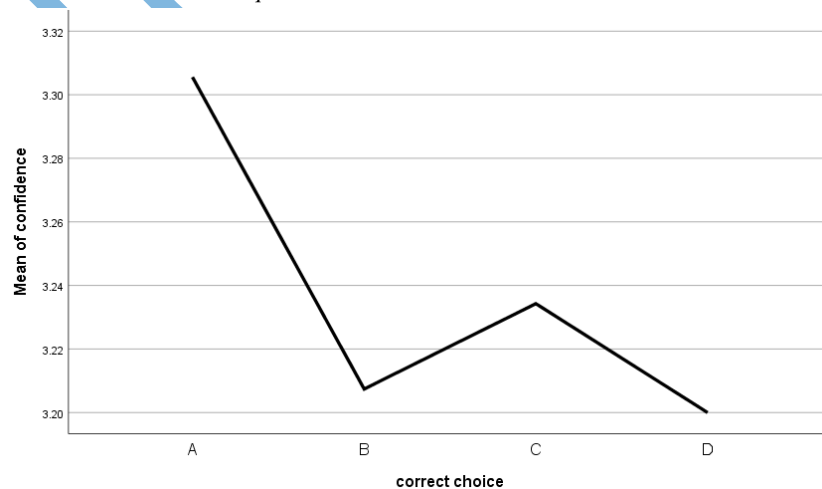
Table 8

ANOVA Results for Confidence Level Across Groups

	Sum of Squares	Df	Mean Square	F	Sig.
Between Groups	.417	3	.139	.670	.571
Within Groups	48.949	236	.207		
Total	49.366	239			

Figure 4

Confidence Level Across the Groups



4.2. The Effect of Response Order on Item Difficulty and Discrimination Across Language Skills and Components

The results of response order effect on item difficulty and discrimination across different language skills and components mostly echoed those of general findings presented in the previous sections; that is, items with the correct answer as A turned out to be the easiest and least discriminating across different language skills and components though the depicted differences through one-way ANOVA were not statistically significant (item difficulty: listening, $F = 0.311$, $p=0.817$; reading, $F=0.400$, $P=0.753$; grammar, $F=2.285$, $P=0.086$; item discrimination: listening, $F = 0.626$, $p=0.600$; reading, $F=2.089$, $P=0.109$; grammar, $F=0.925$, $P=0.433$). Due to space limitations, statistical tables are not reported here.

4.3. The Effect of Response Order on Distractor Functioning Across Language Skills and Components

Table 9 below depicts the findings for distractor functioning on the listening section. It indicates some differences between different test versions with the most noticeable one being between version A versus others. To see whether the differences depicted in different test versions were significant or not, a series of Chi Square tests was employed. The findings indicated non-significant differences for well-functioning options (Chi-square=0.272, $df=3$, $p= 0.965$), significant differences for mal-functioning options (Chi-square=30.528, $df=3$, $p= 0.000$), and non-significant differences for non-functioning options (Chi-square=1.409, $df=3$, $p= 0.704$).

Table 9
Distractor Functioning in the Listening Section

Correct choice	Well-functioning options	Mal-functioning options	Non-functioning options
A	53(22.08%)	0(0%)	27(11.25%)
B	57(23.75%)	4(1.66%)	19(7.91%)
C	52(21.66%)	4(1.66%)	24(10%)
D	55(22.91%)	2(0.83%)	23(9.5%)

Table 10 below reveals a similar pattern of differences between version A and others in the grammar section. Chi Square tests indicated non-significant differences for well-functioning options (Chi-square=1.507, $df=3$, $p= 0.681$), significant differences for mal-functioning options (Chi-square=14.284, $df=3$, $p= 0.000$), and non-significant differences for non-functioning options (Chi-square=4.219, $df=3$, $p= 0.239$).

Table 10
Distractor Functioning in the Grammar Section

Correct choice	Well-functioning items	Mal-functioning items	Non-functioning items
A	45(18.75%)	0(0%)	35(14.58%)
B	57(23.75%)	2(0.83%)	21(8.75%)
C	53(22.08%)	2(0.83%)	25(10.41%)
D	54(22.5%)	2(0.83%)	24(10%)

Finally, the effect of response order was assessed across four test versions in reading comprehension section. Chi Square tests indicated non-significant differences for well-functioning options (Chi-square=3.219, df=3, $p=0.359$), significant differences for mal-functioning options (Chi-square=23.548, df=3, $p=0.000$), and non-significant differences for non-functioning options (Chi-square=5.179, df=3, $p=0.159$).

Table 11*Distractor Functioning in the Reading Section*

Correct choice	Well-functioning items	Mal-functioning items	Non-functioning items
A	49(20.41%)	8(3.33%)	23(9.58%)
B	62(25.83%)	0(0%)	18(7.5%)
C	58(24.16%)	6(2.5%)	16(6.66%)
D	68(28.33%)	2(0.83%)	10(4.16%)

4.4. The Effect of Response Order on Confidence Level (Guessing) Across Language Skills and Components

Tables 12, to 14 below depict descriptive statistics for the effect of response order on guessing across listening, grammar and reading respectively. A series of ANOVA revealed that none of the differences were significant across listening ($F=0.291$, $P=0.832$), grammar ($F=0.206$, $P=0.892$) and reading ($F=0.428$, $P=0.734$).

Table 12*Confidence Level for Different Listening Test Versions*

	N	Mean	Std. Deviation	Std. Error	Lower Bound		Minimum	Maximum
					Upper Bound			
A	20	3.3611	.53848	.12041	3.1091	3.6131	1.94	3.94
B	20	3.2833	.50013	.11183	3.0493	3.5174	2.22	3.94
C	20	3.2611	.46589	.10418	3.0431	3.4792	2.28	3.78
D	20	3.2194	.46428	.10382	3.0022	3.4367	2.17	3.89
Total	80	3.2813	.48644	.05439	3.1730	3.3895	1.94	3.94

Table 13*Confidence Level for Different Grammar Test Versions*

	N	Mean	Std. Deviation	Std. Error	Lower Bound		Minimum	Maximum
					Upper Bound			
A	20	3.4556	.36031	.08057	3.2869	3.6242	2.61	3.94
B	20	3.3639	.38426	.08592	3.1841	3.5437	2.44	3.89
C	20	3.4333	.45083	.10081	3.2223	3.6443	2.28	3.94
D	20	3.4028	.36313	.08120	3.2328	3.5727	2.72	3.94
Total	80	3.4139	.38540	.04309	3.3281	3.4997	2.28	3.94

Table 14*Confidence Level for Different Reading Test Versions*

	N	Mean	Std. Deviation	Std. Error	Lower Bound		Minimum	Maximum
					Upper Bound			
A	20	3.1000	.39473	.08827	2.9153	3.2847	2.39	3.78
B	20	2.9750	.39674	.08871	2.7893	3.1607	2.17	3.61
C	20	3.0083	.36643	.08194	2.8368	3.1798	2.11	3.56
D	20	2.9778	.43759	.09785	2.7730	3.1826	2.17	3.72
Total	80	3.0153	.39531	.04420	2.9273	3.1033	2.11	3.78

5. Discussion

This study was set to investigate the effect of response order in MC items on item characteristics and to explore whether this effect may vary across language skills and components. Concerning the first research question, it was found that the effect of response order on item difficulty was not significant for 90% of the items. This finding is in congruent with studies that reported response order as an insignificant source of item difficulty (e.g., Golda, 2011; Jessel & Sulins, 1975; Marcus, 1963; McLeod et al., 2003; Sonnleitner et al., 2016; Taylor, 2005). On the other hand, item difficulty was changed significantly based on response order for 10% of the items. This is in line with studies that suggested response order is a significant source of item difficulty (e.g., Hagenmüller, 2020; Hohensinn & Baghaei, 2017; Holzkecht et al., 2021; Kiat et al., 2018; Lions, Dartnell et al., 2022). The fact that response order significantly affected only a fraction of items can show the complexity of the response order effect as a multi-faceted factor. It could be argued that response order operates in tandem with other factors such as item content and individual cognitive styles of test takers to create an effect on item difficulty. The mixed findings in the literature can also pinpoint this complexity.

Another explanation could be the proficiency level of the participants. While proficiency was not measured in this study as a variable, the participants were third and fourth-year undergraduates majoring in English and had practiced English for years even before coming to university. So, their proficiency could be estimated to be high-intermediate to advanced. This could be supported by test difficulty values. The mean of test difficulty was 0.19, 0.36, 0.20, and 0.23 for versions A, B, C, and D respectively, indicating that most of the items were not challenging for them. It could be argued that, the effect of response order is more likely to emerge in a test with a diverse range of item difficulties; however, when a test is easy, test takers will not have a challenging job finding the correct answer regardless of its position. In fact, in such a test, the process of finding the correct answer does not require much of comparing and analyzing different options to come up with the correct answer.

Following the relevant literature (e.g., Hagenmüller, 2020; Hohensinn & Baghaei, 2017; Holzkecht et al., 2021; Kiat et al., 2018; Lions, Dartnell et al., 2022), more detailed analyses of the items and options were conducted in the present study to investigate for the presence of any specific

patterns of answering. Although the results did not significantly support primacy effect, a pattern was found indicating that on average, items with option A as the correct answer were the easiest ones and items with option D as the correct answer were the most difficult ones. Items with option B and C as the correct answer did not function differently. Therefore, some traces of the primacy effect (Hohensinn & Baghaei, 2017; Holzkecht et al., 2021; Lions, Dartnell et al., 2022) could be claimed to exist. As the study had employed a small sample size, lack of significant differences could be due to the sample size as a key factor in statistical analyses. Replication studies with larger samples can provide more insights in this regard.

Regarding the effect of response order on item discrimination, no significant effect was found. However, like item difficulty, here again, a primacy effect pattern was observed in further analysis of the options. It was found that items with option A as the correct answer were the least and items with option D as the correct answer were the most discriminating items, and items with either option B or C as the correct answer functioned similarly. The findings of this section are in line with the findings of the effect of response order on item difficulty. This is not unexpected given the relationship that exists between item difficulty and item discrimination (Haladyna & Rodriguez, 2013). That is, the easier an item is, the less discriminating it will be. While previous studies did not focus on item discrimination, findings of the present study on item discrimination could be argued to be in congruence with the relevant literature (e.g., Hagenmüller, 2020; Hohensinn & Baghaei, 2017; Holzkecht et al., 2021; Kiat et al., 2018; Lions, Dartnell et al., 2022) taking into account the relationship between item difficulty and item discrimination.

Regarding the effect of response order on distractor functioning, Chi-Square analysis showed for well-functioning options, the effect was not significant, for mal-functioning options, the effect was significant, and for non-functioning options, the effect was barely significant. This may mean that well-functioning options are effective regardless of their position within the response set. But for mal-functioning options, their effectiveness was significantly influenced by their position within the response set. And for non-functioning options, their effectiveness was minimally but significantly influenced by their position within the response set. Although, to the researcher's knowledge, there are no similar studies investigating the effect of response order on distractor functioning, the findings here provide support for studies reporting a significant effect for response order on item characteristics (e.g., Hagenmüller, 2020; Hohensinn & Baghaei, 2017; Holzkecht et al., 2021) and stand in contrast with the studies suggesting no significant effect for response order (e.g., Golda, 2011; Jessel & Sulins, 1975; Marcus, 1963).

Regarding the effect of response order on the participants' confidence level in answering questions, no significant effect was observed. In other words, the position of correct answer played no significant role in guessing the right answer. However, here again, a pattern was observable. Participants were the most confident when the correct answer listed first (option A) and least confident when it was listed last (option D). The confidence levels were almost the same when the correct answer was either option B or C. This finding is highly in line with the results of item

difficulty and item discrimination. It was found that when the correct answer was option A, the item was the easiest and least discriminating and participants answered the item with the highest confidence level. Conversely, when the correct answer was option D, the item was the most difficult and discriminating and participants answered the item with the least confidence level.

As for the effect of response order on item characteristics across different language skills and components (the second research question), the results revealed that response order effect did not differ across language skills and components. In other words, response order did not behave differently for listening comprehension, grammar, and reading comprehension. It could be argued that response order is related to the position of options as a structural element of test organization regardless of the language skill being assessed. Since the previous studies (e.g., Hohensinn & Baghaei, 2017; Holzknicht et al., 2021; Lions, Dartnell et al., 2022; Sonnleitner et al., 2016) did not compare language skills and components for the effect of response order, the findings cannot be compared against the literature.

6. Conclusions and Implications

While the findings of the current study failed to show significant differences for the effect of response order on item characteristics in most cases, we cannot simply downplay its role in second language testing. The significant effects found for item difficulty (in 10% of the items) and for some distractors' functioning (mal-functioning options) along with the traces found for primacy effect in different areas can all highlight the significance of this issue in second language testing. It could be argued that in a high-stakes test like the Iranian National University Entrance Exam even a small effect of response order as a construct-irrelevant factor can bring about irreparable consequences. Thus, the correct answer position cannot simply be changed in such a test as a test security measure without compromising the validity of the test by introducing construct-irrelevant variance in test results.

The findings imply that the scores coming from different test versions with different response orders are not necessarily comparable, and this can affect the decisions made and the resulting consequences. Some test takers may get lower scores, fail a course or miss a good opportunity in education, occupation, or migration because of receiving a more difficult response order. The findings further imply that changing response order should be double checked as a test security measure. The possibility that it may introduce a construct-irrelevant variance might outweigh the advantages mentioned for reducing cheating and increasing test security.

Finally, caution should be employed in generalizing the findings of this study as the study was limited to a small sample size. Future studies are recommended to replicate this study with larger samples and consider the effect of individual characteristics of the test takers on moderating the response order effect.

References

- Ary, D., Jacobs, L. C., Irvine, C. K. S., & Walker, D. (2018). *Introduction to research in education*. Cengage Learning.
- Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement*, 40(2), 109-128.
- Baghaei, P., & Amrahi, N. (2011). The effects of the number of options on the psychometric characteristics of multiple-choice items. *Psychological Test and Assessment Modeling*, 53(2), 192-211.
- Ballard, L. (2017). *The effects of primacy on rater cognition: An eye-tracking study* [Unpublished doctoral dissertation]. Michigan State University.
- Bar-Hillel, M., Budescu, D., & Attali, Y. (2005). Scoring and keying multiple-choice tests: A case study in irrationality. *Mind & Society*, 4(1), 3-12.
- 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.
- Brown, J.D. (1997). *Testing in language program* (New ed.). The McGraw-Hill Companies
- Bruine de Bruin, W. (2005). Save the last dance for me: Unwanted serial position effects in jury evaluations. *Acta Psychologica*, 118(3), 245-260.
- Burud, I., Nagandla, K., & Agarwal, P. (2019). Impact of distractors in item analysis of multiple choice questions. *International Journal of Research in Medical Sciences*, 7(4), 1136-1139.
- 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.
- Carr, N., T. (2011). *Designing and analyzing language tests*. Oxford University Press.
- Carnegie, J. A. (2017). Does correct answer distribution influence student choices when writing multiple-choice examinations? *Canadian Journal for the Scholarship of Teaching and Learning*, 8(1), 11.
- Christenfeld, N. (1995). Choices from identical options. *Psychological Science*, 6(1), 50-55.
- Cizek, G. J. (1994). The effect of altering the position of options in a multiple-choice examination. *Educational and Psychological Measurement*, 54(1), 8-20.
- Clifton, S. L., & Schriener, C. L. (2010). Assessing the quality of multiple-choice test items. *Nurse educator*, 35(1), 12-16.
- Cohen, J.W. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Dayan, E., & Bar-Hillel, M. (2011). Nudge to nobesity II: Menu positions influence food orders. *Judgment and Decision Making*, 6(4), 333-342.
- DeVore, S., Stewart, J., & Stewart, G. (2016). Examining the effects of testwiseness in conceptual physics evaluations. *Physical Review Physics Education Research*, 12(2), Article 020138.
- Ert, E., & Fleischer, A. (2016). Mere position effect in booking hotels online. *Journal of Travel*

- Research*, 55(3), 311–321.
- Fagley, N. S. (1987). Positional response bias in multiple-choice tests of learning: Its relation to testwiseness and guessing strategy. *Journal of Educational Psychology*, 79(1), 95–97.
- Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, 87(6), 1082–1116.
- Golda, S. D. (2011). A case study on multiple-choice testing in anatomical sciences. *Anatomical Sciences Education*, 4(1), 44–48.
- Hagenmüller, B. (2020). On the impact of the response options' position on item difficulty in multiple-choice items. *European Journal of Psychological Assessment*, 37(4), 290–299.
- Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, 23(1), 17–27.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309–333.
- Haladyna, T. M., & Rodriguez, M. C. (2013). Guidelines for writing selected-response items. In T. M. Haladyna & M. C. Rodriguez (Eds.), *Developing and validating test items* (pp. 89–110). Routledge.
- Hohensinn, C., & Baghaei, P. (2017). Does the position of response options in multiple-choice tests matter? *Psicológica*, 38(1), 93–109.
- Holzknicht, F., McCray, G., Eberharter, K., Kremmel, B., Zehentner, M., Spiby, R., & Dunlea, J. (2021). The effect of response order on candidate viewing behaviour and item difficulty in a multiple-choice listening test. *Language Testing*, 38(1), 41–61.
- Jessell, J. C., & Sullins, W. L. (1975). The effect of keyed response sequencing of multiple choice items on performance and reliability. *Journal of Educational Measurement*, 12(1), 45–48.
- Kiat, J. E., Ong, A. R., & Ganesan, A. (2018). The influence of distractor strength and response order on MCQ responding. *Educational Psychology*, 38(3), 368–380.
- Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response order effects in survey measurement. *Public Opinion Quarterly*, 51, 201–219.
- Lions, S., Dartnell, P., Toledo, G., Godoy, M. I., Córdova, N., Jiménez, D., & Lemarié, J. (2022). Position of correct option and distractors impacts responses to multiple-choice items: Evidence from a national test. *Educational and Psychological Measurement*, 0(0), 1–24.
- Lions, S., Monsalve, C., Dartnell, P., Blanco, M. P., Ortega, G., & Lemarié, J. (2022). Does the response options placement provide clues to the correct answers in multiple-choice tests? A systematic review. *Applied Measurement in Education*, 35(2), 133–152.
- Marcus, A. (1963). The effect of correct response location on the difficulty level of multiple-choice questions. *Journal of Applied Psychology*, 47(1), 48–51.
- McKenna, P. (2019). Multiple choice questions: Answering correctly and knowing the answer.

Interactive Technology and Smart Education, 16(1), 59-73.

- McLeod, I., Zhang, Y., & Yu, H. (2003). Multiple-choice randomization. *Journal of Statistics Education*, 11(1).
- McNamara, W. J., & Weitzman, E. (1945). The effect of choice placement on the difficulty of multiple-choice questions. *Journal of Educational Psychology*, 36(2), 103.
- Schroeder, J., Murphy, K. L., & Holme, T. A. (2012). Investigating factors that influence item performance on ACS exams. *Journal of Chemical Education*, 89, 346–350.
- Shin, J., Bulut, O., & Gierl, M. J. (2020). The effect of the most-attractive-distractor location on multiple-choice item difficulty. *The Journal of Experimental Education*, 88(4), 643-659.
- Sonnleitner, P., Guill, K., & Hohensinn, C. (2016, July). *Effects of correct answer position on multiple-choice item difficulty in educational settings: Where would you go?* [Paper presentation]. The international test commission conference, Vancouver, British Columbia, Canada.
- Taylor, A. K. (2005). Violating conventional wisdom in multiple choice test construction. *College Student Journal*, 39, 141–153.
- Tellinghuisen, J., & Sulikowski, M. M. (2008). Does the answer order matter on multiple-choice exams?. *Journal of Chemical Education*, 85(4), 572-575.
- Wang, L. (2019). Does rearranging multiple-choice item response options affect item and test performance? *ETS Research Report Series*, 2019(1), 1-14.
- Winke, P., & Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing*, 25, 37–53.