Journal of English Language
Teaching and Learning
University of Tabriz
No. 19, 2017

Does number of options in multiple choice tests affect item facility and discrimination? An examination of test-taker preferences*

Karim Sadeghi**
Associate Professor, English Language Department, Urmia University (Corresponding author)
Ghazal Akhavan Masoumi
MA in TEFL, English Language Department, Urmia University

Abstract

Multiple Choice tests are utilized widely in educational assessment because of their objectivity, ease of scoring, and reliability. This study aimed to compare IF and ID of MC vocabulary test items and attempted to find whether these indices are affected by the number of options. To this end, four 20 item stem equivalent vocabulary tests (3-, 4-, 5-, and 6-option MC) were administered to 194 (106 male and 88 female) pre-intermediate students. Besides, an attitude questionnaire was utilized to examine the attitudes of test takers towards MC test format. Results of one-way ANOVA showed that altering number of options in MC tests does not affect Item Discrimination (ID); however, there were significant differences between Item Facility (IF) of 3-, 5-, and 6-option and 4-, 5-, and 6-option MC test but not between 3- and 4-option MC test, suggesting that 6-option test is the most difficult test. Also, the results of questionnaire revealed.test takers preference towards the use of 3-option MC. Findings demonstrated that increasing the number of options makes a test more difficult and that choosing the right number of option for MC tests is controversial. Testers are recommended to consider various factors while choosing the right number of options.

Keywords: multiple-choice item; number of option; test-taker preferences; item facility; item discrimination

*Received date: 2017/01/11 Accepted date: 2017/05/24

**E-mail: k.sadeghi@urmia.ac.ir

Introduction

Testing and evaluation are considered as an indispensible component of all educational programs. Their importance is so high that without necessary information obtained through different modes of assessment (including tests); educational systems cannot make clear-cut decisions about the achievement and progress of students. Indeed, as Good and Brophy (1980) mention, information gained through assessment provides a basis for the selection, placement, certification, benchmarking and several other purposes. The role of assessment becomes vital in that it constructs a measure and criterion that affects the educational life of test takers.

Test is a common measurement tool and according to Haladyna (2004) A test is a measuring device intended to describe numerically the degree or amount of learning under uniform, standardized conditions (p. 17). A good test should enable the tester to locate the areas of difficulty encountered by the test taker and help the tester to diagnose the strengths, weaknesses, and problems of test takers in specific subject areas. It has always been a challenge for testers to identify the most appropriate test format or the best testing technique for a particular purpose, since a test which is used for one situation or context may be entirely useless or inappropriate for another context; in other words, there is not a unique test format that functions well in every situation; accordingly, testing researchers must identify the characteristics of each test format and select the version that most appropriately serves the purpose of a test in a given context.

Multiple Choice (MC) test is considered as a hallmark of selected response item format and is a popular measurement device which is used widely in any context and discipline (Haladyna, 2004). The reason for this popularity lies in numerous strengths that make it well-suited for assessment of mental attributes. MC tests are reliable, objective, unbiased, rapid, and easy to administer and score (Hughes, 2003). Perhaps one of the most principal advantages of MC tests is related to their flexibility for assessing a diversity of contents, abilities and skills (Osterlind, 2002); in other words, MC tests are versatile and have the capacity to measure a wide range of educational objectives and subject-matter areas in a relatively short period of time (Heaton, 1990). MC tests can be scored and assessed easily and quickly, manually or by scoring machines (Simkin & Kuechler, 2005). This feature has a crucial role in high-stakes large scale assessments when the number of test takers is high; the results are of paramount importance for stakeholders; and they are required in short amount of time. Moreover, unlike essay type tests which require a specialist in the field to correct and score the questions, MC tests do not require a specialist and can be scored by anyone. This issue adds objectivity to scoring and immunes the results of the test and also eliminates threats to validity leading to production of strong and desirable educational measurement characteristics (Downing & Haladyna, 2006). A further advantage of MC test is that a test taker cannot hide or mask his/her limited knowledge by producing obfuscated answers; on the other hand, there would not be rater bias and examinees will not be able to bluff their way through content-related material (Osterlind, 2002, p. 164). It should be mentioned that this statement is true when measurement errors like guessing are minimized (Haladyna, 2004). Moreover, MC tests enjoy a high degree of validity. According to Downing and Haladyna (2006), MC or selected response tests mainly encourage content validity evidence by allowing a thorough and representative sampling of the cognitive domain (p. 289) and this characteristic is of utmost importance when large domains of knowledge should be sampled at multiple cognitive levels. This representativeness of sampling strengths validity and decreases construct underrepresentation, which is a major threat to validity (Downing & Haladyna, 2006).

Psychometric properties of a test item: Item Facility and Item Discrimination

Classical Test Theory (CTT) utilizes traditional item and sample dependent statistics for analyzing psychometric properties of items in a test more specifically. For evaluating the effectiveness of individual items in a norm-referenced test systematically, two major techniques are utilized: item facility and item discrimination (Brown, 2005). These two item analysis techniques are two frequently reported item characteristics in language assessment with CTT (Bachman, 2004; Brown, 2005).

Item facility (IF, or item easiness) refers to the percentage or proportion of test takers who answered an item correctly in a given test (Bachman, 2004; Brown, 2005). To calculate IF, the number of test takers that answered the item correctly should be divided by the total number of test takers who took the test; in other words:

$$\text{IF} = \frac{\text{Number of test takers answering correctly}}{\text{Total number of test takers taking the test}}$$

The result of this equation is a value that ranges from 0.00 to 1.00; in this range the tester can interpret the facility of an item easily. In this value, 0.00 signifies that the item in test was so difficult that none of the test takers could answer it correctly and the value of 1.00 shows that the item was very easy and that all the test takers answered it correctly.

The second index of item quality used in item analysis in norm-referenced test is item discrimination. Item discrimination (ID) is the extent to which an item separates or discriminates test takers who scored high on a test as a whole from whose who did poorly (Bachman, 2004; Brown, 2005). ID provides the tester with information to contrast the performance of two groups (upper and lower) of test takers (Brown, 2005), and the tester can calculate ID by subtracting IF of lower group from IF of upper group; in other words:

$$\text{ID} = \text{IF}_{\text{upper}} - \text{IF}_{\text{lower}}$$

The value for ID ranges from +1.00 to -1.00. The value of 0.00 suggests that there is no contrast between the performance of test takers in lower and upper groups. It should be mentioned that ID value of 1.00 indicates that there is a maximum contrast between upper and lower groups of test takers; in other words, the value of 1.00 suggests that all test takers answered an item correctly in the upper group while all the test takers answered that item wrongly in the lower group. For example, if the ID index for an item is 0.6, it indicates that the item is discriminating well between upper and lower test takers, however, if the ID index for an item is -0.4, it can be assumed that the item is measuring something relatively different from the whole test since test takers who scored poorly on the whole test, scored well on this item.

This study uses CTT model instead of other test theories (e.g., Item Response Theory), because CTT is considered to be more flexible and reliable and it can be used in many different circumstances (Chapman, 2007). Furthermore, the new theories and models (e.g. Item Response

Theory) have shown unfulfilled assumptions, uninterpretable parameters, and also estimation difficulties (Alexopoulos, 2007). For these reasons, CTT model was chosen for this study.

Review of Literature

Although the use of MC test items is accepted in large-scale high stakes assessments, it has led to considerable disagreement among scholars when it comes to the number of options. This issue becomes vital when some well-known standardized English tests such as KET and PET use 3-option MC while IELTS and TOEFL use 4- or 5-option MC. Researchers recommend 3-option MC items because of difficulty in writing effective distractors (Rodriguez, 2005); however, there is a controversy among scholars on the optimum number of options for MC tests. According to Heaton (1990):

> The optimum number of alternatives, or options, for each multiple-choice item is five in most public tests. Although a larger number, say seven, would reduce even further the elements of chance, it is extremely difficult and often impossible to construct as many as seven good options. Indeed, since it is often very difficult to construct items with even five, writers recommend using four options for grammar items, but five for vocabulary and reading. (p. 28)

Moreover, Haladyna and Downing (1989) and Haladyna, Downing, and Rodriguez (2002) encourage testers to write as many distractors as they can. Most classroom assessments, large scale and high-stakes assessment in Iran, like university entrance examinations utilize 4-option MC tests. It is the same in important proficiency tests like TOEFL. The number of distractors required for conventional MC item is a matter of some controversy (Haladyna, 2004, p. 69). In spite of the widespread use of 4-option MC items, earlier studies recommended using 3-option MC items (Rodriguez, 2005) while authors of books on assessment struggled for increasing number of options (Haladyna et al., 2002).

Landrum, Cashin, and Theis (1993) investigated the effect of changing number of options in MC tests on the performance of students. Results of the study indicated that students performed significantly better in 3-option MC test. Rogers and Harley (1999) explored susceptibility to test wiseness and psychometric properties of 3 and 4-option MC tests. For this purpose, they administered Mathematics 30 provincial examination to 158 grade 12 students. Results of the study suggested that test wiseness was less affected in 3-option MC test, while item difficulty increased in 3-option MC test. According to Rogers and Harley (1999), teachers preferred 3-option MC test, since it was difficult to write three plausible distractors for 4-option MC test.

Rodriguez (2005) conducted a meta-analysis about the optimal number of options in MC tests. He reported that reducing the number of options caused a significant change in mean item difficulty, and this change was most obvious when number of options was reduced to 2. In other words, reducing number of options increased difficulty index (making the item easier). He also added that this reduction, in most cases, decreased ID. Again the reduction of the number of options to 2 showed the largest change in ID. And finally the investigation of test reliability in these studies revealed that in most cases reducing the number of options resulted in a decrease in the amount of reliability. Rodriguez (2005) concluded that item writing guidelines should change, and he suggested using 3-option MC tests for their numerous advantages. Following Rodriguez, Shizuka, Takeuchi, Yashima, and Yoshizawa (2006) studied the effect of the number of options on psychometric characteristics in an English reading test in a Japanese EFL university entrance examination. Results of their analysis suggested that there were no significant differences between 3-option and 4-option MC tests with regard to IF and ID. In other words, their study revealed that 3-option MC test works nearly as well as 4-option MC test.

Vyas and Supe (2008) conducted a review on the studies done on the optimal number of options. They made a systematic database search using different search engines and electronic resources like: Science Direct, ERIC, Ovid, Sage, Jstor, and Blackwell. They included twenty three articles in their study based on the following question as their framework: How many questions are optimal for multiple choice questions? (p. 130). Results of their review suggested no significant differences in psychometric properties of 3-, 4-, and 5-option MC tests. However, 3-option MC tests compared to 4- and 5-option MC tests showed a higher efficiency since the former needed fewer distracors, took less time to prepare and administer. Vyas and Supe (2008) suggested that 3-option MC test had some qualities as 4- and 5˚ option MC tests and recommended using 3-option MC test because of their considerable advantages. Tarrant, Ware, and Mohammad (2009) examined the quality of MC tests in an undergraduate nursing program in an English language university in Hong Kong. They retrieved MC tests which were administered from 2001 to 2005 in clinical and non-clinical nursing courses. A total five hundred fourteen 4-option MC items were gathered. In order to assess the impact of reducing number of options from 4 to 3, researchers discarded distractors with a choice frequency of zero. Results of the study suggested that there were few differences between 3- and 4-option MC tests with regard to item difficulty and mean test score.

Tarrant and Ware (2010) compared psychometric properties of 4-option MC with their 3-option counterparts. Researchers developed 3-option MC items based on 4-option items by eliminating the distractor with lower response rate. Tarrant and Ware (2010) administered 41 MC items to two cohorts of students over two subsequent academic years and used paired *t*-test and Pearson product correlation to compare mean item difficulty and ID of two tests. Results of their study suggested that 3-option MC was the most feasible and practical choice. They concluded that 3-option MC tests functioned relatively the same as 4-option MC test; however, 3-option MC test required less time to construct and administer. Based on such an observation, they encourage teachers to develop and use 3-option MC tests. Lee and Winke (2013) compared the mean test score of 3, 4, and 5-option MC tests. To this end, they adapted three practice College Scholastic Ability Tests (CSAT) of English listening (which originally has 5-option MC test items) to construct 3- and 4-option MC. Lee and Winke (2013) administered 3 versions of the test to three groups of Korean high school students. They also administered a questionnaire which was made by researchers. Results of tests and analysis from two hundred and sixty four Korean students revealed that the number of options significantly affected mean test scores (level of test difficulty) with 3-option MC being the easiest. Results of comparisons of mean ID showed that there were no significant differences across different formats. The researchers concluded that 3-option MC test may or may not be optimal and testers should consider several other factors in determining optimal number of options. Results of their study suggested that the majority of test takers preferred 3-option MC test.

This study mainly focuses on IF and ID differences in 3-, 4-, 5-, and 6-option MC and examines the preferable number of options for Iranian context, investigating the preferences of test takers as to number of options and reasons for their preferences. In this paper, the following research questions guided the flow of the current study:

1. Does the number of options have any effect on the psychometric properties of MC vocabulary test?

2. What are the attitudes of Iranian pre-intermediate EFL students toward different test formats?

Method

Setting and Participants

The participants of this study were 205 (112 male and 93 female) fourth grade (pre-university) high school students within the age range of 17-18. The participants attended public high schools in Iran, where as part of their compulsory education, they received two hours of English education every week.

Instruments

Proficiency test

An adapted version of KET (which included Reading, Writing and Speaking sections) was administered to homogenize the participants in terms of language proficiency. KET for Schools version utilized in study was updated in 2009; before being used in the main study, the adapted KET was piloted with 25 students similar to the target group and the KR-21 reliability of test was calculated to be 0.78.

Vocabulary Pre-test

This vocabulary test was selected from *Cambridge Key English Test 4 Self Study Pack (KET Practice Tests)* (2006) by Cambridge ESOL. The test consisted of 24 MC vocabulary items with 3-options which was piloted with 24 students and had KR-21 reliability of 0.83.

MC tests

MC tests that were used in study were selected from the national entrance examination in Iran. The MC tests were stem-equivalent; in other words, all of them shared the same stems. It should be mentioned that the contents of these tests were based on the materials students studied and covered at school and matched their level of proficiency. Different versions of MC tests utilized in this study only differed in the number of options. For distractor generation, the researchers administered MC tests as a Constructed Response (CR) test to a group of test takers similar to target students and based on the incorrect responses of students in the CR test, distracters were generated. A 20 item 6-option MC test was constructed and piloted with 30 students similar to the target group (its KR-21 was found to be 0.87). It should be mentioned that the frequency of all the words which acted as answers and distracters were checked against *Collins COUBUILD Advanced Learners' English Dictionary* (2006), and they were of similar frequency. After the 6-option test was administered to the pilot group, the least chosen distractors were omitted and 3, 4, and 5-option MC tests were constructed accordingly. These tests were piloted with 23, 25, and 30 students (their KR-21 was found to be 0.79, 0.82, and 0.85, respectively). It should be mentioned that these four versions of MC tests were also reviewed by two ELT professionals, two high school English teachers and two native speakers of English.

Attitude Questionnaire

To triangulate the quantitative data, a survey questionnaire was used. This questionnaire asked all participants which format (3-, 4-, 5-, or 6-option MC) they preferred. The questionnaire was adapted from Lee and Winke (2013). The survey questionnaire was originally in English but was translated into Persian to avoid the concern that English proficiency may affect the quality of response as Mackey and Gass (2005) caution against. The questionnaire consisted of six items: Demographic questions, one closed-ended item, and four open-ended items.

Procedure

Data collection started in October 2013. Eight English classes at different schools in Urmia (Iran) were chosen randomly. Initially we administered the adapted KET and vocabulary pre-test to ensure the homogeneity of test takers in terms of general language proficiency level and vocabulary knowledge. After elimination the outliers and ensuring homogeneity, the number of test takers decreased to 194 (106 male and 88 female). In this study, four parallel groups received one format of MC test each (one group, one format) as shown in Table 1, and the participants had 15 minutes to answer 20 MC items. And later test takers were asked to fill in a questionnaire about their attitudes toward MC test format.

Table 1. *Descriptive characteristics of participants and procedure*

Group	Male/Female	MC	Session		
			1	2	3
n= 49	26/23	3-option	KET/pretest	MC	Survey
n= 47	27/20	4-option	KET/pretest	MC	Survey
n= 47	24/23	5-option	KET/pretest	MC	Survey
n= 51	29/22	6-option	KET/pretest	MC	Survey

Result

To answer the first research question, two one-way ANOVAs were run to compare the ID and IF of MC tests with 3-, 4-, 5, and 6-option items. Table 2 presents descriptive statistics which show the means and standard deviations of the option types as far as Item Discrimination (ID) is concerned.

Table 2. *Descriptive Statistics for Option Formats based on Item Discrimination*

	N	Mean	Std. Deviation	Std. Error
3-option	20	.2655	.13563	.03033
4-option	20	.2750	.05689	.01272
5-option	20	.2970	.09932	.02221
6-option	20	.3170	.15768	.03526
Total	80	.2886	.11815	.01321

Note: N = 20 (N refer to number of items)

The results of descriptive statistics show that the means of options types are similar to each other, although there is an ascending order from 3-option to 6-option MC test. One-way ANOVA was conducted to examine differences between option types in MC test. Results are presented in Table 3.

Table 3. *ANOVA Results for Option Formats according to Item Discrimination*

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.032	3	.011	.755	.523
Within Groups	1.071	76	.014		
Total	1.103	79			

The results do not show statistically significant differences [F (3, 76) = .76, *p* = .523] among students vocabulary performance with regard to the number of options included as far as ID is concerned.

For the comparison of option differences based on Item Facility (IF), descriptive statistics including means and standard deviations were calculated and then a one-way ANOVA was employed to compare the means. Table 4 presents descriptive statistics which show the means and standard deviations of option types as far as IF is concerned.

Table 4. *Descriptive Statistics for Option Formats based on Item Facility*

	N	Mean	Std. Deviation	Std. Error
3-option	20	.8440	.06202	.01387
4-option	20	.7920	.06502	.01454
5-option	20	.7120	.10066	.02251
6-option	20	.6205	.11133	.02489
Total	80	.7421	.12080	.01351

As the table shows, participants perform better in 3-option items (M= .84, SD = .06) than in other categories. To understand whether the differences are statistically meaningful, a one-way ANOVA was used to examine the exact differences.

Table 5. ANOVA Results for Option Formats according to Item Facility

	Sum of Squares	Df	Mean Square	F	Sig.
Between Groups	.571	3	.190	4.894	.000
Within Groups	.581	76	.008		
Total	1.153	79			

The results show statistically significant differences [F (3, 76) = 24.89, *p* = .000] among number of options with regard to performance on vocabulary tests as far as IF is concerned. A Tukey post hoc test was employed in order to identify the exact points of difference among the option types. Results of Tukey test signify differences between 3-options and 5-options (*p* = .000), 3-options and 6-options (*p* = .000), 4-options and 5-options (*p* = .025), 4-options and 6-options (*p* = .000), 5-options and 6-options (*p* = .008). In other words, results of the study suggest that increasing the number of options to 5 and 6 make the test more difficult but when the number of options decreases, the test becomes easier (with the 3-option MC test being slightly easier than the 4-option test).

An attitude questionnaire was utilized to investigate attitude of learners toward different test formats. This attitude questionnaire was adapted from Lee and Winke (2013) and the same system was utilized for analysis. Written responses of students were analyzed and revealed some noteworthy findings. Regarding the first question of the questionnaire (Which item format do you prefer?), 55% of test takers responded that they preferred 3-option MC, 40% of test takers preferred 4-option MC, 1% preferred 5-option MC, while none of test takers selected 6-option as their preferred format. On the whole, more than half of test takers in the present study preferred 3-option MC as their favorite item format.

More specifically, among 49 respondents in the group that received 3-option MC, 63% (N = 31) of the students preferred 3-option MC while 37% (N = 18) of test takers preferred 4-option MC. Among 47 students in the group that took 4-option MC, 53% (N = 25) of test takers preferred 3-option MC and 43% (N = 20) of them liked 4-option MC as their favorite test format. In the group that were given 5-option MC, among the 47 students, 60% (N = 28) of them preferred 3-option MC and 34% (N = 16) of them liked 4-option MC; and only 6% (N = 3) of them selected 5-option as their preferred test format. In the 6-option group 51% (N = 26) chose 3-option while 47% (N = 24) of them liked 4-option MC.

The second question of the questionnaire which was an open-ended question dealt with the reasons students preferred a specific format. Among 133 respondents who preferred 3-option MC, 53% of test takers wrote that they preferred it because of the item easiness. 16% of the 133 respondents stated that they felt less anxious and more relaxed in 3-option MC, while 13% of them believed that answering 3-option MC was less confusing and they could concentrate better on the options. Among 97 respondents that selected 4-option MC as their preferred format, 50% of them felt that they preferred it because 4-option MC was the proper and standard type of test format; 29% of them mentioned that this format was familiar for them and that they were taking tests with this format for a long time, while 11% of respondents reasoned that 4-option MC assessed their knowledge more accurately and much better than other items. Among 3 respondents that preferred 5-option MC, 2 (65%) of them said that they preferred 5-option because it assesses students much better than other formats do.

The third question of questionnaire asked students about the advantages of their preferred format. Among the respondents who preferred 3-option MC, 30% of them wrote that 3-option MC is not time-consuming and they can save time and answer more questions in exams like entrance examination; 28% of them mentioned that it was less confusing and more to the point. Among students that selected 4-option as their favorite format, 38% of them explained that it is a standard format which is used across the world, and 26% of them responded that it discriminates better between high level and low level students. Finally, 100% of the respondents that selected 5-option MC considered the challenging aspect of it as its important advantage.

The fourth question in this questionnaire dealt with disadvantages of other formats. A total of 33% of respondents to 3-option MC considered other formats as time-consuming in contrast to 3-option MC, while 18% of them felt that other formats are difficult and confusing (especially 6-option MC). A total of 42% of respondents that favored 4-option MC wrote that other formats are unfamiliar and not appropriate for large scale assessments like entrance examination that affects the life of test takers, and 38% of them called other formats as nonstandard and abnormal. Finally, 70% of respondents who favored 5-option MC argued that other MC formats are easy, and 30% of them said it is not possible to use 6-option MC format in large scale assessments.

The last question in this questionnaire asked students about the appropriate format for entrance examinations in Iran. A total of 59% of respondents preferred 3-option MC; 38% of them selected 4-option as their preferred format. Among 142 respondents that preferred 3-option MC, 44% of them wrote that they preferred it because they can easily choose the answer; 25% of them considered it as less time-consuming and said they can save time; and 17% said it is less confusing. Among 94 respondents that preferred 4-option MC, 46% of them explained that it is familiar for them and they are used to it; 25% of them said it is a standard and proper test format; and 17% of them mentioned that they felt less anxious or nervous while answering 4-option MC.

On the whole, questionnaire data revealed that students had positive attitude toward 3-option MC items and using such items in entrance examinations in Iran. This positive attitude can be due to the advantages it has for students such as its easiness, being time-saving, less confusing, etc.

Discussion

This study was an attempt to investigate the effect of number of options on IF and ID which are considered as important psychometric properties of a test. In so doing, this study was meant to find the optimal number of options for MC items. Results of one-way ANOVA for ID suggested no significant difference between ID of 3-, 4-, 5-, and 6-option MC items. However, results of one-way ANOVA for IF revealed significant differences between IF of all options, except between 3 and 4 options, and indicated that 6-option items are the most difficult MC format. In other words, results of the study suggested that increasing number of options to 5 and 6 made the test more difficult while with a decrease in the number of options, the test became easier (with the 3-option MC test being slightly easier than the 4-option test).

Findings of the current study are in line with findings of Currie and Chiramanee (2010) and Rodriguez (2005). Currie and Chiramanee (2010) examined 3-, 4- and 5-option MC items in an English structure test. Their findings showed that decreasing number of options to 3-options made the test slightly easier, since the performance of test takers was slightly better in 3-option MC. In his meta-analysis, Rodriguez (2005) reviewed 27 studies and found that generally decreasing the number of options increases facility index and makes the test easier. He further suggested that 3-option MC test be used for assessment purposes because of its numerous advantages.

Lee and Winke (2013) compared the performance of test takers in 3-, 4-, and 5-option MC listening test. Their findings suggested mean IF increases with reducing number of options and indicated that 3-option MC was significantly easier; however, they also found that there were no significant differences in mean ID indices of 3-, 4-, and 5-option MC items. They took a conservative position and concluded that 3-option may or may not be optimal and for deciding on optimal number of options for a specific test, testers should consider several other factors, too.

Some of the prior studies done in the area of number of options compared only 3- and 4-option MC items. In this regard, findings of this study are also consistent with findings of Shizuka et. al. (2006). Their study found no significant differences in mean IF of 3- and 4-option MC items in a reading test; also the analysis of mean ID of 3- and 4-option MC items suggested no significant difference between them. Based on these conclusions, they suggested using 3-option MC items for assessing reading ability because it is economical and practical. Likewise, Tarrant and Ware (2010) compared the psychometric properties of 3- and 4-option MC tests, and their results suggested no significant difference in item difficulty and ID of 3- and 4-option MC items. They encouraged test developers to use 3-option MC items because it was less time consuming and more practical.

Results of the current study appear to contradict findings of Rogers and Harley (1999). They compared item difficulty of 3- and 4-option MC test in a test of mathematics and their results revealed that item difficulty increased in 3-option MC items. Own and Froman (1987) compared the performance of test takers in 3- and 5-option MC items. Their results suggested that item difficulty and ID were not significantly different in 3- and 5-option MC items.

Our findings suggest that test takers performance is better on MC items with fewer options (3 and 4 options). One logical justification for such an observation is that in these items test takers do not deal with confusing options and they do not spend time to delete non-plausible distracors; in this regard, 3- and 4-option MC items put less cognitive burden on test takers for deleting the distractors and selecting the right answer, while in 5- and 6-option MC test, test takers are forced to read options as fast as possible within the limited time and choose a correct answer. No difference between 3- and 4-option MC could mean that items were of appropriate level of difficulty for the test-takers, and also indicate that 3- and 4-option MC items are less affected by construct irrelevant factors such as anxiety and unfamiliarity.

Data gathered from the survey convey a similar finding: test takers believed that 5 and especially 6-option MC items made them perplexed and confused. Construct irrelevant factors create variances or fluctuations in scores due to the factors unrelated to the construct being measured (Kaplan & Saccuzzo, 1997). One of the reasons that test takers perform poorly in 5- and 6-option MC items may be due to some construct irrelevant factors such as test anxiety (Lee & Winke, 2013). In doing any assessment or test, test takers generally feel anxious or mainly experience test anxiety and this anxiety increases when they encounter a new type of the format (5- and 6-option MC) different from the conventional and familiar format. In the Iranian context, test takers are not familiar with 5-option MC test which might be a familiar format for test takers in other countries (e.g., Japan), so Iranian test-takers may be affected by anxiety or stress which affects their performance negatively (test takers in this study felt quite surprised when they were given 5- and 6-option MC items, they seemed perplexed and said it is abnormal and it is an awkward test). The mean ID was not significantly different among MC tests used in the current study, because as prior studies (e. g., Lee & Winke, 2013; Shizuka et al., 2006) conducted in English as a Foreign Language (EFL) and English as a Second Language (ESL) contexts highlighted, the stem (i.e., the core of the question), the problem and the correct response are the same in all these formats and only the number of options differs. In other words, only did the process of selecting the correct response vary based on the number of options, with more options putting more burden on test takers for eliminating the non-plausible distractors. So it can be concluded that the number of options in an MC test does not change the power of a test in discriminating among test takers even though the number of options affects IF.

Most of the prior studies on the optimal number of options suggested using 3-option MC items (Haladyna & Downing, 1993; Rodriguez, 2005; Rogers & Harley, 1999; Tarrant & Ware, 2010; Vyas & Supe, 2008), because it offers different benefits for teachers and testers. First, use of fewer options saves time and decreases testing time (Tarrant et al., 2009); second, with fewer options, more items can be added to tests and it enables the tester to increase the sampling content while keeping testing time constant (Tarrant & Ware, 2010). Furthermore, since writing more plausible distractors is difficult and time consuming (Haladyna & Downing, 1993), with fewer options testers can write more plausible distractors in a relatively shorter time and put their effort to increasing the number of items rather than the number of options (Lee & Winke, 2013). Based on the findings of the current study, it seems that 3-option MC test seems cost-effective, since there are no differences in ID between 3-, 4-, 5-, and 6-option MC items, and taking into account the difficulty of writing 4 or 5 plausible distractors, it would be beneficial to use MC tests with fewer options. However, 3-option MC tests may or may not be optimal and testers should consider several other statistical and contextual factors while choosing an appropriate number of options for MC items.

This study tried to investigate the attitude and preferences of test takers towards different test formats. To this end, an adapted attitude questionnaire based on Lee and Winke (2013) was utilized. Questionnaire data suggested that test takers mostly preferred 3-option MC items. They preferred 3-option MC because they assumed that it was easier than other versions and that they felt less anxious and less confused while doing MC items.

Finally, results of this study corroborate findings of Lee and Winke (2013). They examined the preferences of test takers toward 3-, 4-, and 5-option MC tests, and found that test takers preferred 3-option MC for assessment and also for their entrance examination in Japan.

In general, findings of the current study revealed that test takers preferred 3-option MC. Test takers are mostly neglected in assessment and their ideas and opinions are not paid enough attention in making assessment-related decisions, while assessment outcomes directly affect their educational life. It is accordingly necessary for testers to take into account and consider their attitude and opinions while developing tests, so as to develop fair tests.

Conclusions and Limitations

In this study, we examined the effect of number of options on psychometric properties (IF and ID) of MC tests. Overall, the findings of current study suggest that increasing number of options makes MC tests more difficult, but it does not affect discrimination power of the test. In addition, it was found that test takers mainly preferred 3-option MC.

As with most research studies, the current study is subject to some limitations by virtue of possible methodological and practical restrictions which were imposed on it. This study was, of course, limited in the number test items investigated, due to the practicality and timing issues.

References

Alexopoulos, D. S. (2007). Classical test theory. In N. J. Salkind, (Ed.), *Encyclopedia of measurement and statistics* (pp. 140-143). Thousand Oaks: SAGE Publications.

Bachman, L. F. (2004). *Statistical analyses for language assessment.* Cambridge: Cambridge University Press.

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

Chapman, E. (2007). True Score. In N. J. Salkind, (Ed.), *Encyclopedia of measurement and statistics* (pp. 1014-1016). Thousand Oaks: SAGE Publications.

Currie, M., & Chiramanee, T. (2010). The effect of the multiple-choice item format on the measurement of knowledge of language structure. *Language Testing, 27*(4), 471-491. doi: 10.1177/0265532209356790

Downing, S. M., & Haladyna, T. M. (2006). *Handbook of test development.* Mahwah: Lawrence Erlbaum Associates.

Furnharm, A., Christopher, A., Garwood, J., & Martin, N. G. (2008). Ability, demography, learning style, and personality trait correlates of students preference for assessment method. *Educational Psychology, 28*(1), 15-27.

Good, T.L., & Brophy, J.E. (1980). *Educational psychology: A realistic approach* (2nd ed.). New York: Harper and Row.

Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*(1), 37˚ 50.

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement, 53*(4), 999-1010. doi: 10.1177/0013164493053004013

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurment in Education*, *15*(3), 309-334.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Mahwah: Lawrence Erlbaum Associates.

Heaton, J. B. (1990). *Writing English language tests*. New York: Longman.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

Kaplan, R. M., & Saccuzzo, D. P. (1997). *Psychological testing: Principles, applications,and issues*. Pacific Grove: Cole Publication.

Landrum, R. E., Cashin, J. R., & Theis, K. S. (1993). More evidence in favor of three-option multiple-choice tests. *Educational and Psychological Measurement, 53*(3), 771-778. doi: 10.1177/0013164493053003021

Lee, H., & Winkle, P. (2013). The differences among three-, four-, and five-option-item formats in the context of a high-stakes English-language listening test. *Language Testing, 30*(1), 99-125.

Osterlind, S. J. (2002). *Constructing test items multiple-choice, constructed-response, performance, and other formats*. Dordrecht: Springer.

Owen, S. V., & Froman, R. D. (1987). What s wrong with three option multiple-choice items? *Educational and Psychological measurement, 47*(3), 513-521.

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), 3-13. doi: 10.1111/j.1745-3992.2005.00006.x

Rogers, W. T., & Harley, D. (1999). An empirical comparison of three-and four-choice items and tests: Susceptibility to testwiseness and internal consistency reliability. *Educational and Psychological Measurement, 59*(2), 234-247. doi: 10.1177/00131649921969820

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

Simkin, M. G., & Kuechler, W. L. (2005). How well do multiple choice tests evaluate student understanding in computer programming classes? *Journal of Information Systems Education, 14*(4), 389-399.

Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. *BioMed Central.* Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2713226/.

Tarrant, M., & Ware, J. (2010). A comparison of the Psychometric properties of three- and four-option multiple-choice questions in nursing assessments. *Nurse Education Today. 30*(6), 539-543.

Tozoglu, D., Tozoglu, M. D., Gurses, A., & Dogar, C. (2004). The students' perception: Essay versus multiple-choice type exams. *Journal of Baltic Science Education, 2*(6), 52-59.

Vyas, R., & Supe, A. (2008). Multiple choice questions: A literature review on the optimal number of options. *The National Medical Journal of India*, *21*(3), 130-133.

Zeidner, M. (1987). Essay versus multiple-choice type classroom exams: The students' perspective. *The Journal of Educational Research. 80*(6), 352-358.