International Journal of Applied Linguistics & English Literature
E-ISSN: 2200-3452 & P-ISSN: 2200-3592
www.ijalel.aiac.org.au

The Role of Order and Sequence of Options in Multiple-choice Questions for High-stakes Tests of English Language Proficiency

Talip Karanfil, Steve Neufeld*

Middle East Technical University, Northern Cyprus Campus, 99738 Kalkanlı, Güzelyurt, Mersin 10, Turkey

Corresponding author: Steve Neufeld, E-mail: steve@metu.edu.tr

ARTICLE INFO

Article history
Received: July 20, 2020
Accepted: September 29, 2020
Published: November 30, 2020
Volume: 9 Issue: 6
Advance access: November 2020

Conflicts of interest: None
Funding: None

ABSTRACT

High-stakes and high-volume English language proficiency tests typically rely on multiple-choice questions (MCQs) to assess reading and listening skills. Due to the Covid-19 pandemic, more institutions are using MCQs via online assessment platforms, which facilitate shuffling the order of options within test items to minimize cheating. There is scant research on the role that order and sequence of options plays in MCQs, so this study examined the results of a paper-based, high-stakes English proficiency test administered in two versions. Each version had identical three-option MCQs but with different ordering of options. The test-takers were chosen to ensure a very similar profile of language ability and level for the groups who took the two versions. The findings indicate that one in four questions exhibited significantly different levels of difficulty and discrimination between the two versions. The study identifies order dominance and sequence priming as two factors that influence the outcomes of MCQs, both of which can accentuate or diminish the power of attraction of the correct and incorrect options. These factors should be carefully considered when designing MCQs in high-stakes language proficiency tests and shuffling of options in either paper-based or computer-based testing.

Key words: Multiple-choice Questions (MCQs), Option Order in MCQs, Sequence Priming in MCQs, Order Dominance in MCQs

INTRODUCTION

The assessment of learning, which is a notoriously time-consuming and challenging aspect of education in general, is even more problematic when the subject is learning a foreign language (Bachman et al. 1996). High-stakes standardized language proficiency tests are often needed for hiring and career advancement, entry requirements for professions, immigration and citizenship, and universities where English is the medium of instruction. Institutions can use third-party commercial testing services or develop their own in-house language proficiency tests. Commercial testing services offer computer-based testing, while in-house testing is normally paper-based, either hand-marked or scanned optic answer sheets. Despite concerns about washback (Messick 1996), many educational institutions use paper-based multiple-choice questions (MCQs) due to their reliability, validity, and ease of scoring. The Covid-19 pandemic of 2020 has meant that many institutions are converting paper-based tests to online testing platforms, one of the most common being the open-source learning management system (LMS) MOODLE. The quiz module of MOODLE defaults to shuffling options within multiple-choice questions to minimize cheating, so that each student sees a different order of options throughout all the items in one test. While there has been considerable research to inform test designers about the ideal number of options in one question, and the order of the questions within a test, there has been extraordinarily little research on the influence of the order and sequence of options within a test item as measured by classic test theory analysis.

Features of Multiple-choice Questions and Exams

There are well-established guidelines for creating MCQ test items (Haladyna et al., 2002; Gierl et al., 2017) to ensure maximum reliability and validity. An MCQ consists of a stem that provides the context for the question, and several options, including the option that correctly completes the stem as well as several incorrect but plausible answers, referred to as distractors. The skills that lend themselves to MCQ format are listening and reading, while alternative assessment strategies are required to assess productive skills, such as writing and speaking.

The number of options in an MCQ generally varies from three to five. Sadeghi and Masoumi (2017) report findings that indicate the difficulty of an MCQ increases with the number of options. However, they found that there is no significant difference between MCQs with three and four options in English language proficiency tests. Rodriguez

Published by Australian International Academic Centre PTY.LTD.

http://dx.doi.org/10.7575/aiac.ijalel.v.9n.6p.110

(2005) reviewed research over the past eighty years on the optimal number of options in an MCQ, reporting broad support for three-option MCQs. Shizuka et al. (2006) report that for different skills, such as reading, there are factors other than the number of options that affect the level of difficulty of an MCQ. In assessing vocabulary for English for Academic Purposes (EAP), Oruç, Ertürk and Mumford (2017) point to considerations related to experiential factors of the test takers and designers in addition to objective test design guidelines. Nevertheless, regarding the number of options, language proficiency tests predominately have three- or four-option MCQs, and it is generally recognized that 5-option MCQs will be more difficult.

This exam format is vulnerable to cheating or malpractice. To minimize this risk, multiple versions of the test can be produced, consisting of the same items but in a different order (Davis, 2017). General findings show the order of items is not a significant factor in reliability (Davis, 2017; Stout and Heck n.d.). In one such study, researchers created three different versions of the same test by scrambling the order of items. They found that changing the order of items did not affect student performance or the overall difficulty of the test (Satti et al., 2019). In contrast, another study by Ollennu and Etsey (2015) suggests that changing the order of items may have a significant effect on overall test difficulty. It is recommended to have items ordered easy first and more difficult at the end (Hambleton et al. 1974)—a procedure difficult to implement in English language proficiency tests, in which questions are based on listening or reading texts which determine the order of the test items.

Item Analysis According to Classic Test Theory

After a test is administered, one approach to determine its effectiveness is to analyze psychometric properties using item and sample dependent statistics. Classical Test Theory (CTT) employs two main techniques: item facility and item discrimination, which are frequently reported item characteristics in language assessment (Brown, 2005; Bachman, 2004).

Item facility, sometimes referred to as Ease Index (EI), refers to the percentage of test-takers who answered an item correctly. This is a simple calculation of the number of test-takers that answered the item correctly divided by the total number of test-takers and can have a value from 0.00 (so difficult that no-one got the correct answer) to 1.00 (so easy that everyone got the right answer).

Item discrimination, sometimes referred to as Discrimination Index (DI), reflects the extent to which the item differentiates test-takers who scored high on a test from those who did poorly. This is calculated by identifying the top and bottom third of the test-takers based on their overall score on the test, and for each item subtracting the EI of the lower third from the upper third. The value for DI ranges from +1.00 (all the test-takers in the upper group correctly answered the question, while none of the bottom third did) to -1.00 (all the test-takers in the lower group correctly answered the question, while none of the upper third did) The value of 0.00 indicates that there is no contrast between the performance of test-takers.

When interpreting the item analysis statistics, EI and DI must be considered together. An item that has a high EI will typically have a low DI, i.e., it will not discriminate between the upper and lower thirds, while an item that has a low EI should have a high DI, discriminating positively for the upper third. In all cases, a negative DI would point to a problematic item.

Item Analysis of Identical MCQs with Different Option Order

The AUTHORS'_INSTITUTION produces its own paper-based English Proficiency Exam (EPE). This high-volume, high-stakes exam is administered simultaneously across three campuses (X, Y, and Z) to over 4,000 students on five separate occasions during one calendar year. Each EPE is bespoke, and none of the questions are recycled or reused in any subsequent test.

To minimize the potential for cheating and maximize the efficiency of invigilation during the exam, two versions of the test are prepared for the listening, reading, and vocabulary sections, which all consist of MCQs. The questions are in the same order for both versions. However, the order of options in the original version (TEST A) is randomly changed to produce the second version (TEST B). The focus when randomizing the order of options is to "*vary the location of the right answer"* (AUTHORS'_INSTITUTION testing team, personal correspondence) and keep a balanced distribution among the questions, a common approach recommended in most guidelines (Haladyna et al., 2002; Gierl et al., 2017). Thus, overall, the location of options that are correct are in balance, i.e., each position is to be used the same number of times for the correct response throughout the test.

After the item analysis of the EPE, taken by 163 test-takers, a striking difference in the results between TEST A and TEST B was observed for several items. In particular, the very last question in the exam, a 5-option MCQ on vocabulary, stood out. The stem was a gapped sentence, and the five options were words to fill the gap. In TEST B, the correct option was in position E, and the option in position C was the dominant distractor--the option which was most chosen of the remaining options. In TEST A, these two options were swapped, leaving the other options in the same position. In TEST A, 36% of the test-takers got the correct answer when it was in position C, but only 6% chose it when it was in position E for those who took TEST B. Likewise, the distractor in position E in TEST A was selected by 12% of the test-takers, while 30% opted for the distractor when it moved from position E to position C in TEST B. The DI for each test format was 0.07, and overall was 0.02, i.e., the item provided virtually no discrimination generally expected for this level

	Question	KEY	A	B	C	D	E
TEST A	82	C	15%	23%	36%	11%	12%
TEST B	82	E	17%	26%	30%	21%	6%

Figure 1. Item analysis of one MCQ showing extreme variance between test versions.

of difficulty between the students with scores in the top and bottom third overall.

In both test versions, the order of the other options was the same, and despite the difference in difficulty between the two versions, the low DI was the same for each version. When the correct option was moved to the last position, the difficulty increased dramatically. Also, the option that was the dominant distractor changed from the option in position B in TEST A to the option in position C in TEST B. This suggests that the options are not independent of each other; rather, there is a relationship between the options that is determined by the order and sequence in which they appear. Specific research about such a relationship between options is scarce, most dating from the last half of the twentieth century, and the scant literature is contradictory (Marcus 1963; McNamara and Weitzman 1945; Cizek 1994; Mosier and Price 1945). More recent research focuses on different aspects regarding the options such as the effect of the dominant distractor's location (Tellinghuisen and Sulikowski, 2008; Shin et al., 2019; Hohensinn and Baghaei, 2017). With respect to assessing listening, Holzknecht et al. (2020) confirmed the primacy effect, where the correct answer is chosen more often when it occurs in the first option. They used eye-tracking software to show that in MCQ listening tests, scores were lower when the correct options were in later positions, i.e., the position of the correct answer can affect the level of difficulty of the question. Aside from these, research in the early twenty-first century mostly focuses on the order of questions in a test and the number of options, varying between 3, 4, and 5, but not the order of options within questions (Satti et al., 2019; Shizuka et al., 2006).

Research Questions

Over two years, we conducted item analyses of our high-stakes, high-volume MCQ EPEs, administered face-to-face in TEST A and TEST B versions. All questions in both versions appeared in the same order, but the order of options for each question in the second version was changed. The analyses consistently revealed differences in the difficulty and discrimination indexes of some of the questions in the two different versions.

Since the question stems/items were identical and in the same order in both booklets, the change in order of the response options was likely the reason for the difference, bearing in mind that item parameters calculated using classical test theory are sample dependent. Consequently, our research questions are:

1. To what extent does the order of options within an MCQ affect the ease and discrimination indexes of that question?
2. How does the influence of the order of options within an MCQ vary when assessing listening and reading?
3. What types of interdependency exist between options based on the order and sequence in which they appear?

METHOD

To test our hypotheses, we needed to administer a test with the same MCQs in two versions to two similar groups of students, each group seeing the same MCQs in the same order, but with the options in each in a different order. The METU EPE provided exactly these two versions. Still, we needed to form the two groups of test-takers so that we could identify and discount any sample dependency that might contribute to the variance we would observe in the CTT item analysis. In other words, to what extent could any difference in EI or DI revealed be attributed solely to the change in the option order and sequence. The ETP course at METU NCC presented a situation that provided an ideal context for our study.

Context of the study

The students

There were 73 students in the ETP course. All ETP students were high school graduates who had completed one full year of the English Preparatory Program (EPP) and achieved at least a 65% average of end-of-year work to be eligible to take the METU EPE. All had attempted the EPE at least once and achieved a score between 50% and 59% in at least one attempt, approximating a B1 CEFR level of proficiency. However, since the minimum required threshold is 59.5% to move on to the department, these students had to continue with their English learning, and they attended the ETP course.

For this study, this group presented a very homogenous group. Their age, cultural and educational background, native language (Turkish), English language knowledge, exposure to English outside of class, and motivation were all very similar.

The course

The English Towards Proficiency (ETP) course runs over a 16-week semester, four hours a day, five days a week, and is geared towards proficiency exam skills preparation in listening, reading, and writing, with an emphasis on active knowledge of academic vocabulary.

Of the 73 students enrolled in the course, 64 took the EPE and 44 of these students passed, a "success" rate of 69% for the actual test-takers and 60% of all who had enrolled. This high success rate reputation of the program also helps student motivation, which positively affects attendance, participation, and attention.

The course syllabus is tightly confined to exam preparation, with 80% minimum attendance required, so there was virtually no variation in the amount of input each student received before they took the EPE.

The exam

The English Proficiency Exam (EPE) is an in-house paper-based English language proficiency exam. It consists of 74 multiple choice questions for listening, reading and vocabulary, a writing task, and a listening/reading synthesis task. For the multiple-choice sections of the EPE, an MCQ optical form is given to each student to code their answers.

The listening section consists of twenty-five 3-option items, one 5-option item, and four 6-option items, each worth one point. Reading consists of 22 3-option items and two 5-option questions, each worth one point, Vocabulary consists of 20 5-option items, each worth 0.5 points.

The test is produced from scratch by a testing team at METU Ankara, which has no connection with teaching or running the ETP at METU NCC, so there is no subjectivity in designing the test based on the ETP student profile. After all team members and the administration approve the test, it is produced in two versions, TEST A and TEST B, to minimize the opportunity for cheating. When making these two versions, all the questions are kept in the same order, but the order of the options within each question is different, mixed at random while maintaining a balance between the position of correct options.

Exam seating arrangement

The students were placed in three classrooms to take the EPE. They were seated according to their rank in the course work, based on the overall average of the midterms given through the course. Therefore, the upper third of the group were together in one class (21 students), the middle third of the group were in another (22 students), and the bottom third was in a third class (21 students).

During the EPE, the students sit in four rows in each class in alphabetical order by surname. TEST A booklets are distributed to the first and third rows, and TEST B booklets are distributed to the second and fourth rows.

Distribution and Comparison of TEST A and TEST B booklets

The seating arrangement grouped the top and bottom thirds into separate classrooms according to ETP midterm results, ensuring an even distribution of TEST A and TEST B to each. Therefore, when conducting the CTT item analyses of the two test versions, we were sure that the test versions had been evenly distributed. The respective top and bottom sections of the overall results in the EPE were comprised of, according to rank in the ETP midterm results, the 22 students ranked at the top and the lowest 23 students.

The student profile, the course of study, the seating arrangement, and the distribution of the two versions of the test, in which the questions were all in the same order, all ensure that the different order of options within each question is the main variable that would affect differences in the CTT indexes for ease and discrimination between TEST A and TEST B.

Sample Variables between the Two EPE Test Version Cohorts

To determine precisely how well the exam grouping and booklet distribution provided equally balanced cohorts, we reviewed the sampling according to student performance in the ETP course as well as the EPE to make sure there were no coincidental sample variables that might interfere with the results.

ETP Midterms Profile

The ETP Midterm MCQ profile in Figure 2 below shows the high degree of homogeneity of the 64 students that were eligible to take the English Proficiency Exam.

The MT averages are for the overall midterm results, including writing and note-taking. These two sections are the only two sections in which students are required to use their productive skills and write a paragraph. The listening, reading, and vocabulary are all MCQ sections in the midterm, which is a simulation of the English Proficiency Exam (EPE).

Normalized distribution between ETP overall average and EPE MCQ performance for all students

The relative similarity of the normalized distribution between the overall average of three ETP midterms and EPE MCQs in Figure 3 supports the notion that there appears to be no significant sample dependent variables to consider when comparing the discrimination index between TEST A and TEST B in our study outside the order and sequence of options within each question.

ETP and EPE performance for CTT top and low thirds of TEST A and TEST B cohorts

In addition, we further compared the performance of the students who took TEST A with those who took TEST B by comparing their year work total of the ETP midterms (MCQ parts only - Listening, Reading, and Vocabulary) and the overall EPE score (MCQ parts for the same three sections). We found that the students who took TEST A scored slightly better than the students who took TEST B in both the EPE and ETP, but the correlation between the two was remarkably high.

The comparison between the TOP THIRD students who took TEST A and TEST B concerning their ETP year work and EPE score is shown in Figure 4. The relative difference between the groups in ETP and EPE is almost identical, with the students taking TEST A slightly outperforming the students taking TEST B.

The comparison between the LOW THIRD students who took TEST A and TEST B regarding their ETP year work and EPE score is shown in Figure 5. The relative difference between the groups in ETP and EPE is almost identical, and like the TOP THIRD group, the LOW THIRD students taking TEST A slightly outperform the students taking TEST B.

CTT sampling analysis

Furthermore, to investigate if there are any sample dependency effects within the three CTT item analysis bands, Figure 6 shows the comparisons between TEST A and TEST B groups for TOP, MID, and LOW thirds. The radar graphs below show that the TOP, MID, and LOW third CTT sampling of the EPE groups based on the ETP ranking has produced consistent samples within each third of the CTT analysis bands.

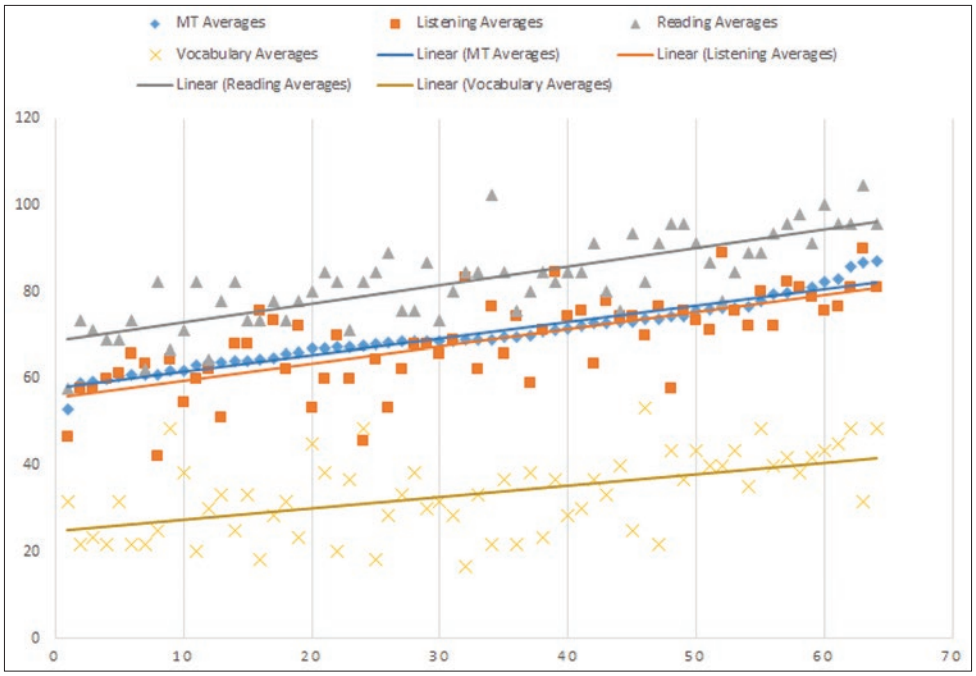


Figure 2. ETP Midterm MCQ profile

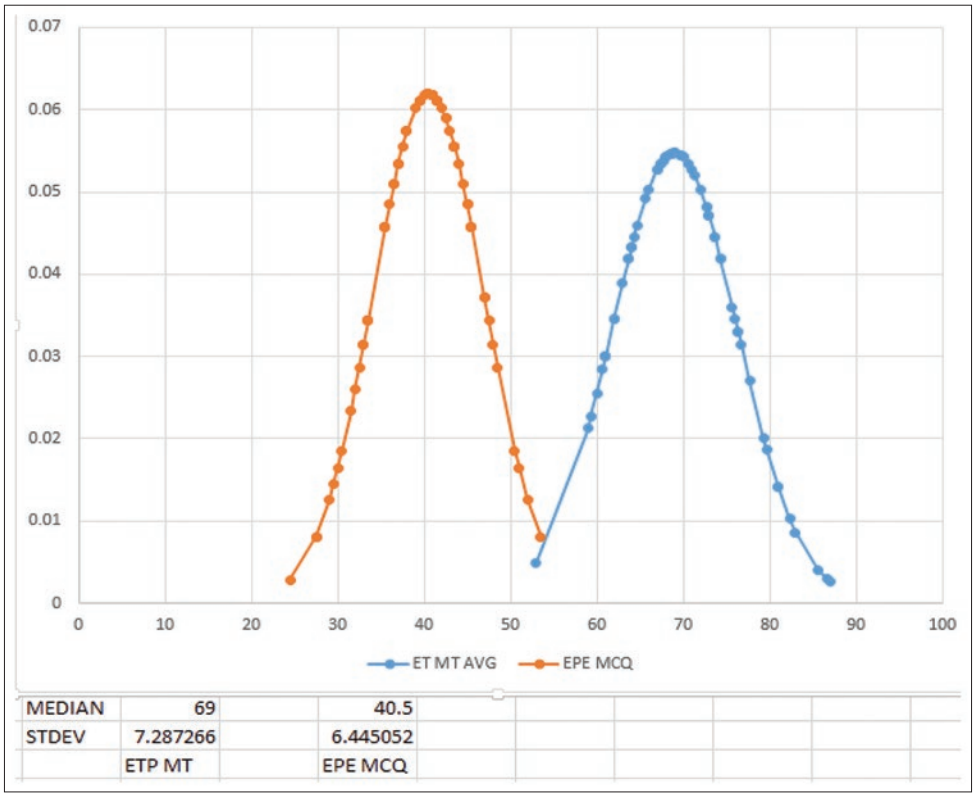


Figure 3. Normalized distribution of ETP Midterm and EPE MCQs

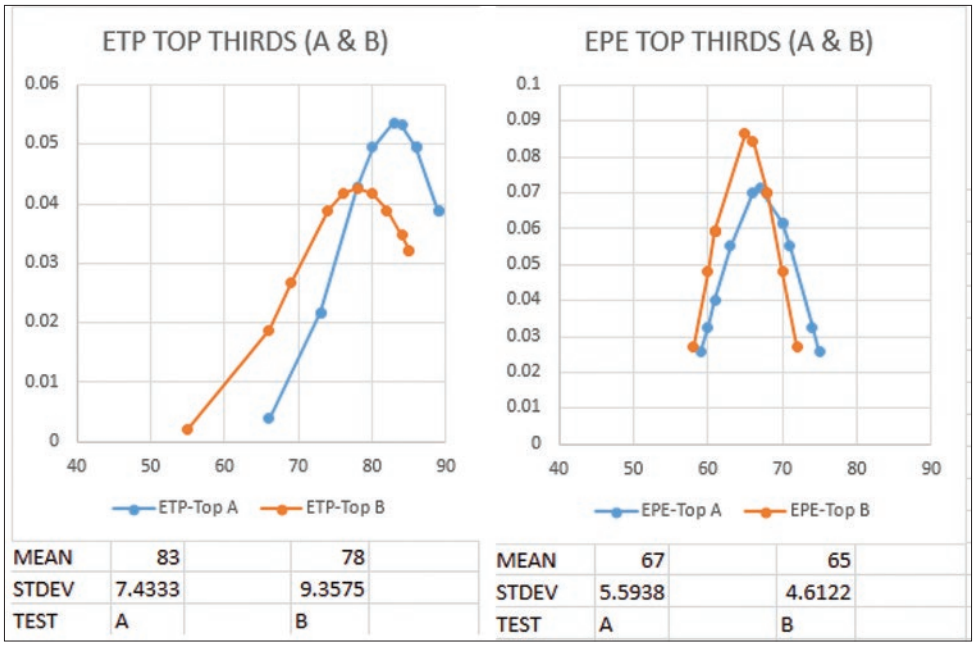
The ETP group profile provides an ideal opportunity to examine the impact of the order and sequence of options for individual questions in TEST A and TEST B performance as measured by CTT

Option Analysis

Having determined that the primary variable that differentiates TEST A and TEST B is the order and sequence of the options within each question, we examined this variable in more detail.

Option number variance

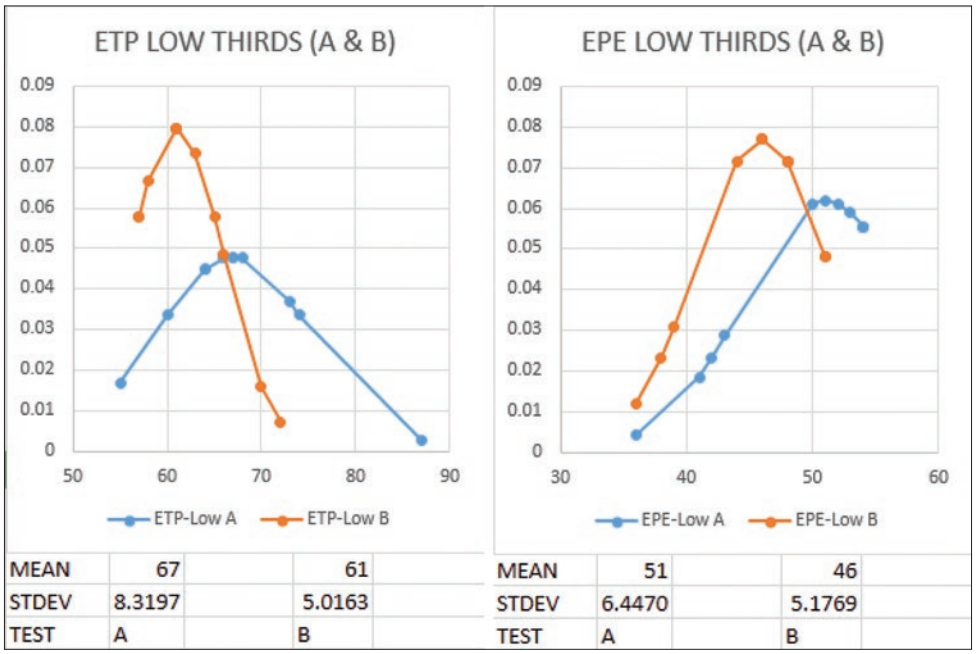
Our focus is to investigate the impact of different option order and sequence between the questions in TEST A and TEST B. The general principle employed by the METU test writers is to design questions with only three options,



MEAN	83	78
STDEV	7.4333	9.3575
TEST	A	B

MEAN	67	65
STDEV	5.5938	4.6122
TEST	A	B

Figure 4. TOP third (Test A and B) performance in ETP and EPP



MEAN	67	61
STDEV	8.3197	5.0163
TEST	A	B

MEAN	51	46
STDEV	6.4470	5.1769
TEST	A	B

Figure 5. LOW third (Test A and B) performance in ETP and EPP

except in the vocabulary section in which each question has five options. However, in the listening section, there was one question that had five options and a set of four related questions, each having six options. We removed these questions from our analysis because the level of difficulty changes with the number of options, as noted in our introduction. Therefore, for the specific focus of this paper, we only considered the 25 questions in the listening section that had three options. Similarly, in the reading section, there was a set of two related questions that each had five options. We removed these from the analysis to focus only on 3-option MCQs, which would allow us to compare question behavior with listening.

Option sequence patterns

As our research questions focus on the order and sequence of the options within a question, we classified each question based on the relative position of the correct option to the dominant distractor. The test designers reported that when creating the options, they did not consciously predict which of the distractors would be dominant. However, CTT analysis for difficulty provides a percentage value for how many test-takers chose each option. Therefore, we used these percentages to identify the option, which was the dominant distractor in each question, and then observed the position of the dominant distractor in relation to the correct option in each version of the question in TEST A and TEST B.

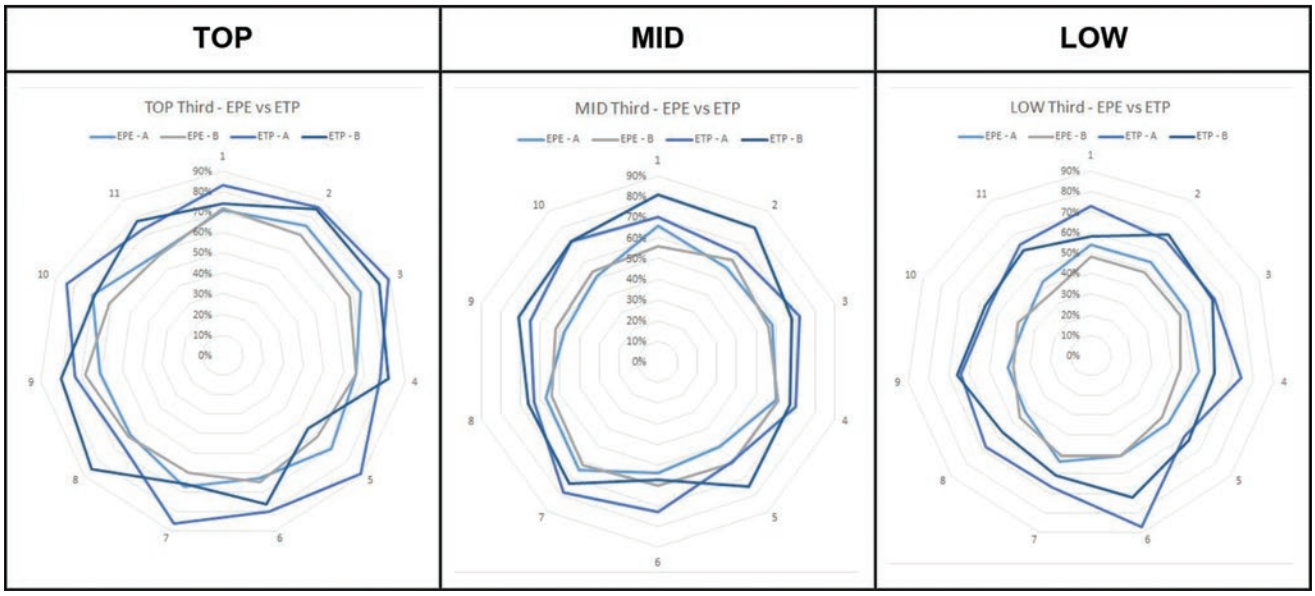


Figure 6. Radar graph analysis of CTT sampling

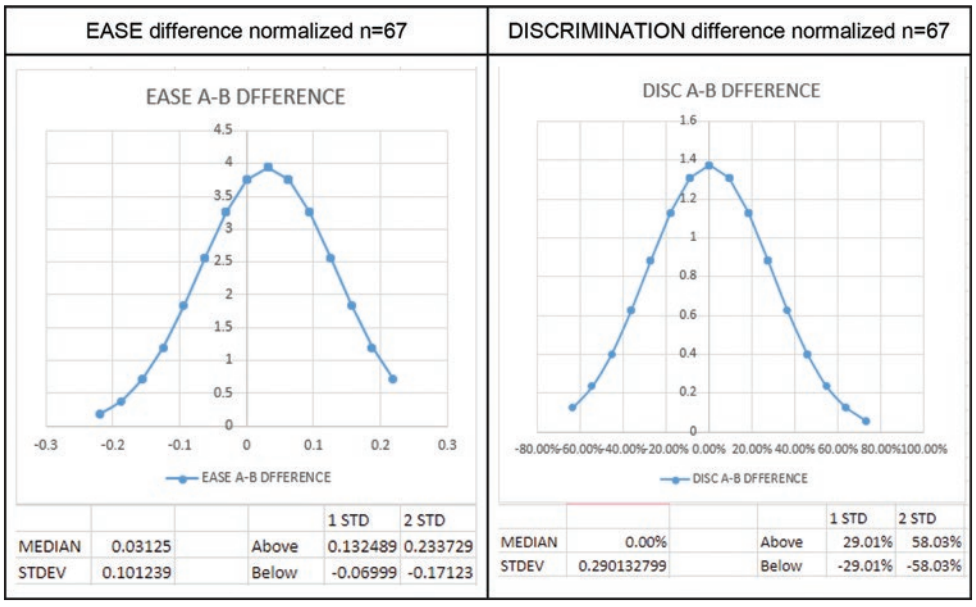


Figure 7. Normalized distribution of EI and DI differences between TEST A and TEST B

Accordingly, we classified the questions in both TEST A and TEST B according to the four patterns explained in Table 1.

The threshold of problematic variance between TEST A and TEST B questions

To determine a threshold to distinguish abnormal question behavior between the two test versions, we calculated the DISCRIMINATION and EASE differences between Test A and Test B for the forty-seven 3-option and twenty 5-option questions combined, disregarding the seven MCQs in listening and reading that were not 3-option. We used the differences to create a normalized distribution, as shown in Figure 7. Because of the homogeneous nature of the group, we flagged all the questions that were outside one standard deviation as problematic questions to investigate.

Questions whose difference between EASE and DISCRIMINATION in TEST A and TEST B was within one standard deviation were assumed to exhibit normal behavior in terms of CTT. Questions outside this threshold were investigated. The relative percentage of such questions in reading was over double that of either listening or vocabulary.

Table 2 below gives a summary of the 15 questions that were outside one standard deviation, with the raw data and the CTT indexes. Question numbers up to 30 are for listening, and question numbers above 54 are for vocabulary, with the rest for reading. The same data for the 52 questions that were within one standard deviation can be found in APPENDIX 1: CTT item analysis for MCQs with normal behavior.

Table 1. Option sequence patterns

Pattern	Classification code
The correct option preceded the dominant distractor in both TEST A and TEST B.	AB
The correct option preceded the dominant distractor in TEST A but occurred after the dominant distractor in TEST B.	A*
The correct option preceded the dominant distractor in TEST B but occurred after the dominant distractor in TEST A.	*B
The dominant distractor preceded the correct option in both TEST A and TEST B.	**

RESULTS

Having established that there is minimal interference between the TEST A and TEST B cohorts, other than a slight tendency for the TEST A cohorts (top and low thirds) to perform marginally better than their counterparts in TEST B, the following results were gleaned from the CTT item analysis.

PATTERN impact on EASE and DISCRIMINATION indexes

When all 74 MCQs are analyzed without differentiating the option patterns, the EASE index is virtually identical between TEST A and TEST B, as shown in Figure 9 below. Likewise, there is extraordinarily little difference in the discrimination index when comparing TEST A and TEST B. However, when these indexes are calculated to consider

Table 2. Raw data and CTT indexes for problematic questions

		2	18	23	26	38	41	43	44	45	47	51	54	57	66	70
GRADE		1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.5	0.5	0.5
BOOKLET A		C	C	B	B	C	A	C	B	A	B	C	B	B	A	A
BOOKLET B		A	A	C	A	B	C	B	B	C	A	C	B	C	D	B
1st KEY		B	B	A	B	B	A	B	-	A	B	-	-	A	A	A
Sequence		*B	*B	**	*B	**	A*	**	A*	A*	*B	**	**	**	A*	A*
		~			~		~	~			~			~	~	
EASE A		41%	75%	34%	28%	81%	47%	72%	44%	38%	63%	50%	47%	53%	9%	34%
EASE B		59%	72%	34%	6%	78%	63%	59%	38%	41%	84%	41%	50%	75%	25%	31%
EASE ALL		50%	73%	34%	17%	80%	55%	66%	41%	39%	73%	45%	48%	64%	17%	33%
		~	~	~	~	~	~	~	~	~	~	~	~	~	~	~
DISC A		45%	9%	64%	55%	55%	18%	64%	9%	-18%	55%	36%	36%	-18%	18%	-9%
DISC B		0%	45%	18%	9%	9%	55%	-9%	64%	18%	-9%	0%	-36%	45%	-45%	36%
DISC ALL		24%	29%	43%	29%	33%	33%	24%	33%	-5%	29%	24%	0%	14%	-10%	10%
COUNT TOP A		6	8	8	6	11	7	10	5	2	9	6	8	5	2	4
COUNT LOW A		2	8	1	1	6	5	4	4	4	4	3	4	8	0	5
COUNT TOP B		7	11	5	1	9	9	6	8	6	9	3	3	11	0	6
COUNT LOW B		7	6	3	0	8	3	7	1	4	10	3	7	6	5	2
COUNT ALL	#	32	47	22	11	51	35	42	26	25	47	29	31	41	11	21
EASE A-B	#	-0.19	0.03	0.00	0.22	0.03	-0.16	0.13	0.06	-0.03	-0.22	0.09	-0.03	-0.22	-0.16	0.03
DISC A-B	#	0.45	-0.36	0.45	0.45	0.45	-0.36	0.73	-0.55	-0.36	0.64	0.36	0.73	-0.64	0.64	-0.45
TA-A	0	0.47	0.13	0.47	0.47	0.09	0.47	0.09	0.22	0.38	0.22	0.28	0.41	0.28	0.09	0.34
TA-B	0	0.13	0.13	0.34	0.28	0.09	0.28	0.16	0.44	0.13	0.63	0.22	0.47	0.53	0.31	0.09
TA-C	1	0.41	0.75	0.19	0.25	0.81	0.25	0.72	0.34	0.5	0.16	0.5	0.13	0.03	0.19	0.22
TA-D	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.22	0.28
TA-E	0	0	0	0	0	0	0	0	0	0	0	0	0	0.16	0.19	0.03
TA-F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TB-A	0	0.59	0.72	0.13	0.06	0.13	0.09	0.28	0.31	0.13	0.84	0.25	0.38	0.09	0.16	0.38
TB-B	1	0.28	0.13	0.53	0.5	0.78	0.28	0.59	0.38	0.47	0.09	0.34	0.5	0.06	0.22	0.31
TB-C	0	0.13	0.16	0.34	0.44	0.09	0.63	0.13	0.31	0.41	0.06	0.41	0.13	0.75	0.28	0.19
TB-D	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.25	0.06
TB-E	0	0	0	0	0	0	0	0	0	0	0	0	0	0.09	0.09	0.06
TB-F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TA-TOT	1	1	1	1	1	1	1	0.97	1	1	1	1	1	1	1	0.97
TB-TOT	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

the influence of the option patterns, it becomes clear that when the correct option occurs before the dominant distractor in both TEST versions (pattern AB), both the EASE and DISCRIMINATION indexes are optimal. As can be seen in Figure 10 below, there is significant negative impact on EASE and DISCRIMINATION when the dominant distractor occurs before the correct option (patterns *B, A*, or **).

Option Balance Analysis

The following analyses examine the impact of the actual position of the correct option in TEST A and TEST B without considering the sequence in relation to the dominant distractor. Due to the different nature of the vocabulary 5-option MCQs, we have focused only on the 3-option MCQs from this point. The effect of the correct option position overall is contrasted with the profile for the problematic and normal behavior questions. The skills of listening and reading are also contrasted in the two different test versions.

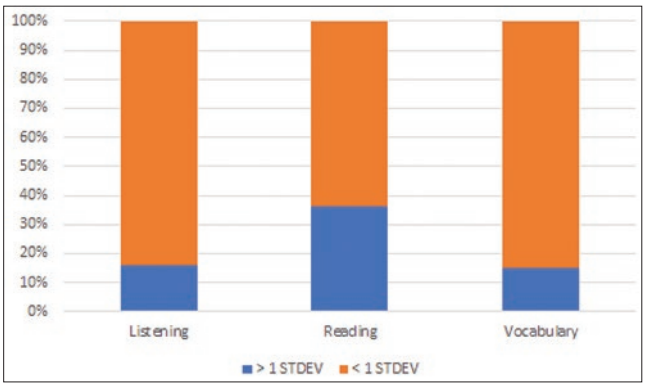


Figure 8. Percentage of abnormal questions breakdown by skills

Option balance in 3-option MCQs overall

Of all the 47 3-option MCQs selected, TEST B had two more correct options in positions A and B, and 4 less in position C.

Option balance in 3-option MCQs in reading and listening between problematic and normal behavior

Differentiating the option balance between the problematic and non-problematic questions in Figure 12 shows less balance in the problematic questions for listening. None of the four problematic listening questions in TEST A had the correct answer in position A, while three out of four questions in TEST B had the correct answer in position A. In contrast, the balance in reading questions in both TEST A and TEST B was similar to the questions with normal behavior as shown in Figure 13.

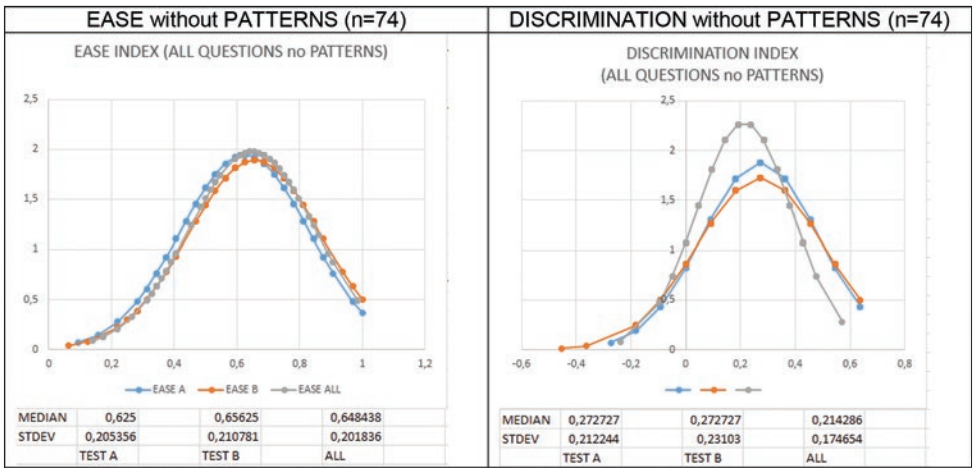


Figure 9. EASE and DISCRIMINATION without option patterns

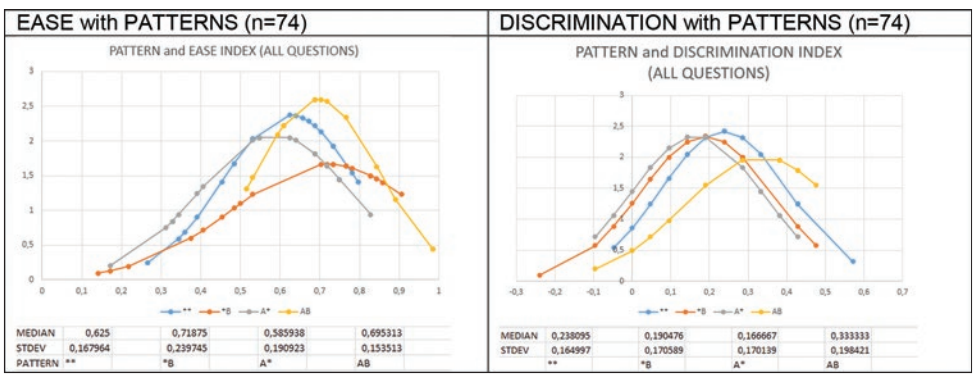


Figure 10. EASE and DISCRIMINATION with option patterns

Option Sequence Analysis

The following analysis of the 3-option and 5-option MCQ examines the impact of the actual position of the correct option in TEST A and TEST B, taking into consideration its relationship to the dominant distractor. The effect of the patterns of the relationship between the correct option and the dominant distractor is contrasted for the problematic and normal behavior questions. In the 3-option MCQ analysis, the skills of listening and reading are also contrasted in the two different test versions.

Sequence patterns of problematic and normal behavior questions

Table 3 below gives a breakdown of 15 problematic questions we identified in terms of differences between the TEST A and TEST B groups and the 52 questions that exhibited normal behavior.

The overall sequence patterns comparing problematic with normal behavior in Figure 14 below shows that none of the questions in either Test A or Test B in which the correct option occurred in a position before the dominant distractor exhibited abnormal behavior.

When looking at the distribution of problematic questions in the skills areas, as shown in Figure 15 below, a common feature distinguishing abnormal question behavior is that the dominant distractor occurs in a position before the correct option. When this happened in both Test A and Test B, it was a common problem in all three skills areas, but the impact is most noticeable in reading. When this happened in Test A, it was problematic for listening; when it happened in Test B, it was problematic in vocabulary.

Sequence patterns of problematic and normal behavior questions

These results delve into the sequence patterns considering the CTT item analysis and contrast the features of the top, middle, and bottom thirds of each group taking TEST A and TEST B.

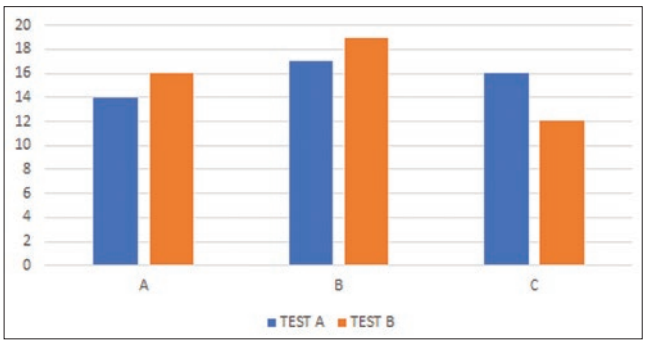


Figure 11. Option balance for all 3-option MCQs

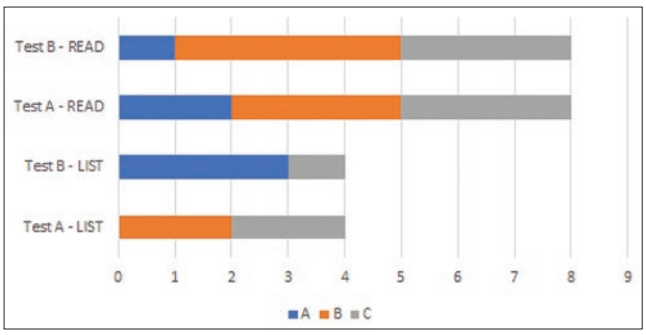


Figure 12. Option balance in problematic questions – 3-options in Listening and Reading

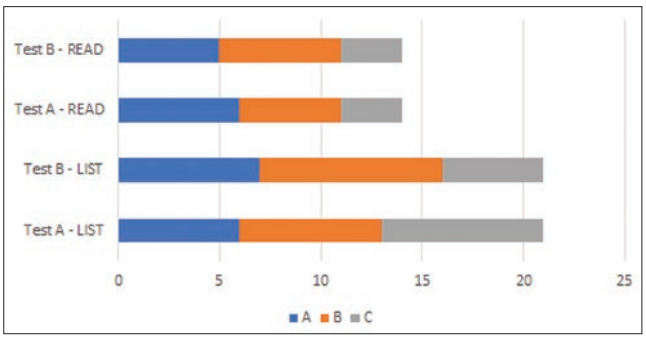


Figure 13. Option balance in normal behavior questions – 3-options in Listening and Reading

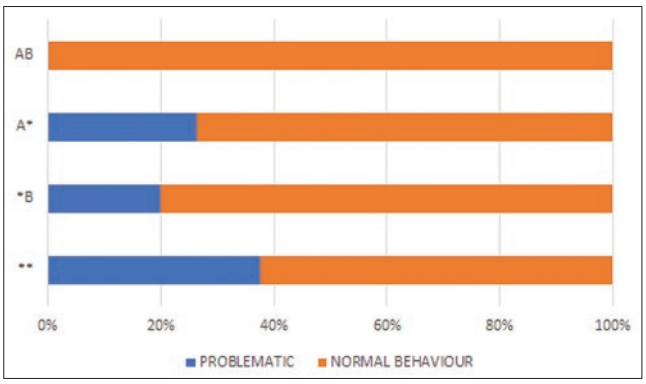


Figure 14. Option sequence patterns in normal and abnormal questions

Table 3. Breakdown of questions with normal and abnormal behavior

PATTERN	Question with problematic behavior				Questions with normal behavior			
	Listening	Reading	Vocabulary	SUBTOT	Listening	Reading	Vocabulary	SUBTOT
**	1	4	1	6	4	2	4	10
*B	3	1	0	4	6	4	6	16
A*	0	3	2	5	5	4	5	14
AB	0	0	0	0	6	4	2	11
TOTAL	4	8	3	15	20	14	17	51

Top thirds

The sequence patterns are sorted for each problematic question, and the number of students who got the question right in each test version is plotted in Figure 16 below.

There is a clear trend in favor of the students in Test A performing consistently better than their colleagues answering the same question in Test B when the dominant distractor occurred before the correct option in both test versions. However, this trend is reversed and the students taking Test B scored consistently higher when the correct option occurred before the dominant distractor in their version, but after it in Test A. In both cases of clear and consistent trends, the questions are from the reading section, aside from one listening. In the pattern where the correct option occurs be-

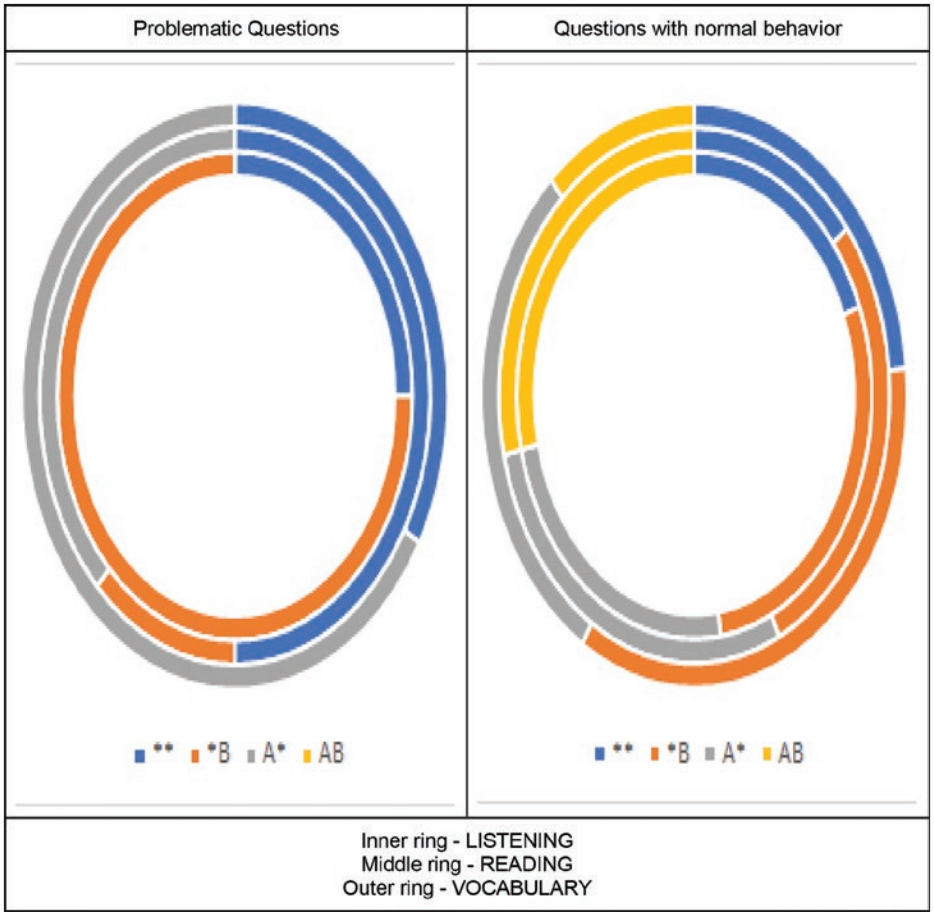


Figure 15. Distribution of pattern sequences in normal and abnormal questions by skills

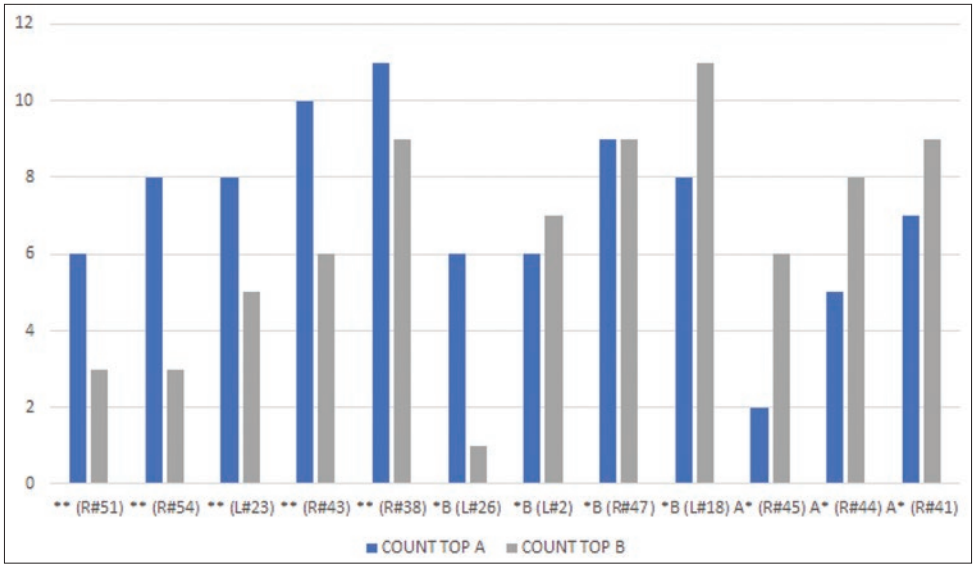


Figure 16. Number correct answers in TOP third cohorts for TEST A and TEST B in problematic questions

fore the dominant distractor, the students taking Test B score better or as well as those from Test A in three out of four questions. However, the trend is not as straightforward as in the other two patterns. It should be noted that in this pattern, there are only questions from the listening section.

The two cohorts scored almost identically overall on the questions that exhibited normal behavior, and there is no obvious skills bias as shown in Figure 17 below.

Low thirds

Like the previous section, for each problematic question, the number of students in the low thirds who got the question right in each test version is plotted in Figure 18 below.

The results are a mirror image of the results of the top third above. There is a clear trend in favor of the students in Test B performing consistently better than their colleagues

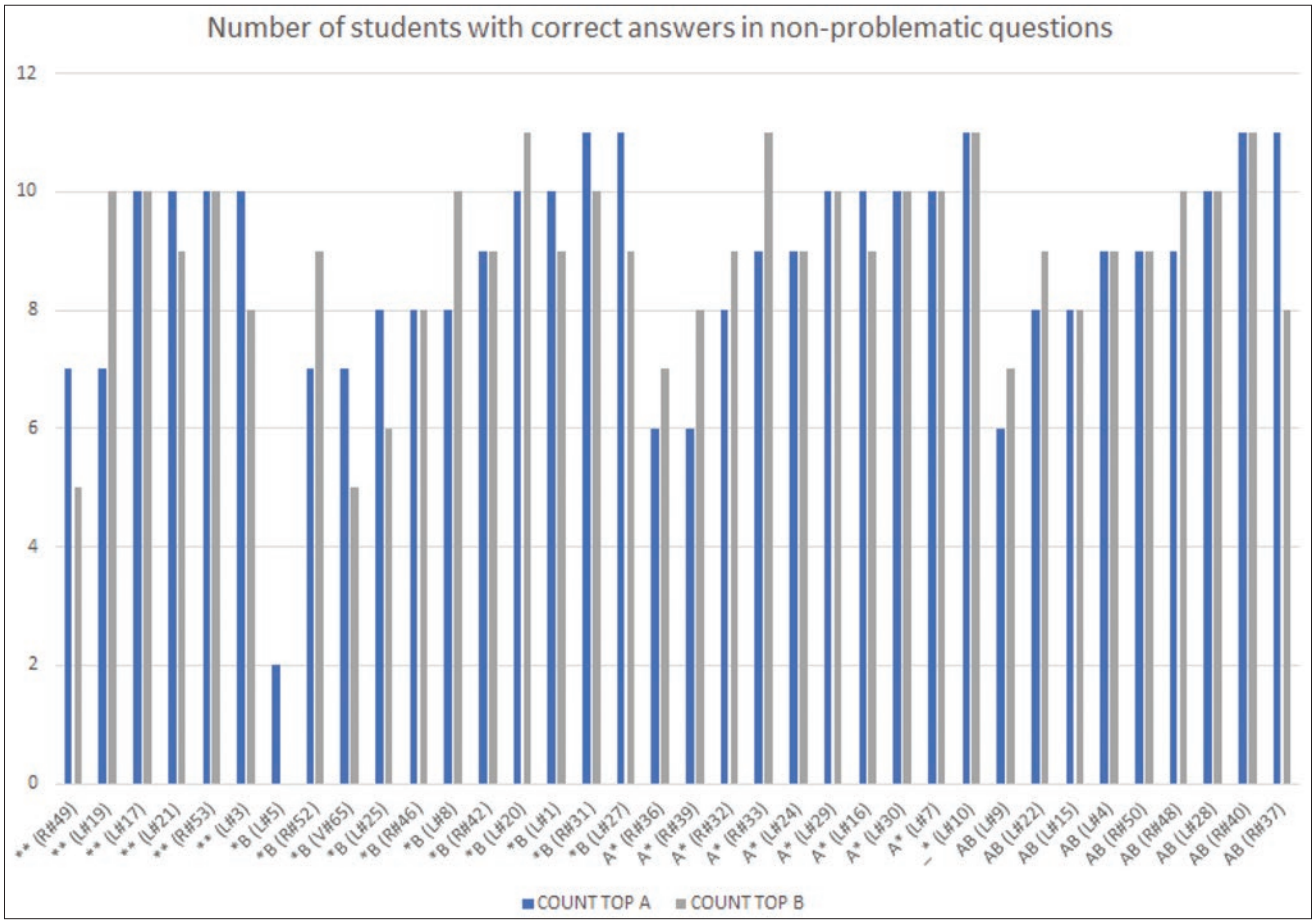


Figure 17. Number correct answers in TOP third cohorts for TEST A and TEST B in normal questions

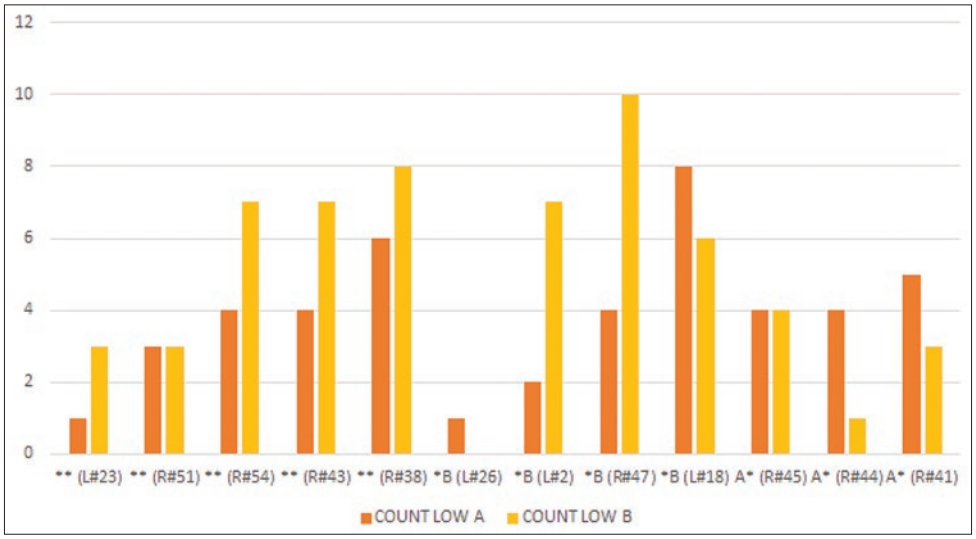


Figure 18 . Number correct answers in LOW third cohorts for TEST A and TEST B in problematic questions

answering the same question in Test A when the dominant distractor occurred before the correct option in both test versions. However, this trend is reversed and the students taking Test A scored as high or higher when the correct option occurred before the distractor in their version, but after in Test B. In both cases of clear and consistent trends, the questions are from the reading section, aside from one listening. In the pattern where the correct option occurs before the dominant distractor in Test B and not Test A, the students taking Test B score better than those from Test A. However, the trend is not as clear as in the other two patterns. It should be noted that in this pattern, there are three questions from the listening section and only one from reading.

In the questions that exhibit normal CTT behavior, there appears to be a slight trend for the students taking Test B to do better when the correct option occurs before the distractor in Test B but the opposite in Test A. Likewise, there is a similar slight tendency for students taking Test A to do better when the correct option occurs before the distractor in Test A but the opposite in Test B.

Performance between CTT thirds in TEST A and TEST B

Distribution of the CTT thirds in normal questions

The TOP, MID, and LOW thirds for the NON-PROBLEMATIC questions in both TEST A and TEST B have compatible distributions, as shown in Figure 20 below. In both test formats, TOP A thirds performed significantly better, suggesting that these questions provided better discrimination between thirds. There is a clear distinction in performance between the thirds, replicated for both TEST A and TEST B, with clear demarcations in terms of medians and standard deviation.

Distribution of the CTT thirds in problematic questions

Unlike the case for the NON-PROBLEMATIC questions, the TOP, MID, and LOW thirds for the PROBLEMATIC questions have a much higher standard deviation, and the variance between the medians of all the groups is not marked, aside from the LOW A third, which is dramatically lower than the corresponding LOW B third, with its median two full points lower. Overall, the graph of all the thirds, aside from LOW A, indicates that there is little significant discrimination between them in the problematic questions. The aberration of the LOW A group suggests that there can be significant differences in the impact of the change in order and sequence between the CTT thirds. That is, an effect may be evident in some of the CTT thirds, but absent from others, suggesting complex influences in option relationships.

Impact of Order and Sequence of Primary and Secondary Distractors

Although the general procedure is to change the position of all the correct options between TEST A and TEST B, there were three questions in which only the position of the distractors were changed. Interestingly, all three of these questions were flagged as problematic by our criteria. This gives

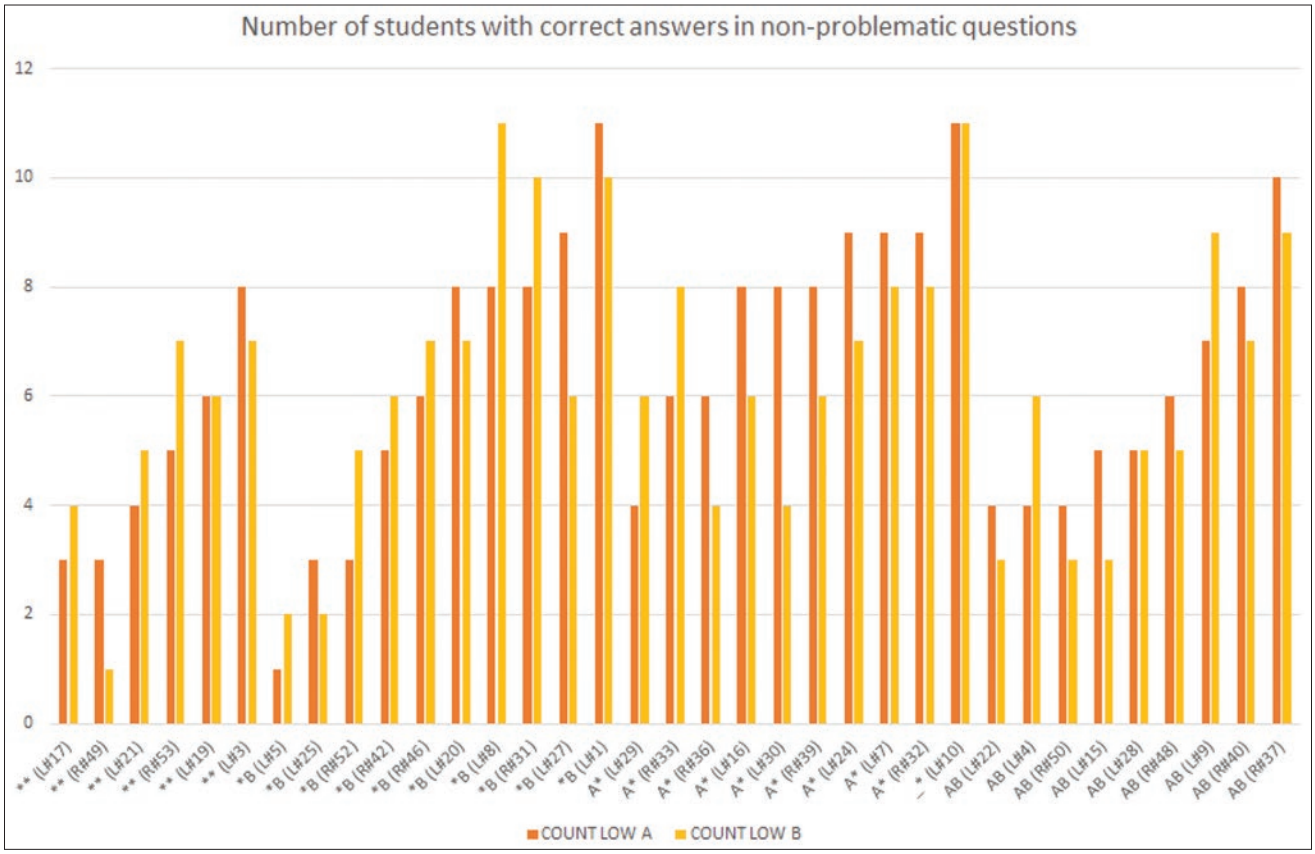
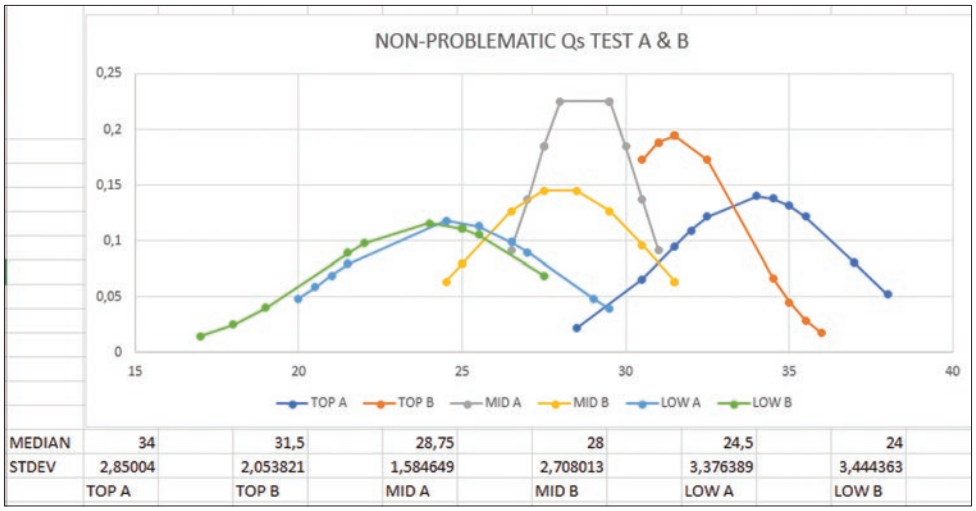
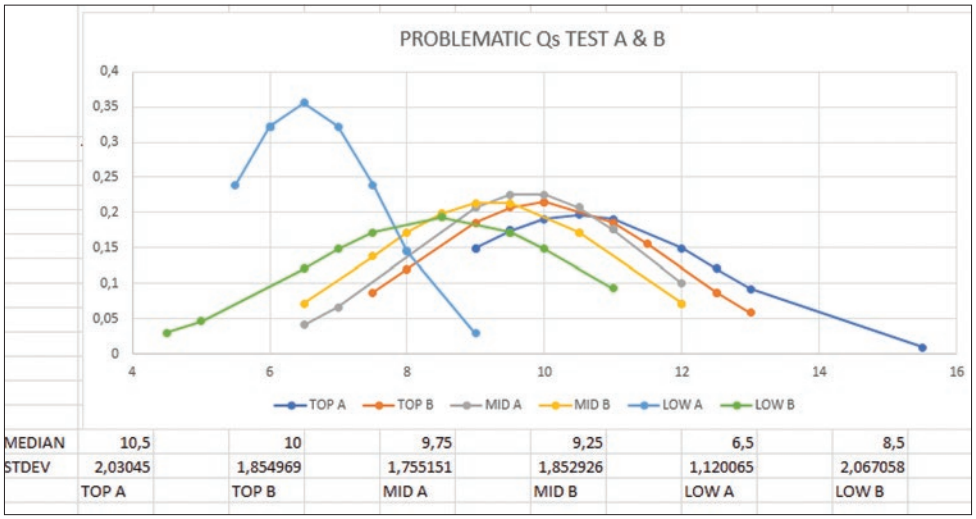


Figure 19. Number correct answers in LOW third cohorts for TEST A and TEST B in normal questions



MEDIAN	34	31,5	28,75	28	24,5	24
STDEV	2,85004	2,053821	1,584649	2,708013	3,376389	3,444363
	TOP A	TOP B	MID A	MID B	LOW A	LOW B

Figure 20. Normalized distribution of the CTT thirds in normal questions



MEDIAN	10,5	10	9,75	9,25	6,5	8,5
STDEV	2,03045	1,854969	1,755151	1,852926	1,120065	2,067058
	TOP A	TOP B	MID A	MID B	LOW A	LOW B

Figure 21. Normalized distribution of the CTT thirds in abnormal questions

us a unique opportunity to gain insights into the role and impact of changing the position of distractors while the correct option remains in the same position.

In Table 4 below, the complete CTT item analysis is given for separate cohorts for TEST A, TEST B, and the entire cohort. In addition, the specific item choices for each student in each of the TOP, MID, and LOW cohorts for TEST A and TEST B are shown and color-coded according to the correct option, the second and third choices according to the percentage chosen within each test version. Following the table are three graphs with a detailed discussion of three phenomena observed in MCQs that we have coined as 'sequence priming,' 'order domination,' and 'power of attraction.'

Order domination is when the position determines the likelihood of the student to be attracted to the option. This tends to be less dependent on the CTT third, i.e., order domination will affect all the students regardless of the CTT third to which they belong, although it can be a more significant factor for students in the LOW and MID thirds. Sequence priming and power of attraction are more dependent on language proficiency and tend to be a more significant factor for the TOP third. Sequence priming is when the distracting option influences or 'primes' the student to be aware of the qualities that will distinguish the correct option. The power of attraction depends on the level of language sophistication and awareness required to differentiate a distractor and correct option. For higher-level proficiency test-takers, the less nuanced the difference, the more powerful the attraction, while the more pronounced the difference, the less powerful the attraction. For lower-level proficiency test-takers, this is just the opposite. Of course, none of these are mutually exclusive, and all can impact to varying degrees the performance for students in all CTT thirds.

Sequence priming and power of attraction disproportionate effect on TOP third cohort

Figure 22 below demonstrates how the nature of the distractors and their position can have a disproportionate effect across the CTT thirds. In question 54, the correct option is in POSITION B in both test versions. Due to the marked difference in performance between TEST A and TEST B top third cohorts, it appears that CHOICE 2 is a weaker distractor than CHOICE 3. When CHOICE 2 occurs in POSITION A in TEST A, most of the students in that cohort appear to have understood not only that it is incorrect, but

Table 4. CTT analysis and choice selection of three questions with same correct option in both TEST A and TEST B

CTT item analysis

	#44	#51	#54
BOOKLET A	B	C	B
BOOKLET B	B	C	B
Sequence	A*	**	**
EASE A	44%	50%	47%
EASE B	38%	41%	50%
EASE ALL	41%	45%	48%
	~	~	~
DISC A	9%	36%	36%
DISC B	64%	0%	-36%
DISC ALL	33%	24%	0%
CORRECT TOP A	5	6	8
CORRECT LOW A	4	3	4
CORRECT TOP B	8	3	3
CORRECT LOW B	1	3	7
CORRECT ALL	26	29	31
EASE A-B	0,06	0,09	-0,03
DISC A-B	-0,55	0,36	0,73
TA-A	0,22	0,28	0,41
TA-B	0,44	0,22	0,47
TA-C	0,34	0,5	0,13
TB-A	0,31	0,25	0,38
TB-B	0,38	0,34	0,5
TB-C	0,31	0,41	0,13

Actual Choices

CTT Band	TOP THIRD						MIDDLE THIRD						LOW THIRD					
TEST version	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B
Question	44		51		54		44		51		54		44		51		54	
correct	B		C		B		B		C		B		B		C		B	
1st dist (%)	C	A	A	B	A	C	C	A	A	B	A	C	C	A	A	B	A	C
2nd dist (%)	A	C	B	A	C	A	A	C	B	A	C	A	A	C	B	A	C	A
S 1 in cohort	B	B	C	C	B	A	C	A	C	C	A	B	C	C	A	C	A	B
S 2 in cohort	B	B	C	C	B	A	A	C	C	C	A	A	B	A	A	C	C	C
S 3 in cohort	C	B	A	A	A	A	B	A	C	B	A	B	C	B	C	B	A	B
S 4 in cohort	A	C	A	B	B	A	B	A	C	C	B	A	B	C	B	A	A	A
S 5 in cohort	C	B	B	A	B	B	B	C	C	C	A	B	C	A	B	C	B	B
S 6 in cohort	A	B	A	C	B	B	B	B	C	A	B	B	B	A	B	A	B	B
S 7 in cohort	B	B	B	A	B	B	C	C	A	C	C	A	A	C	C	B	A	B
S 8 in cohort	B	B	C	B	B	A	C	B	B	B	A	B	A	C	A	B	A	B
S 9 in cohort	A	B	C	B	A	C	C	A	C	C	C	B	C	A	C	B	B	A
S 10 in cohort	B	C	C	B	B	C	B	B	A	C	B	C	A	C	B	A	B	A
S 11 in cohort	C	A	C	B	A	A							B	A	A	A	C	B

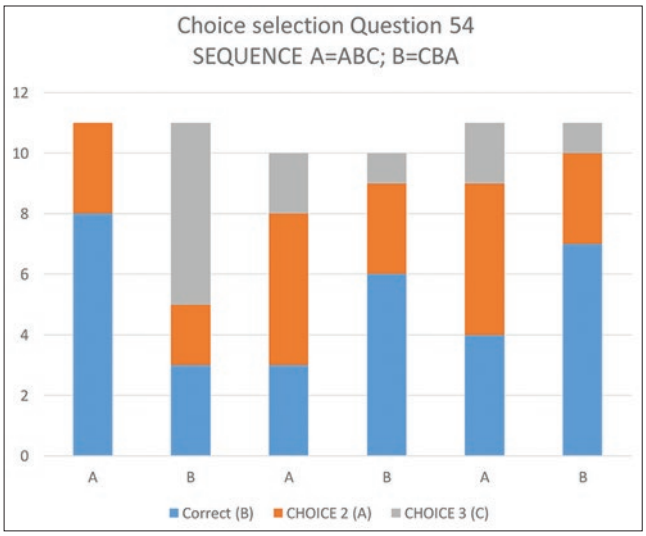


Figure 22. Sequence and order priming distortion in TOP third cohort

because the reason appears reasonably apparent to them, it also primed them to know what to look for in the correct choice. Therefore, because of this 'sequence priming', the majority of students were ready to accept the next choice in POSITION B as being correct without being concerned or distracted by CHOICE 3, which is in POSITION C. The 'sequence priming' was so strong, combined with order domination, that the 'power of attraction' of CHOICE 3 to distract the students was lost completely. However, when CHOICE 3 was in POSITION A in TEST B, the power of its attraction was much stronger than CHOICE 2. The students in TEST B appear to be sufficiently 'conflicted' about the distinction between it and the correct option. It seems that the 'power of attraction' in CHOICE 3 plus its 'order domination' (its appearance first in POSITION A before the correct option in POSITION B) persuaded more of the top cohort of students taking TEST B to choose it instead of the correct option. Having the language proficiency level to understand the subtle difference between CHOICE 3 and CHOICE 2 appears to be an influence in the top third cohort. In the MID and LOW thirds, the students appear to have less language knowledge to appreciate the discrimination between CHOICE 2 and CHOICE 3. As a result, the power of attraction of CHOICE 3 is diminished, while the power of attraction of CHOICE 2 is heightened. The effect of 'sequence priming' and 'order domination' is evident between TEST A and TEST B in the MID and LOW groups, but because the power of attraction of CHOICE 3 appears beyond their language proficiency level, the priming and order of attraction has just the opposite effect when compared to the TOP third cohort.

Sequence priming and power of attraction diminished effect on MID and LOW third cohort

The limited impact of sequence priming and power of attraction on the LOW and MID CTT thirds compared to the TOP third is quite evident in Figure 23 below. Concerning the LOW and MID thirds, the order of domination appears to come into effect to distinguish between them. The LOW third cohorts in both TEST A and TEST B appeared to give almost equal weight to the correct option in POSITION C with the distractors either in POSITION A or POSITION B. However, the MID third cohorts appeared to be influenced by sequence priming. Still, the priming affects the choice of distractor and not the correct option. While both TEST A and TEST B correctly chose the option in POSITION C, the sequence priming effect slightly magnified the order of domination, evident as CHOICE 2 and CHOICE 3 were more likely to be chosen when they occurred in POSITION A, and less likely when they were in POSITION B. In the TOP third cohorts, CHOICE 2 appears to have primed the TEST A group to distinguish the correct option in POSITION C from CHOICE 3 in POSITION B. The order domination of effect seems to have increased the power of attraction of both CHOICE 3 in POSITION A and CHOICE 2 in POSITION B for the TEST B group.

Sequence priming and power of attraction mirror effect on MID and LOW third cohort

The analysis of question 44 In Figure 24 below shows that the profile of the TOP third cohorts between TEST A and TEST B is reversed in the MID and LOW thirds. In the TOP thirds, the order of domination and power of attraction is clear when CHOICE 3 in TEST A appears in POSITION A. Although CHOICE 2 is in POSITION C in TEST A, it has a higher power of attraction than when it appears in POSITION A in TEST B. However, it appears to have a much greater sequence priming in TEST B, as the correct option has been selected much more, with CHOICE 3 in POSITION C almost neglected completely. Then, for the MID and LOW groups, this pattern is exactly reversed – the reversal is more evident in the MID group, but it is evident in the LOW group. In both TEST A and TEST B, the sequence priming of the correct choice in POSITION B, creating a higher power of attraction for CHOICE 3, which follows in POSITION C.

DISCUSSION

We embarked on this investigation after noticing contradictory results of CTT analyses of two versions of a METU EPE in which questions are identical except for the order of the options. About one-quarter of the 3-option MCQs exhibited abnormal CTT indexes in difficulty and discrimination. Instead of designing a specific research project to investigate this further, we gathered data from another EPE, administered at the end of the ETP course. This course is offered only at METU NCC and offered an ideal opportunity to study CTT analysis of an EPE with a highly controlled set of parameters. Because CTT analysis is entirely sample dependent, we needed to ensure there was minimal potential for discrepancies between the students taking the two different test versions. As we have shown in our description and analysis of the ETP students, the cohorts taking TEST A and TEST B performed within near-identical parameters. Consequently, we can confidently assert that differences in the CTT analysis of the questions in the two test versions are primarily due to the impact of the different order and sequence of options in MCQs.

As reported in the literature for MCQs in general, our study confirms that the difficulty of a question increases as the correct option moves toward the end of the options, which we have called 'order domination.' Our analysis revealed that the discriminating qualities of a question are also affected not only by the position of the correct option but also by the sequence of the options, which we have called 'sequence priming'. We have shown that the ability of a distractor to attract students to choose it instead of the correct

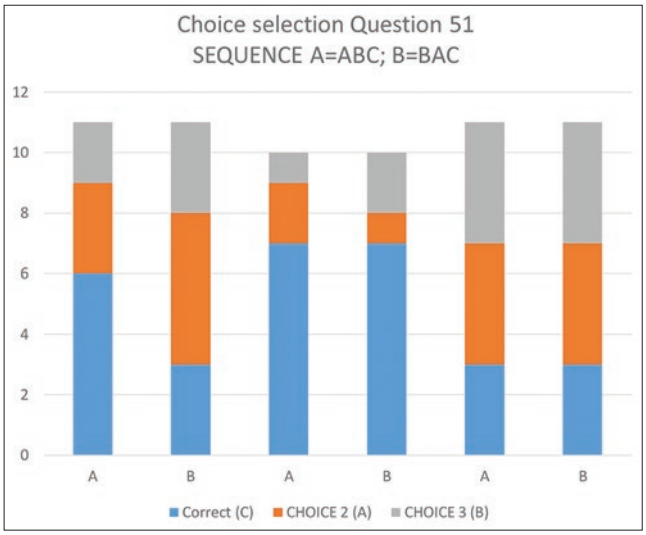


Figure 23. Cohort dependent tendencies for sequence, order priming and power of attraction

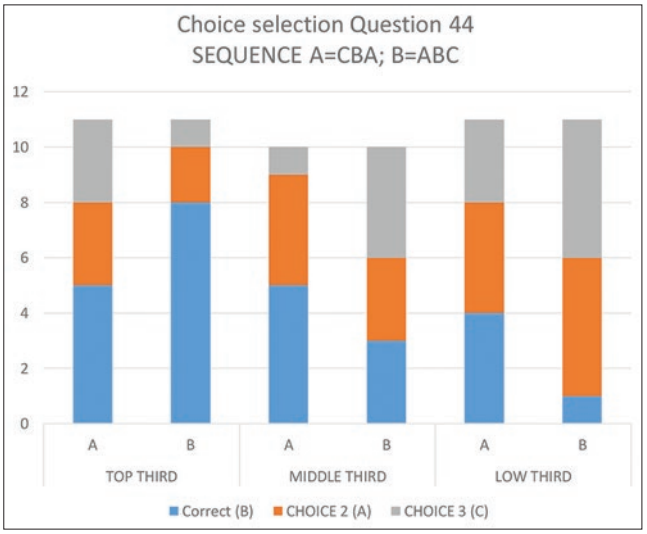


Figure 24. Mirror impact of TOP third on MID and LOW thirds in sequence priming, order domination and power of attraction

option, a quality we have called 'power of attraction,' can be accentuated or diminished by sequence priming. Sequence priming and power of attraction can have a significant influence on higher-level students but limited impact on the mid to low level students, while order domination is a more influential factor on the mid and low-level students.

Finally, within the context of a high-stakes test of English language proficiency, our study has shown that the impact of sequence priming, power of attraction, and order domination can vary between the two skills of listening and reading. In listening tasks, the test taker hears the listening text only once and cannot review what has been heard when choosing the correct option. The spontaneous nature of listening appears to limit the strength of sequence priming and power of attraction as the task gives the test taker little time for reflection and no opportunity to review. This finding is in keeping with the findings of Holzkenecht, et al. (2020) which is based on a different method of monitoring eye-tracking. Reading, on the other hand, has opportunity for review and reflection, so this skill is more prone to influence from all the three phenomena on MCQ choice selection, and this can manifest quite differently between the high and lower levels of students. In contrast to listening, where the question stem and options are simple in terms of language, in reading there is a higher cognitive load to read and understand the options in addition to reviewing and reflecting on several sentences that contain the clues to determine the correct answer.

Our research shows that creating multiple versions of a test by shuffling the order of option can negatively affect reliability, which is an essential concern about high-stakes exams. Although a CTT of the exam as a whole may suggest that the exam has an acceptable normalized distribution, the performance of students in the different thirds may be imbalanced. In our study, students in the LOW third who took TEST A were disadvantaged compared to the counterparts who took TEST B by 2 points over the 47 3-option questions, which would amount to 5% out of the total points in the test. Similarly, the students in the TOP third who took TEST B were slightly less discriminated from the LOW third when compared to their counterparts who took TEST A. In effect, the change in the order of options can make the outcomes of the question in one test version behave dramatically differently in the other test version. In our sample, this would call into question the reliability of about one-quarter of the 3-option questions.

While we focused solely on 3-option MCQs and the order and sequence of options within individual questions, we acknowledge that there are other influences at play. Test-taking strategies and algorithms employed for guessing may affect the outcomes of some of the questions we analyzed. For example, a test-taker may feel uncomfortable selecting the correct option from the same position in more than two questions in a row. Also, guessing may take into account factors such as length of option or differences in structures. Because our sample was small, there is a chance that some of these external factors may manifest

Finally, with 3-option MCQs, there are only six permutations of options possible. With a 4-option MCQ, the total number of permutations is 24, so it is likely to add a broader range of influence of sequence priming and order domination. While finding three meaningful distracting options may be difficult, the extra option in a 4-option MCQ, which may have limited value as a distractor, would nevertheless serve as a placeholder that could be used to create more variation within each pattern.

CONCLUSION

Creating reliable MCQs requires a clear test specification that covers the learning objectives which inform the nature of the question stem and guides the crafting of the correct option and the distractors. However, our research shows that beyond these essential building blocks, the test designer also must consider what order to arrange the options and the position of the correct choice. Holzkenecht, et al. (2020) is the only other study to show a signification correlation between the position of the correct response and the level of difficulty of the question in foreign language assessment of listening. Our findings suggest that not only the difficulty but also the discrimination in the assessment of listening and reading of any one test item can be manipulated simply by considering the principles of sequence priming, order domination, and power of attraction and arranging the order of the options accordingly. It is difficult to predict precisely how these factors will react with any specific group of students. Still, over time the test items could be evaluated and refined until the reliability of the configuration gives the expected outcomes, particularly in consideration of the other CTT thirds. Therefore, over some time, an item test bank can be built up with a broad set of questions that have an acceptable degree of reliability. In this scenario, a computer adaptive testing (CAT) framework drawing on categorized items from the test bank would be best to reduce the risk of cheating and maximizing the most accurate assessment of an individual's language proficiency.

Some institutions create a bespoke proficiency test for each iteration. In our institution's case, there are five unique EPE created each year. In such a case, individual question configuration cannot be refined to optimize reliability as the test items are not recycled. If the test is paper-based and two versions created by shuffling options, there is a possibility that up to 25% of the questions in either version could suffer from lack of reliability, and therefore put the overall assessment of any one student in question, which may vary further according to their level of proficiency. In such a context, where the desire to reduce the chance of cheating must be balanced by the lack of reliability that results from such measures, the CTT results of the exam should be carefully examined and compared between the two versions and any items that are flagged which have unacceptable CTT indexes for difficulty and discrimination discounted from the final result. Creating 4-option MCQs might ameliorate the effect of changing the order of options, as more permutations are available, and therefore more questions in each test version could have compatible option sequence patterns.

Much greater concern about our research may be relevant to computerized MCQ testing platforms, such as the MOO-

DLE QUIZ module. When administering a computerized MCQ test, typically by converting a traditional paper-based test, a single tick can set the parameter for the options in each question to be shuffled for each student. If the same set of questions are being used for all test-takers in the same order, it is conceivable that each student would see an entirely different test and could suffer unpredictable outcomes skewed by the influences of sequence priming and order domination for the unique combinations that they are presented with in accordance with their language proficiency level.

Given the findings of our research, and the unpredictability of the outcomes of an MCQs due to the influence of sequence priming, order domination, and power of attraction, using an MCQ on its own as a high-stakes assessment of language proficiency would appear liable to unfair or unreliable outcomes for individual students. If MCQs are needed because of the convenience of large-scale testing, it would be advisable to combine the MCQ test result with another source of assessment of language proficiency, such as their overall work and authentic performance throughout the course of study that leads up to taking the proficiency test. If the test-taker is coming from outside the institution with no acceptable record of language proficiency performance, on the spot authentic tasks can be assigned, such as an impromptu speaking task or a reading and summary task, to ensure that the student's language proficiency assessment is reliable and accurately describes their actual level.

FUTURE RESEARCH

In our study, we have shown that options in an MCQ do not exist in isolation; there is a relationship between the choices governed by their position and sequence in relation to each other. This relationship can be a factor in deciding not only the difficulty of a question, but also the discrimination in CTT indexes. Our domain was high-stakes English proficiency tests, so one area of further research would be to determine if order domination, sequence priming, and power of attraction are also evident in tests of proficiency in other languages. The influence of L1 could also be investigated, as our sample was drawn from young adult Turkish speaking students. Perhaps investigating speakers of different language groups may reveal an L1 variable. In addition, the extant research that suggests there is no significant difference between 3- and 4-option MCQs could be reviewed by considering the effects of the number of permutations in the order and sequencing of options on both difficulty and discrimination. Finally, our study looked at listening and reading, but another area of the use of MCQs in high-stakes exams is assessment of vocabulary, which draws on different cognitive skills and memory, depending on the nature of the stem.

REFERENCES

Bachman, L.F. (2004). *Statistical analyses for language assessment*. Cambridge University Press. https://doi.org/10.1017/CBO9780511667350

Bachman, L.F., Lyle, F. and Palmer, A.S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Brown, J.D. (2005). *Testing in language programs: a comprehensive guide to English language assessment*. McGraw-Hill College.

Cizek, G.J. (1994). The Effect of Altering the Position of Options in a Multiple-Choice Examination. *Educational and psychological measurement* 54(1), pp. 8–20. https://doi.org/10.1177/0013164494054001002

Davis, D.B. (2017). Exam question sequencing effects and context cues. *Teaching of Psychology* 44(3), pp. 263–267. https://doi.org/10.1177/0098628317712755

Gierl, M.J., Bulut, O., Guo, Q. and Zhang, X. (2017). Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. *Review of Educational Research* 87(6), pp. 1082–1116. https://doi.org/10.3102/0034654317726529

Haladyna, T.M., Downing, S.M. and Rodriguez, M.C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education* 15(3), pp. 309–333. https://doi.org/10.1207/S15324818AME1503_5

Hambleton, R.K., Traub, Ross E. and Traub, Rose E. (1974). The Effects of Item Order on Test Performance and Stress. *The Journal of Experimental Education* 43(1), pp. 40–46. https://doi.org/10.1080/00220973.1974.10806302

Hohensinn, C. and Baghaei, P. (2017). Does the position of response options in multiple-choice tests matter? *Psicológica*.

Holzknecht, F., McCray, G., Eberharter, K., Kremmel, B., Zehentner, M., Spiby, R., & Dunlea, J. (2020). The effect of response order on candidate viewing behaviour and item difficulty in a multiple-choice listening test. *Language Testing*, https://doi.org/10.1177/0265532220917316

Marcus, A. (1963). The effect of correct response location on the difficulty level of multiple-choice questions. *The Journal of applied psychology* 47(1), pp. 48–51. https://doi.org/10.1037/h0042018

McNamara, W.J. and Weitzman, E. (1945). The effect of choice placement on the difficulty of multiple-choice questions. *Journal of educational psychology* 36(2), pp. 103–113. https://doi.org/10.1037/h0060835

Messick, S. (1996. Validity and washback in language testing. *Language Testing* 13(3), pp. 241–256. https://doi.org/10.1177/026553229601300302

Mosier, C.I. and Price, H.G. (1945). The arrangement of choices in multiple choice questions and a scheme for randomizing choice. *Educational and psychological measurement* 5(4), pp. 379–382. https://doi.org/10.1177/001316444500500405

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

Oruç Ertürk, N. and Mumford, S.E. (2017). Understanding test-takers' perceptions of difficulty in EAP vocabulary tests: The role of experiential factors. *Language Testing* 34(3), pp. 413–433. https://doi.org/10.1177/0265532216673399

Rodriguez, M.C. 2005. Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research. *Educational Measurement: Issues and Practice* 24(2), pp. 3–13. https://doi.org/10.1111/j.1745-3992.2005.00006.x

Sadeghi, K. and Masoumi, G.A. (2017). Does number of options in multiple choice tests affect item facility and discrimination? An examination of test-taker preferences. *Journal of English Language Teaching and Learning*, *9*(19), 123-143.

Satti, I., Hassan, B., Alamri, A., Khan, M.A. and Patel, A. (2019). The effect of scrambling test item on students' performance and difficulty level of mcqs test in a college of medicine, KKU. *Creative Education* 10(08), pp. 1813–1818. https://doi.org/10.4236/ce.2019.108130

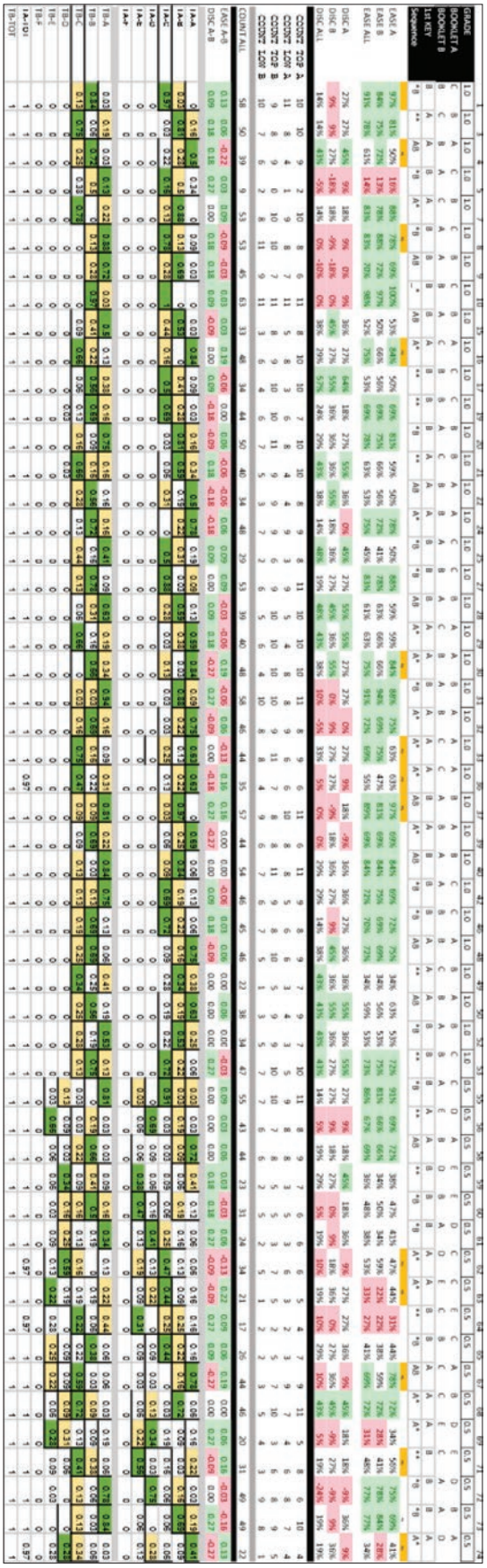
Shin, J., Bulut, O. and Gierl, M.J. (2019). The Effect of the Most-Attractive-Distractor Location on Multiple-Choice Item Difficulty. *The Journal of Experimental Education*, pp. 1–17. https://doi.org/10.1080/00220973.2019.1629577

Shizuka, T., Takeuchi, O., Yashima, T. and Yoshizawa, K. (2006). A comparison of three- and four-option English tests for university entrance selection purposes in Japan. *Language Testing* 23(1), pp. 35–57. https://doi.org/10.1191/0265532206lt319oa

Stout, D., & Heck, J. (1995). Empirical Findings Regarding Student Exam Performance and Question Sequencing: The Case of the Cumulative Final. *Journal of Financial Education, 21*, 29-35.

Tellinghuisen, J. and Sulikowski, M.M. (2008). Does the Answer Order Matter on Multiple-Choice Exams? *Journal of chemical education* 85(4), p. 572. https://doi.org/10.1021/ed085p572

APPENDIX



Appendix 1. CTT Item Analysis for MCQs with Normal Behavior