Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 13, No. 2, October 2023
ISSN 2476-5880

Investigating Different Kinds of Stems in Multiple-Choice Tests: Interruptive Vs. Cumulative

Sharareh Sadat Sarsarabi[1]*, Zeinab Sazegar[2]

Received: 19 April 2023 Accepted: 13 Jun 2023

Abstract
The statement stated in a multiple-choice question can be developed regarding two types of sentences: Interruptive (periodic) and cumulative (or loose). This study deals with different kinds of stems in designing multiple-choice (MC) items. To fill the existing gap in the literature, two groups of teacher students passing general English courses at Farhangian University were selected based on Cambridge Placement Test. The design of this study was a comparison group design. To verify the effectiveness of the stems, i.e., interruptive and cumulative stems, two types of tests based on the book entitled Thoughts and Notions 2, which was taught in General English classes, similar in content, but different in their stems, were designed. Each test contained 40 items, 25 vocabulary items, and 15 items of reading comprehension. The first group of students was given the test designed using only interruptive sentences as stems. The second group participated in the test, being prepared using only cumulative sentences as stems. After the data analysis via an independent t-test, it became apparent that the first group outperformed the second. Therefore, it was concluded that interruptive sentences as a stem in multiple-choice tests were more reliable and valid than cumulative ones. One of the study implications is that the interruptive stems can be used to assist policymakers and material designers, and language teachers to be considered for future decision making, and designing materials.

Keywords: Cumulative stems; general English course; interruptive stems; multiple-choice tests

1. Introduction

Assessment is essential at all stages of education, primarily for higher education, where university admission criteria become competitive and demanding. Even though assessment is defined differently for different goals, it is self-evident that the new approaches to evaluation have a significant impact on learning (Sudakova et al.,2022). Language testing is always considered an inseparable part of language learning and teaching, regardless of the kind of testing theory which is taken. Also, testing development as a critical multidimensional task (Polat, 2020) has always been changing accordingly to that of linguistics (Mao, 2022). Assessment development as a critical and dynamic process fosters revision, change, and improvement of different dimensions of language education and that is why testing influences educational practices and determines learning goals and curriculum in various ways (Cheryl et al., 2017).

[1] Assistant professor, Department of English, Farhangiyan University, Mashhad, Iran, Email: dr.sarsarabi@cfu.ac.ir
[2] Department of English, Farhangian University, Mashhad, Iran, Email: Zeynab_sa@yahoo.com

Generally, tests are at hand choices for teachers who wish to assess learners favored proficiency levels in a particular course where the considered ability is measured via test items (Brualdi Timmins, 1998). As a result, tests are traditional assessment methods. The desire to assess pre-thought educational outcomes utilizing methods that are more reliable and valid has resulted in the production of different test techniques in second and foreign language assessment (Polat, 2020). These assessment techniques are classified based on cognitive and affective factors, and also their test preparation methods like multiple-choice tests, open-ended questions, matching questions, true/false sentences, and completion questions (Black et al., 2003). It is clear that these question types have some advantages and disadvantages, and each of them would be appropriate for specific educational objectives (Cheryl et al., 2017).

Having global popularity, preference, and widely used factors in mind, multiple-choice questions were considered in this study. Generally, multiple-choice tests (MCTs) are widely used to examine learners' foreign language proficiency, specifically in high-stake tests. Multiple-choice tests are also utilized to assess knowledge outcomes and different kinds of learning outcomes. They are most broadly used for assessing knowledge, comprehension, and application outcomes (Golvardi Yazdi et al., 2021). MCTs are preferred types of tests within courses of study and entrance examinations of different institutions (Ibbett & Whldon, 2016).

They are types of tests that provide the answer among their items and do not require the test takers to answer but choose the correct option from a set of alternatives (Güler, 2017). These tests are not expensive, rapid, and facile for administration but lack the possibility of measuring abundant cognitive skills (Brown, 2005). Many argue that MCTs can be efficient if they save time and provide faster feedback for students and effective if they provide reliable and valid results (Bacon, 2003; Douglas et al., 2012; Glass & Sinha, 2013; Nicol, 2007). One of the main disadvantages of MCTs is the probability of guessing the correct answer, which reduces its reliability, validity, and effectiveness (Golvardi Yazdi et al., 2021). Dávid (2007) took into account several reasons for the relative unpopularity of the recent research on the MC format. MC tests are generally regarded as a partial, limited, and decontextualized form of assessment. Accordingly, inadequate for many testing situations since they could not measure all parts of the construct. Another deficiency of MC formats is related to the difficulty of writing correct and acceptable MC items. Despite the shortcomings, many educational systems rely heavily on multiple-choice tests to fulfill their assessment demands due to the large number of test takers, the need for fast scoring, and the convenience and reliability of multiple-choice tests (Currie & Chiramanee, 2010).

In a multiple-choice question, which is composed of a stem, the best answer, and the distractors, the stem is the question that can be a problem or an incomplete statement. The statement stated in a multiple-choice question can be developed regarding two types of sentences: Interruptive (periodic) and cumulative (or loose). Interruptive or periodic sentences are complex sentences that bring the main idea of the sentence at the end of it, having the dependent clauses in front of the main independent clause, and their purpose is to encourage the readers to read the complete sentences to understand their main idea (Ward & Murray-Ward, 1999). On the other hand, cumulative or loose sentences put the main clause at the beginning of the sentence, and the dependent clause goes after. It has the aim of creating the

effect of immediacy and naturalness. It's important to note that English writers utilize cumulative sentences much more often than interruptive ones (Cunningham, 1998).

This study is considered to be among the few attempts to fill the gap of research in the investigating different kinds of stems in multiple-choice tests. In the absence of probing the value of these two types of stems, the aim of this article is to probe the value of interruptive vs. cumulative stems in multiple-choice tests by verifying the effectiveness of the stems using two types of tests. It is expected that the results help language test developers to choose test stem types that are more effective and help test takers to perform better.

2. Review of Literature

Due to the popular view of structuralism in language instruction, McNamara (2003, p.10) stated, "a tendency to atomize and decontextualize the knowledge to be tested and to test aspects of knowledge in isolation". With the application of psychometrics to language testing, a new approach to testing, known as discrete point testing, appeared (Mao,2022). In accordance with the discrete point testing, there must be only one testing point in each item, and the tests of grammar would be separated from other language components; Furthermore, materials in the tests were presented with minimal context, and to suit these characteristics, multiple-choice was selected as the main item type for the testing, and there was always an adequate sampling of the items to achieve validity (Mao, 2022).

Sigott (2004) and Spolsky (1978) considered three phases for the history of language testing, namely the pre-scientific phase, psychometric-structuralist phase, and sociolinguistic phase. In the pre-scientific phase of language testing, in order to provide the validation of tests, testing specialists tried to answer one question: "Does the test look as if it measures what it purports to measure?" Positive answer of testing experts, test users or test takers to this question proved the validity of the question test. Nowadays, according to Sigott (2004), the validity of tests involves two perspectives. On the one hand, it is argued that since this kind of validity (validity in the pre-scientific phase of language testing) has no theoretical foundations, it may result in highly idiosyncratic and subjective evaluations of a test by different people. On the other hand, in communicative testing, considerable attention is paid to this concept, and the emphasis is to represent real-life language use in language test tasks as much as possible.

The study of Baghaei and Dourakhshan showed that dichotomously scored double-response MC items were significantly harder than their single-response MC counterparts. Double-response items had equal reliability to the single-response items and had a better fit to the Rasch model. Polytomously scored double response items, however, were easier than single-response MC items but substantially more reliable. The study goal of Kaddoura and Al Husseiny (2021) was to identify students' satisfaction with using Nearpod as a tool for interactive learning in higher education. The responses were collected through an online questionnaire. Analysis of students' responses shows their interest in using online learning tools in class. Analysis of students' responses proves the necessity to change teaching styles in higher education to be more interactive and increase students' engagement in the lesson.

Rashidi and Safari (2014) examined the effect that various test formats used for testing grammar have on the measurement of the trait. Specifically, they investigated the characteristics of a special type of multiple-choice (MC) test, called multitrak, in comparison

to the properties of the standard MC test and the constructed-response (CR) test. The results indicated that each of the testing methods yielded different degrees of difficulty for the test takers, with the standard MC test being the easiest and the multitrak test being the most difficult. In other words, the study found that the multitrak test was easier than the CR test and more difficult than the standard MC test. Accordingly, in a multitrak test, the test taker gets involved more on the higher levels of grammar; so, the focus of multitrak items goes beyond the syntactic level to come close to semantic and pragmatic levels, whereas a standard MC test mostly covers the narrow concept of grammar which is in syntactic level.

2.1. Multiple-Choice Tests

Multiple-choice tests have dominated educational assessment in various parts of the world. Leaning achievements are related to multiple-choice testing formats because giving immediate feedback is better than other traditional delayed feedback formats (Wood et al., 2022). They are sometimes used at universities to gain a high evaluation efficiency (Giusto et al., 2019). Butler (2018, p.1) mentioned that "a multiple-choice item consists of a stem (the content, context, or question the test-taker is required to answer) and a set of possible responses, only one of which is correct (the other responses are commonly referred to as lures or distractors)." Research literature suggests including a minimum of 3 answer choices. Considering the popularity and practicality of multiple-choice tests, it is not weird that a great number of studies have been performed to find the best ways to develop and utilize them (Haladyna et al., 2004). Many of these studies investigated the ways through which the reliability and validity of such tests can be enhanced (Butler, 2018).

According to the findings of Esmaeeli et al. (2021), the psychometric properties of three-choice questions are similar to four- or five-choice questions, and the validity and reliability of the test or the coefficients of difficulty and differentiation do not change significantly with decreasing the number of options. Therefore, reducing the number of questions can reduce the time needed to design tests and take exams, saving the time and energy of the faculty and students. Most studies have concluded that it is cost-effective to use a three-choice question if it does not change the psychometric properties of the test by reducing the number of options.

2.2. Interruptive vs Cumulative Multiple-Choice Tests

MCQs consist of two elements: the stem and the suggested responses (Giusto et al., 2019). According to Pongweni (2017), an interruptive sentence suspends the meaning of grammatical structure until the end of it in order to trigger and keep curiosity in the reader, and to motivate them to go on reading till its pick ending. He defines the cumulative sentence as "the one in which clauses go after one another and the main idea is put at the beginning of the sentence with the aim of bombarding and posing the risk of drowning the reader" (p. 22). Accordingly, cumulative and interruptive stems differ regarding the place where the answer has been omitted. The examples below extracted from the book entitled Thoughts and Notions 2, written by Ackert and Lee (2005), reveal the difference between these two types of stems.

2.2.1. Interruptive or Periodic Stems

Example 1: Sometimes your face gets red when you feel …….

a. unknown b. embarrassed c. famous d. dangerous

Example 2: Will there be enough food for all people, or will we have food……?

a. extinct b. shortage c. credit d. absence

2.2.2. Cumulative or loose stems

Example 1: Cattle are …… in my part of China, so there are no dairy products there.

a. special b. religious c. heated d. rare

Example 2: He …… a risk when he jumped into the ocean to save the child. He could die.

a. made b. took c. did d. received

This study deals with interruptive and cumulative stems and their desirable effects on the reliability and validity of such tests. No studies were found that investigated the effects of the stems on the reliability or validity of multiple-choice items, which is the thrust of this study. Therefore, the present study gained its significance.

To check the effectiveness of either of these two types of stems, this study poses and addresses the research question:

1. Do interruptive stems multiple-choice items measure students' competence of vocabulary, and reading comprehension more accurately than those of the cumulative ones?

Since research has not provided us with any definite answers to the question of stems and their role in promoting the reliability and validity of our classroom tests, the above questions are changed into its null hypothesis:

H.0: Cumulative multiple-choice items measure students' competence in vocabulary, and reading comprehension as accurately as interruptive ones.

3. Method

3.1. Research Design

This study was concerned the effect of two types of stems, interruptive and cumulative stems, on multiple-choice tests. The design of this study was a comparison group design that evaluators have frequently used to assess program impacts (Henry, 2015). Sixty university teacher students took part as participants in two groups (each group of 30 teacher students). Group comparison research involves comparing the mean scores of two or more groups of research participants on one or more dependent variables (Whitley, 2002).

3.2. Participants

To investigate the effect of cumulative and interruptive stems on measuring students' competence via multiple-choice items, sixty teacher students from Farhangian University participated in this study. The teacher students were all female. Their average age was 20 ranging from 19 to 22 years. They were studying at the end of the first semester of 2022 of the academic year. The student's mother tongue was Persian, and they were all majoring in Educational Science. After taking the CPT, these teacher students were known to be homogeneous. They were divided into two homogeneous groups: They have divided into two homogeneous groups: 30 participants belonged to the interruptive stems group, and 30 participants belonged to the cumulative stems group based on the type of multiple-choice item tests given to them by the researchers.

ISSN 2476-5880

3.3. Instruments

3.*3.1. Cambridge Placement Test (CPT)*. A CPT was run on a group of 70 female teacher students to homogenize the members according to their proficiency level. CPT is a 25-question online English test. The researchers administered this test to teacher students at Farhangian University in Mashhad. Those members who have located one standard deviation above and below the mean were chosen to take part in this research. Then according to their scores on the CPT, sixty teacher students were selected for this study. The reliability index of CPT was assessed by Kuder-Richardson formula 21 as 0.86. Thus, the CPT indicated an acceptable index of reliability (Kline, 2000).

3.3.2. Vocabulary and Reading Comprehension Pre-test and Post-test. To measure the effectiveness of the two types of multiple-choice item tests of the study, the researcher-made test consisting of 40 multiple-choice items was employed as a pre-test and post-test. Two types of multiple-choice item tests were prepared using two types of stems: interruptive multiple-choice items and cumulative multiple-choice items. Both tests were based on the book entitled Thoughts and Notions 2. The same information was reflected in both of them. Each of the tests consisted of 40 items with two sub-tests as well: vocabulary and reading comprehension, respectively. This test had 25 vocabulary and 15 reading comprehension items. The first group was prepared using only interruptive stems (See Appendix A), and the second one was prepared using only cumulative stems for multiple-choice items (See Appendix B). The content validity of the tests was determined by experts in the field of EFL teaching. However, to check the reliability of the tests, the researchers piloted the tests before the main administration at two Farhangian campus in Mashhad based on convenience sampling. After piloting the tests on forty teacher students different from the participants of this research, the tests' reliability was estimated by Cronbach Alpha Coefficient (α).

Table 1
Reliability of the Pre and Post-test

Instruments	Number of teacher students	Number of items	R
Pre and Post-test	40	40	0.820

As displayed in Table 1, the pre and post-test showed a reliability of 0.82. The pre and post-test indicated an acceptable index of reliability (Kline, 2000). Thus, this test is regarded as a reliable and valid test.

3.4. Data collection procedures and analysis

As the primary step within the present research, the researchers piloted the tests on 40 teacher students different from the members of this study before the main administration. The second step was the selection of the participants based on the CPT. The researchers developed a CPT. The instructor, first of all, gave the test to teacher students to see whether they were homogeneous or not. After administering the test, the two groups were known to be homogeneous, with the means of 29.10 and 30.10 for the first and second groups, respectively.

Both groups were taught by the same instructor, in the same class, for two sessions each week during the same instructional semester. The data were analyzed, then sixty teacher students were selected for the objective of this investigation.

Following the CPT, two other types of multiple-choice item tests were developed, each of which consisted of 40 items. These tests had different formats, but their content was the same. Next, participants were assigned to interruptive stems and cumulative stems groups, with 30 teacher students in each group. In the first type of the test, only interruptive stems were used to measure the students' competence, while in the second type, only cumulative stems were used to do this. Teacher students were in advance told that their participation was voluntary and the tests would not affect their course grades. They were also told that the purpose of the study was to examine the types of answers given on the test to see if there was a difference in the way questions were approached.

The researchers gave the interruptive stem multiple-choice items to one of the groups and the cumulative stem multiple-choice items to the other group. It is worth mentioning that the two groups were supposed to answer the questions for 30 minutes. Both interruptive and cumulative stem multiple-choice items were administered simultaneously. Therefore, the time and environmental variables were the same for both groups. After administering and scoring the tests, the data were collected by carefully studying the answers on the answer sheets. Having the data collected, the researchers processed the data using the statistical package for social sciences (SPSS) version 25. Then, to compare the results and establish the differences, the independent t-test was used to determine the differences between the two groups.

4. Results

A T-test was administered for the interruptive stems and cumulative stems groups before the treatment to determine the homogeneity of the two groups. The results obtained were collected and registered in Table 2 as follows:

Table 2
T-Test for Both Groups' Performance on the Proficiency Pre-Test

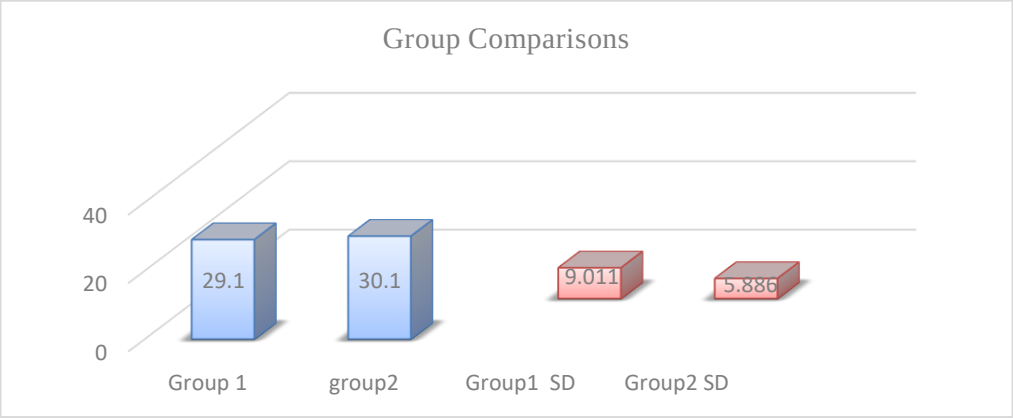
Groups	Mean	SD	N	Df	t-observed	critical t-value
Group 1	29.10	9.011	30	29	0.509	2.00
Group 2	30.10	5.886	30	29		
Total			60	58		

As the above results indicate, the difference between the two groups is not meaningful. That is, the t-observed is much smaller than the critical t-value at the $p< 0.05$ level of significance. Based on these results, it can be concluded that both groups are nearly homogeneous. Figure 1 illustrates the means for both groups as follows:

ISSN 2476-5880

Figure 1

Comparison between the Performance of Group 1 and Group 2 on the Proficiency Pretest (Mean ± SEM)



Following the pre-test, all the mentioned participants were told to study unit one of Thoughts and Notions carefully. Two weeks later, they were given two types of multiple-choice item tests which were the same for their contents but different for their stem's formats. In other words, the first test was designed by using only interruptive stems, and the second one was designed by using only cumulative ones. Each test sample contained the same number of items, i.e., 40 questions: 25 items of vocabulary, and the last 15 items of reading comprehension, which were equal to the other one in number. Group one (N= 30) participated in the multiple-choice items with interruptive stems. Group two (N= 30) participated in the multiple-choice items with cumulative stems.

As mentioned, both groups favored similar conditions during testing administrations. Following the administration of both cumulative and interruptive stem multiple-choice items, these results were gained. Based on the data in Table 3, another t-test was run. In other words, if the calculated t-test exceeded the critical t-value (2.00) at the 0.05 level of probability for d.f.=58, the null hypothesis might be rejected; otherwise, it might be contributed to other factors. The results of the second independent t-test have been illustrated in Table 3 as follows.

Table 3

T-Test for Both Groups' Performance on the Interruptive and Cumulative Stem Multiple-Choice Items

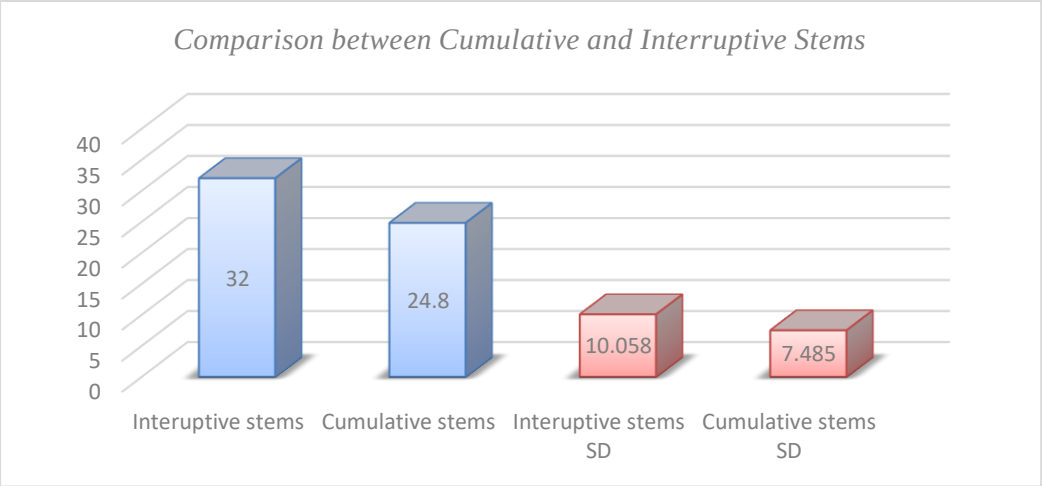
Stems	Mean	SD	N	Df	t-observed	critical t-value
Interruptive Stems	32	10.058	30	29	3.145	2.00
Cumulative Stems	24.8	7.485	30	29		
Total			60	58		

According to Table 3, for the cumulative stems group, the calculated mean and the standard deviation were respectively 24.8 and 7.485, and for interruptive stems group, they were respectively 32 and 10. 058. The t-observed were 3.145. The means for both interruptive and cumulative stems have been illustrated in Figure 2 as follows:

ISSN 2476-5880

Figure 2

Comparison between Cumulative and Interruptive Stems on Both Groups' Performance (Mean ± SEM)



Considering (t-observed 0.509 < critical t-value =2.00) at the pre-test stage and (t-observed = 3.145 > critical t-value = 2.00), a remarkable difference came about between the two groups at the post-test stage.

After comparing the two mean scores through independent sample t-test calculations, the researchers felt justified and safe that the null hypothesis could be rejected; the two groups scored differently on the post-test, and the difference was statistically significant. The researchers' interpretation was that interruptive stems had been proven to be effective. The two groups were not significantly different at the outset of the study; however, they performed differently on the final test. Therefore, it seems to justify the idea that interruptive stems have served the intended purpose. The researchers are satisfied to state that the final calculated t-test (3.14) at the $p<0.05$ level of probability is due to the independent variable (interruptive stems). That is, the group that participated in the multiple-choice stems being designed by using only interruptive stems outperformed the other group participating in the cumulative stem multiple-choice test. The researchers' interpretation is that interruptive stems can measure the test takers' knowledge of vocabulary and reading comprehension more accurately than those of the cumulative ones.

5. Discussion

The current study aimed to compare the efficiency of interruptive stems with cumulative stems in multiple-choice items. After comparing the two mean scores through t-test calculations, the researchers felt quite justified and safe that the null hypothesis could be rejected; the two groups scored differently on the posttest, and the difference was statistically significant. This finding is in contrast with the study of Yanagawa and Green (2008) which indicated no significant difference, caused by test items, on learners' performance in a listening test. One possible explanation for this inconsistency might be the skills the test items have intended to measure. The researchers' interpretation was that interruptive stems had been

proven to be effective. The two groups were not significantly different at the outset of the study; however, they behaved differently on the final test. The researchers' interpretation is that interruptive stems can measure the test takers' knowledge of vocabulary and reading comprehension more accurately than those of the cumulative ones. For this reason, it seems that a test containing such types of stems can be both valid and reliable. This study is considered to be among the few attempts to fill the gap of research in the construction of MC item writing.

From a theoretical perspective, the findings support different functions interruptive and cumulative sentences as the stems of multiple-choice items (Pongweni, 2017). The effectiveness of interruptive stems compared to cumulative stems can be justified by the motivating role of these items. In effect, as earlier studies have shown, interruptive stems have the potential to suspend the meaning of each grammatical structure until the end of a stem with the aim of motivating the learners and increasing their curiosity to read the whole stem (Ackert & Lee, 2005; Pongweni, 2017). Accordingly, the learners would have a more detailed understanding of test items. In addition, the less-effective role of cumulative stems is in support of previous research which indicates cumulative items enhance the risk of drowning the students (Pongweni, 2017). Accordingly, cumulative and interruptive stems differ regarding the place where the answer has been omitted.

In sum, the present study lends more credence to multiple-choice testing as an enduring, objective, and efficient form of educational assessment (Haladyna, 2004). Yet, this study highlights the importance of writing efficient stems that need the learners read the whole item in order to perform well. Complementing the findings of previous studies which have focused on the development of distractors and the number of options (Shin et al., 2019; Thanyapa & Currie, 2014), the current study implied that developing stems can also exert a significant influence on learners' performance on a multiple-choice formatted test. Accordingly, test developers are suggested to consider this issue and its impact on learner' cognitive load during a multiple-choice formatted test.

6. Conclusion

The primary purpose of the stem of the multiple-choice items is to present the problem clearly and concisely (Heaton, 1988). Test takers should be able to obtain a complete idea or thought from the stem and then answer the question. At the same time, the stem or lead of the multiple-choice items should not contain irrelevant or extraneous information, which makes test takers go astray while selecting the correct answer. That is, it should not make the test takers beat around the bush (Jones, 2021).

This study sought to examine the value of interruptive versus cumulative stems in multiple-choice tests by verifying the effectiveness of the stems using two types of tests. To reach this goal, two groups of student teachers studying at Farhangian University were given these two types of tests. The result of the analysis showed that those who participated in the interruptive stem multiple-choice items outperformed the other participants who took the test with cumulative stems.

In the same vein, the present study focused on two types of stems, which can be useful and effective in writing multiple-choice items - interruptive and cumulative stems. Since the t-observed was much greater than the critical t-value, the null hypothesis, i.e., stem types do not

have any effect on the accuracy of measurement of the multiple-choice items, was rejected. Therefore, it could be cogently argued that by using interruptive stems rather than cumulative ones, test writers can enhance the validity and reliability of the multiple-choice items-two important characteristics of a good test. It was supposed that interruptive stems would measure the test takers' knowledge of vocabulary and reading comprehension more accurately than those cumulative stems. Interruptive stems can give the test takers a complete idea or thought of the problem at the end of the stem. Aiken (1987) suggests that test writers should place the blank in an incomplete statement at the end. Interruptive stems, too, have their blanks at the end or nearly the end of the stem. For this reason, test takers can have a complete idea or thought of the problem before selecting the right or best answer.

Cumulative stems, on the other hand, have their blanks at the beginning or nearly the beginning of the stem. Such stems cannot provide test takers with a complete idea or thought of the problem before selecting the right or best answer. Therefore, they are not as appropriate as the interruptive stems are in constructing multiple-choice items. Of course, it seems that the scope of this study covered only two types of stems-interruptive and cumulative stems of the multiple-choice items.

It seems that placing the blanks of the stems at the end or very end can be very effective on the reliability and validity of multiple-choice tests. Therefore, test writers are advised to use interruptive stems in planning multiple-choice incomplete statement stems. The same case has been observed in TOEFL tests. Having carefully studied 40 items of Model Test One (structure and written expression part) of the TOEFL, the researchers noticed that nearly 90% of such stems have their blanks at the end.

Developing good MCQs to test learners' knowledge and hinder test takers from simply guessing the correct answer is challenging and time-consuming (Giusto et al., 2019). Besides considering some points when writing multiple-choice questions, such as avoiding purposeless or negative stems, which may lower the reliability and validity of the tests (Aiken, 1987), careful attention to the language of items to ensure that the reading level is appropriate to the test-takers for whom they are designed (Gaspar, 2004), the suggestion of this paper is to motivate test developers to use interruptive stem multiple-choice rather than cumulative ones to gain better results. Clearly, if test writers do not observe the above-mentioned points, they may violate the practicality principle which is an important characteristic of a good test (Bachman & Palmer, 2022).

Last but not least, the decision as to what point to raise in the stem and to word the choices is left to the discretion of the test writers (Jafarpur, 2003). If they want to get better results, they should word the stems of the multiple-choice items concisely so that there remain no ambiguities of the problem for the test takers.

The interruptive stem multiple-choice items' outperformance in this study may be used to advance the existing body of knowledge in the field of assessment and planning. It may help to determine how MCs are perceived by teachers to be one of the factors that correlate with their other latent variables. Individuals who are in charge of designing tests and policymakers should pay more attention to the various components of test designing and their potential effects on learners.

One of the limitations of this study is the fact that this study centered on teacher students. Future research could validate the instruments within other university settings and contexts to enhance the generalizability of the findings or to find the differences. Furthermore, the ideas and the concepts in EFL teaching are dynamic; thus, it would be developed with other participants and different statistical populations in various educational contexts.

Declaration of Conflicting Interests

We declare no potential conflicts of interest with respect to this study.

Funding

The authors received no financial support for this research.

References

Ackert, P., & Lee, L. (2005). *Thoughts and notions (2)*. Thomson Heinle.

Aiken, L. R. (1987). Testing with multiple-choice items. *Journal of Research and Development in Education, 20*, 44-58.

Bachman, L., & Palmer, A. (2022). *Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford University Press.

Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context. *Journal of Marketing Education, 25*(1), 31-36.

Baghaei, P., & Dourakhshan, A. (2016). Properties of single-response and double-response multiple-choice grammar items. *International Journal of Language Testing*, *6*(1), 33-49.

Black, P., Harrison, C., Lee, C., Marshal, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice.* McGraw-Hill Education.

Brown, J. D. (2005). *Testing in language programs*: *A comprehensive guide to English language assessment.* Prentice-Hall.

Brown, H. D., & Abeywickrama, P. (2004). *Language assessment: Principles and classroom practices.* Pearson Education.

Brualdi Timmins, A. C. (1998). Classroom questions. *Practical Assessment, Research, and Evaluation, 6*(1). 1-3. https://doi.org/10.7275/05rc-jd18

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

Cunningham, G.K. (1998). *Assessment in the Classroom*. Falmer Press.

Currie, M. & Chiramanee, T. (7111). The effect of the multiple-choice item format on the measurement of knowledge of language structure. *Language Testing*, *72*(2), 221–291. https://doi.org/ 11.11220174331771913429

Douglas, M., Wilson, J., & Ennis, S. (2012). Multiple-choice question tests: a convenient, flexible and effective learning tool? A case study. *Innovations in Education and Teaching International*, *49*(2), 111-121.

Esmaeeli, B., Shandiz, E. E., Norooziasl, S., Shojaei, H., Pasandideh, A., Khoshkholgh, R., Fazli, B. & Ahmadi, F. B. (2021). The Optimal Number of Choices in Multiple-Choice Tests: A Systematic Review. *Medical Education Bulletin, 2*(3), 253-260. https://doi.org/10.22034/MEB.2021.311998.1031

Gaspar, M. (2004). *Assessment using multiple-choice: Implications for testing international students in an under graduate commerce subject*. Deakin University.

Giusto, F., Müller Werder, C., Reichmuth, A., Adams-Hausheer, D., & Christian, J. (2019). *Multiple-choice questions: Teaching guide for higher and professional education.* ZHAW School of Management and Law.

Glass, L. A., Sinha, N. (2013). Multiple-choice questioning is an efficient instructional methodology that may be widely implemented in academic courses to improve exam performance. *Current Directions in Psychological Science, 22*(6), 471-477. https://doi.org/10.1177/0963721413495870

Golvardi Yazdi, M., Haghighat Shoar, M., Sobhani, Gh., Vafi Sani, F., Khoshkholgh, R., Mousavi Bazaz, N., & Mansourzadeh, A., (2021). Factors affecting students' guesswork in multiple-choice questions and corrective strategies. *Medical Education Bulletin, 2*(4), 341-349. https://doi.org/10.22034/MEB.2021.312176.1032

Güler, M. (2017). The effect of goal orientation on student achievement. In E. Karadag (Ed.), *The factors effecting student achievement* (pp. 291–307). Springer. https://doi.org/10.1007/978-3-319-56083-0_18

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2004). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-333.

Heaton, J. B. (1988). *Writing English language tests*. Longman.

Henry, G. T. (2015). Comparison group designs. In Wholey J., Hatry H., Newcomer K. (Eds.), *Handbook of practical program evaluation* (pp. 137-157). Jossey-Bass.

Ibbett, L. N., Wheldon, J. B. (2016). The incidence of clueing in multiple choice test bank questions in accounting: some evidence from Australia. *e-Journal of Business Education and Scholarship of Teaching, 10*(1), 20-35.

Jafarpur, A. (2003). Is the test constructor a facet? *Language Testing, 20*(1), 57-87.

Jones, G. (2021). Designing multiple-choice test items. In P. Winke and T. Brunfaut (Eds.), *Handbook of Second Language Acquisition and Language Testing* (pp.90-101), Routledge. A helpful overview of issues in developing MC items for language assessment.

Kaddoura, S., & Al Husseiny, F. (2021). An approach to reinforce active learning in higher education for IT students. *Global Journal of Engineering Education, 23*(1), 43-48.

Kline, P. (2000). *The handbook of psychological testing*. (2nd ed.), Psychology Press.

Mao, A. M. (2022). Literature review of language testing theories and approaches. *Open Access Library Journal, 9*(5), 1-5. https://doi.org/10.4236/oalib.1108741.

McNamara, T. (2003). *Language testing*. Shanghai Foreign Language Education Press.

Nicol, D. (2007). E-assessment by design: Using multiple-choice tests to good effect. *Journal of Further and Higher Education, 31*(1), 53-64.

Polat, M. (2020). Analysis of multiple-choice versus open-ended questions in language tests according to different cognitive domain levels. *Novitas-ROYAL(Research on Youth and Language), 14*(2), 76-96.

Pongweni, A. (2017). Negotiating meaning through the labyrinthine meanderings of periodic and cumulative English sentences. *Journal of Language and Literature, 28*, 21-40.

Rashidi, N., & Safari, F. (2014). Does the type of multiple-choice item make a difference? The case of testing grammar. *International Journal of Language Testing*, *4*(2), 175-186.

Sigott, G. (2004). *Towards identifying the C-Test construct*. Peter Lang International Publishers Frankfurt.

Spolsky, B. (1978). *Educational linguistics: An introduction*. Newbury House.

Sudakova, N. E., Savina, T. N., Masalimova, A. R., Mikhaylovsky, M. N., Karandeeva, L. G., & Zhdanov, S. P. (2022). Online formative assessment in higher education: Bibliometric analysis. *Education Sciences*, *12*(3), 209. https://doi.org/10.3390/educsci12030209

Ward, A.W., & Murray-Ward, M. (1999). *Assessment in the Classroom*. London: Wadsworth.

Whitley, B. E., Jr. (2002). Group comparison research. In M. W. Wiederman & B. E. Whitley, Jr. (Eds.), *Handbook for conducting research on human sexuality* (pp. 223–254). Lawrence Erlbaum Associates Publishers.

Wood, E., Klausz, N., & MacNeil, S. (2022). Examining the influence of multiple-choice test formats on student performance. *Innovative Higher Education, 47*, 515–531. https://doi.org/10.1007/s10755-021-09581-7

Appendix A

Part I- Vocabulary

1. I don't have anything sweet for ……., but we can have some fruit.
a. chemical b.dessert c.snack d.dairy
2. One of the most popular hobbies in the world is …… stamps.
a. counting b. posting c. collecting d. crossing
3. Some countries did not want to …… letters with stamps from other countries.
a. deliver b. except c. expect d. accept
4. Sometimes your face gets red when you feel …….
a. unknown b. embarrassed c. famous d. dangerous
5. Will there be enough food for all people, or will we have food……?
a. extinct b. shortage c .credit d. absence
6. Some people fast for……..reasons.
a. functional b. logical c. usual d. political
7. Lead is a very heavy, soft, dark gray …….
a. object b. piece c. metal d. liquid
8. Many Hindus fast on special ……,as do some Christians and Buddhists.
a. conditions b. occasions c. functions d. attention
9. No one knows why the Ronoke settlers ……..
a. settled b. returned c. disappeared d. deserted
10. He bought a new T. V because his old one didn't ….well.

a. hold b. shape c. function d. pour
11. They seemed to speak a/an……kind of English.
a. eager b. touch c. odd d. training
12. I'm eager to hear what you did in my ……..
a. fight b. descendants c. absence c. hostile
13. Foods that are well known to you may not be …… to people from other countries.
a. forbidden b. scary c. foreign d. familiar
14. Easter Island is one of the most…….places on earth.
a. smallest b. isolated c. populated d. nearest
15. Which one is not the opposite?
a. excellent/very bad b. decrease/increase c. accept/allow d. filled/emptied
16. When you rent an apartment for a year, you have to …… the last month's rent.
a. prepay b. prepare c. perform d. prove
17. Sometimes your face gets red when you feel …….
a. unknown b. embarrassedc. famous d. dangerous
18. Qatar has a desert climate, but Indonesia and Malaysia have a tropical one.' Here the pronoun 'one' refers to …….
a. Qatar b. desert c. climate d. Malaysia
19. Some people fast for……..reasons.
a. functional b. logical c. usual d. political
20. Lead is a very heavy, soft, dark gray …….
a. object b. piece c. metal d. liquid
21. People didn't like pencils with lead. They used bird's …… as pens instead.
a. wings b. flights c. feathers d. beaks
22. We could see the footprints of a large animal in the snow.
a. tracks b. elbows c. ears d. hooks
23. All the tropical countries are located near the …….
a. jewelry b. forest c. equator d. nature
24. When the captain went to ……the ship, no one came to meet him.
a. damage b. inspect c. appear d. elect
25. He was late for class, but he had a good……..
a. violence b. explanation c. appearance d. credit

Part II-Comprehension

A: Many years later, more settlers came to North Carolina. One of them met a Native American group called the Lumbee. They were unusual looking compared to the black-haired, brown-eyed Native Americans in the north. Some Lumbee had blonde hair and grey eyes. Then he listened to their speech and almost fell off his horse. They seemed to speak an odd kind of English.

1. Many years later, more settlers moved south and met a group of people called….
a. Ronoakes b. Carolinas c. Lumbee d. Europeans
2. They seemed to speak an odd kind of English probably means that:
a. their grandparents were able to read

b. their grandparents talked from a book
c. they were the descendants of the Roanoke settelers
d. they were natives

3. They were unusual looking compared to the black-haired, brown-eyed Native Americans in the north. They refer to……

a. Ronoakes b. Carolinas c. Lumbee d. Europeans

B: The population of the world is increasing rapidly. By 2020, there could be 75 million people on Earth. Will there be enough food for all these people, or will we have a food shortage? Some scientists think fish farming could solve this problem. However, other scientists worry that fish farming could cause serious environmental problems.

4. The population of the world is increasing……..

a. slowly b. a little c. quickly d. rarely

5. Fish farming could solve this problem. This refers to……

a. Food shortage c. population increase
b. enough food d. environmental problem

C: Chandra is a dentist in Texas. She is from India. "I'm afraid to try new foods because they might contain beef. I'm a Hindu, and my religion forbids me to eat meat from the cow. That's why I can't eat hamburgers or spaghetti with meatballs."

6. The Hindu religion forbids the eating of……

a. green vegetable b. beef c. chemicals d. candy

7. Chandra can't eat beef because of her…….

a. family b. health c. religion d. salary

D: The umbrella is a very ordinary...8 …..It keeps the rain and the …8…off people. Most umbrellas …9…., so it is easy to ..10..them.

a. fold up b. object c. carry d. sun

E: They chose one ten-millionth of the ….11…… from the……12…… to the north pole. They called this distance the ……13….

a. equator b. length c. meter d. distance

F: Human resources management is one of the most important 14….. in the field of ..15……management.

a. tourism b. functions c. scopes d. managerial

Appendix B

Part I- Vocabulary

1. For us, ……. is a candy, but once it was a medicine.

a. chocolate b. pepper c. dessert d. snack

2. …… stamps are one of the most popular hobbies in the world.

a. Counting b. Posting c. Collecting d. Crossing

3. Some countries did not want to …… letters with stamps from other countries.

a. deliver b. except c. expect d. accept

4. Please go to your office ……. You have a long-distance phone call.

a. carefully b. immediatelyc. internationally d. popularly

5. Zippers are so …… now that we forget how wonderful they are.

a. common b. fastened c. simple d. attached
6. There are about a …… students in class. It's a small class.
a. couple b. dozen c. little d. number
7. How …… did you pay for your refrigerator?
a. many b. far c. much d. long
8. Write a/an.. …… of your city, and tell us what it looks like!
a. importance b. pollution c. description d. collection
9. My friend is in an……..program for people with knee problems.
a. experimental b. experiments c. experimentd. experimented
10."What do want to do this weekend? "
"I …… go skiing in the mountains with my friends."
a. will be b. would c. am going to d. am used to
11. There are only a few …… in the bread: flour, water, yeast, and a little sugar.
a. chemicals b. events c. purposes d. ingredients
12. He …… a risk when he jumped into the ocean to save the child. He could die.
a. made b. took c. did d. received
13. When she ….the chemicals in the water, she was shocked.
a. discovers b. discovered c. discovery d. discovering
14. Would you please …….the map on the wall?
a. bring attention to c. pay attention
b. give attention to d. get attention
15. When the captain went to ……the ship, no one came to meet him.
a. damage b. inspect c. appear d. elect
16. When our cat died, we …… it under the apple tree in the garden.
a. buried b. tasted c. scared d. competed
17. Many people keep a /an……..in which they write down all their secrets.
a. curse b. value c. diary d. amount
18. Puffer fish …… a poison many times more dangerous than cyanide.
a. avoid b. contain c. remove d. blow up
19. Cattle are …… in my part of China, so there are no dairy products there.
a. special b. religious c. heated d. rare
20. What can we do to……..the birds from returning and eating fruit?
a. allow b. realize c. isolate d. prevent
21. The main …….in the cake is chocolate.
a. solid b. milk c. ingredient d. eliminate
22. A fastener …… along and joins the hooks together.
a. bends b. opens c. solves d. slides
23. His old T.V didn't ….well, so bought a new one.
a. hold b. shape c. function d. pour
24. People used bird's …… as pens because they didn't like pencils with lead.
a. wings b. flights c. feathers d. beaks
25. All the …….. countries are located near the equator
a. jewelry b. forest c. tropical d. nature

Part II-Comprehension

A: Many years later, more settlers came to North Carolina. One of them met a Native American group called the Lumbee. They were unusual looking compared to the black-haired, brown-eyed Native Americans in the north. Some Lumbee had blonde hair and grey eyes. Then he listened to their speech and almost fell off his horse. They seemed to speak an odd kind of English.

1. A group of people called…., were met by more settlers many years later
a. Ronoakes b. Carolinas c. Lumbee d. Europeans
2. ……., So they spoke an odd kind of English.
a. Their grandparents were able to read
b. Their grandparents talked from a book
c. They were the descendants of the Roanoke settelers
d. They were natives

B: The population of the world is increasing rapidly. By 2020, there could be 75 million people on Earth. Will there be enough food for all these people, or will we have a food shortage? Some scientists think fish farming could solve this problem. However, other scientists worry that fish farming could cause serious environmental problems.

3. The ……..of the world is increasing quickly.
a. population b. length c. people d. environment
4. This problem could be solved by fish farming.
a. Food shortage c. population increase
b. enough food d. environmental problem

C: Chandra is a dentist in Texas. She is from India. "I'm afraid to try new foods because they might contain beef. I'm a Hindu, and my religion forbids me to eat meat from the cow. That's why I can't eat hamburgers or spaghetti with meatballs."

5. The eating of…… is forbidden by the Hindu religion
a. green vegetable b. beef c. chemicals d. candy
6. Because of her……. Chandra can't eat beef
a. family b. health c. religion d. salary

D: The……7…. is a very ordinary object. It keeps the rain and the sun off people. Most of them…8…very easily, so it is easy to …9….. them.
a. fold up b. forbid c. umbrella d. carry

E: They chose one ten ……10……of the ….11…… from the equator to the north pole. They called this distance the meter.
a. meter b. length c. millionth d. distance

F: When preparing a/an ….12…tour, the essential first step is to ….13…attractive group rates and scheduling from a carrier.
a. inclusive b. increase c. obtain d. personal

G: Human……14…..management is one of the most important functions in the field of ….15….. management.
a. tourism b. resources c. scopes d. managerial