Psychological Test and Assessment Modeling, Volume 53, 2011 (2), 192-211

The effects of the number of options on the psychometric characteristics of multiple choice items

Purya Baghaei[1] *& Nazila Amrahi*[2]

Abstract

This study aims to determine the optimal number of options for multiple-choice test items. A common item equating design was employed to compare item statistics, distracter performance indices, person statistics, and test reliabilities across three similar vocabulary test forms, A, B, and C, which were different only in the number of options per item. Forms B and C were constructed by randomly deleting one distracter from each item in Form A and Form B respectively. Form A, Form B, and Form C, each containing 30 multiple-choice vocabulary test items with five, four, and three options per item, were randomly given to 180 graduate and undergraduate English major university students. The three test forms were linked by means of ten common items using concurrent common item equating. The Rasch model was applied to compare item difficulties, fit statistics, average measures, distracter-measure correlations, person response behaviors, and reliabilities across multiple-choice vocabulary test items with five, four, and three options per item. Except for discrimination power of distracters which was revealed to be inversely affected by the number of options per item, no significant change was observed in item difficulties ($p < 0.05$), item fit statistics, person response behaviors, and reliabilities across the three test forms. Considering the amount of time and energy needed for developing multiple-choice tests with more distracters, three options per item were concluded to be optimal.

Key words: Optimal number of options, Multiple-choice tests, Rasch model

[1] *Correspondence concerning this article should be addressed to:* Purya Baghaei, PhD, English Department, Islamic Azad University, Mashhad Branch, 91886 Mashhad, Iran; email: pbaghaei@mshdiau.ac.ir.

[2] Urmia University, Iran

1. Introduction

Multiple-choice tests are of considerably widespread use as a means of objective measurement. The main reason behind such popularity is the many dominant advantages associated with multiple-choice tests. They can be used for diagnostic as well as formative purposes and can assess a broad range of knowledge. In addition, they are scored easily, quickly, and objectively either by human-beings or by scoring machines. These and many similar advantages make multiple-choice tests suitable for a wide range of purposes ranging from classroom achievement testing to large-scale standardized tests. Thus, improving the quality of multiple-choice test items appears to be of a lot of importance.

In their taxonomy of multiple-choice item-writing guidelines, Haladyna and Downing (1989) suggest 43 guidelines of which 10 are concerned with general item writing, six are related to stem development, and 20 refer to option development. The fact that a good number of guidelines have dealt with the issue of option development is an indicator of the importance of this last concern.

One of the most frequently mentioned guidelines with regard to option development deals with the number of options to be written for each item. Haladyna, Downing, and Rodriguez (2002) recommend writing as many plausible distracters as possible. Traditionally, it is recommended to use four or five options per item in order to reduce the effect of guessing. Most classroom achievement tests and international standardized tests (e.g. TOEFL) usually follow the rule of four options per item. In spite of the widespread use of four or five options per item advocated by many authors and test developers, most of the studies carried out to investigate the optimal number of options have ended with recommending the use of three-choice items (Aamodt & McShane, 1992; Crehan, Haladyna & Brewer, 1993; Delgado & Prieto, 1998; Haladyna & Downing, 1993; Landrum, Cashin & Theis, 1993; Rodriguez, 2005; Shizuka, Takeuchi, Yashima, & Yoshizawa, 2006; Straton & Catts, 1980).

Significance of the study

In every testing situation, including language testing, test developers are always concerned with two main issues: *what* of testing and *how* of testing. *What* of testing is often specified in terms of test blue-prints developed in almost all testing situations, be it a simple classroom achievement test or a large scale standardized test. However, deciding on *how* of testing is not so straightforward. Traditionally, deciding on how of testing was considered to be limited to a few test methods like oral interview, multiple-choice, self-rating, etc. (Bachman, 1990). Bachman (1990) rejecting such a monolithic view of test methods, presents a framework in which several facets for test method are introduced which comprise the how of testing. One of these facets which has received an overriding attention in most test development procedures is item format (categorized under the label of *facets of the expected response* in Bachman's framework). Rodriguez (2002) points out, "the choice of item format is reduced to one between multiple-choice and constructed-response formats or some combination of the two" (p. 213). However, as men-

tioned before, such a monolithic view of item format should be replaced with a broad view which takes into account the existing variations in each of these two item formats.

For a multiple-choice test item, Bachman (1990) distinguishes variations like whether the item involves matching or completing the missing information or whether it is a direct question. He also refers to the length of the item as well as the task specified by the item (e.g., selecting the correct or the best answer, identifying or selecting the incorrect part) as existing variations in a multiple-choice test item. The present study aims to point out and investigate another variation associated within a multiple-choice test item, namely the number of options. The number of options as a feature affecting the characteristics of the test method is used to elicit test performance which in turn is an indicator of the test-takers' ability. It is extremely important to understand whether test method affects examinee performance and the what of testing, i.e., the construct. It is argued that the test method should not interfere with the construct that we want to measure. This study aims at demonstrating the effect of test method, which is rendered as the number of options per item in this study, on examinee performance and test quality statistics.

2. Previous studies

Straton and Catts (1980) compared some of the characteristics of multiple-choice economics achievement test items which contained two, three and four distractors. Time required to complete the forms was compared as well. Though classical analysis of the results manifested the greatest reliability for three-choice form, there was not a notable difference between the reliability of scores on the four-choice test form and those on the three-choice form. The two-choice form was indicated to be quite poor with regard to test reliability. Mean item difficulty calculated for all test forms revealed that the more distracters per item, the more difficult the item was. The mean item discrimination was highest for four-choice form followed by three-choice and two-choice forms. In addition, the standard error of measurement was lowest for the three-choice form and highest for the two-choice form. Moreover, the study revealed that mean time taken to complete the forms increased with the number of items while the mean time per item increased with an increase in the number of choices. Thus, Straton and Catts (1980) concluded that three options per item are optimal for classroom achievement tests unless there is a need for easy items in which case two options would serve better.

Aamodt and McShane (1992) investigated the effect of the number of options per item in multiple-choice tests on test scores and test completion time. Eight studies with fourteen samples were included in their meta-analysis. The results indicated that the number of options in an item does not affect the test scores greatly. Three-choice items were indicated to be only slightly easier than four-choice items. However, it was revealed that three-choice items took significantly less time to complete. So it was inferred that instead of four-choice items, more three-choice items could be included in an exam without any increase in testing time, and thus more content valid tests could be developed. Taking into account that three-choice items do not affect the test scores greatly and also take less

time to construct and administer, the researchers concluded that they are preferable to their four-choice counterparts.

Crehan, Haladyna, and Brewer (1993) carried out item analysis and test analysis comparing multiple-choice tests composed of four-choice items with those containing three-choice items. Forty-eight items were selected from the instructor's manual for a psychology course by the course instructor and were rewritten in the form of both four-choice items and their three-choice counterparts. The three-choice form was constructed by eliminating the least functional distracters which were considered to be the distracters with the lowest discrimination indices based on the already available item data. Approximately 110 psychology students responded to the two test forms as their final exam, with about 55 examinees taking each test form. Item and test analysis comparing the two test forms were carried out using methods of Item Response Theory. The results revealed that the three-choice items were less difficult than four-choice items. However, the four-choice test form turned out to be more reliable than the three-choice form. No important difference was revealed in item discrimination indices of the two test forms. Crehan et al. (1993) point out the need for further research using methods of item response theory in order to obtain more conclusive results.

Haladyna and Downing (1993) investigated the frequency of occurrence of effective and ineffective distracters for several different testing programmes. Three multiple-choice tests, each representing a different test use, were selected. The first test was a standardized test used as part of a graduate medical education programme which consisted of 200 five-choice items. The second test was part of the ACT Assessment used to predict college grades. It contained 75 four-choice items on reading and 52 four-choice items on social studies. The third test was a state certification in the health sciences and consisted of 150 four-choice items.

The tests were given to 1110, 500, and 247 test-takers, respectively. The performance of distracters was studied by investigating trace lines (option characteristic curves) constructed from the relevant option-by-score group contingency tables. It was expected that the correct option would have a monotonically increasing trace line and the distracters would have monotonically decreasing trace lines. The results indicated that the number of effectively performing distracters per item was approximately one. Moreover, product-moment correlations revealed a direct relationship between the number of effective distracters for an item and item discrimination while no relationship was detected between the number of effective distracters for an item and item difficulty. Since it was indicated that few items had more than two effective distracters, three-choice test items were concluded to better serve the testing purposes, particularly considering the amount of time spent for test construction.

In an attempt to solve the guessing problem involved in multiple-choice tests Kubinger and Gottschall (2007) examined a type of multiple choice items, called, the "*x* of 5" format. In this format the items have five options with multiple correct answers. In order to get a point for an item the test-takers have to mark all the correct options and none of the wrong options. The number of correct options can vary across the items. Kubinger and Gottschall (2007) demonstrated that the "*x* of 5" format is significantly more difficult

than the traditional multiple choice items with one correct response out of six options. Their study also showed that "*x* of 5" is as difficult as the constructed response format. They conclude that the "*x* of 5" format is less guessing-prone and can be used instead of traditional multiple choice items.

In another study Hohensinn and Kubinger (2009) demonstrated that the three above-mentioned response formats, i.e., "1 of 6", "2 of 5" and constructed response format measure the same latent traits and the response format does not affect the construct of interest.

Kubinger, Holocher-Ertl, Reif, Hohensinn, and Frebort (2010) compared two multiple-choice response formats, namely one answer out of six options ('1 of 6') and two answers out of five options ('2 of 5') with free-response format of a mathematics test. Only if examinees had marked both correct answers and none of the distracters in the '2 of 5' format the items were considered as correct. Kubinger et al. (2010) demonstrated that the free-response and the '2 of 5' formats were significantly harder than the '1of 6' format. The free-response format was slightly harder than the '2 of 5', not statistically significantly though. Kubinger et al. (2010) conclude that the reason why '1 of 6' format turned out to be easier than the '2 of 5' was the large of amount guessing that is involved in answering single response multiple-choice items even when there are five distracters and recommend double or multiple-response multiple-choice items to eliminate the guessing problem associated with canonical multiple-choice items.

Landrum, Cashin, and Thesis (1993) compared three-choice and four-choice test forms in terms of student as well as test-item performance. Five 50-item four-choice tests (each test administered at the end of one unit of the course content) and five 50-item three-choice tests were given to a group of undergraduate students of psychology in two successive semesters. Only 57.6 % (i.e., 144 of 250) of items were identical (except for the number of options). The researchers based their study only on the common items whose number of options differed (i.e., was reduced from four to three) in two successive semesters. The mean scores as well as the paired t-test results indicated a general pattern of improvement in students' performance on three-choice tests in comparison with their four-choice counterparts. The results revealed three-choice test items were more difficult than or at least as difficult as their four-choice counterparts. Thus, the improvement of students' performance was associated with the superiority of three-choice test items, or as the researchers state the three-choice items are "perhaps more valid test of student knowledge" (Landrum, Cashin, & Thesis, 1993, p. 777).

Delgado and Prieto (1998) studied the effect of the number of options on item difficulty, item discrimination, and test reliability from a classical test theory perspective. Two versions of three different computerized tests, with four and three options per item, each containing 30 items were given to a total of 433 first-year university students in two successive years as classroom assessment tests for an introductory course on Research Methodology in Psychology. The three-choice test forms were formed by deleting the least frequently endorsed option in their four-choice counterparts. Analysis of the results revealed no notable change in item difficulty, item discrimination, and test reliability. Thus, further evidence favoring three-choice items was provided.

Rodriguez (2005) conducted a meta-analysis of 27 empirical studies. The results revealed a slight decrease in item difficulty as a result of reducing four options to three options, while items became considerably easier when the number of options was reduced to two. Reducing the number of options per item resulted in a reduction in item discrimination power, except for when the number of options was reduced from four to three in which case a slight increase in item discrimination was observed. The greatest changes were obtained when the number of options was reduced to two. Reduction in the number of options mainly led to a decrease in test score reliability, except for when the number of options was reduced from four to three in which case a slight increase in test score reliability was observed. Out of those 27 studies which were included in this meta-analysis, only two studies dealt with criterion-related validity. The results of both studies indicated that reducing the number of options from five to three and from five to four to three led to a statistically negligible change in criterion-related validity of the test. Considering the results of this meta-analysis as well as issues like time needed for test construction and administration, Rodriguez concluded three options are optimal.

Shizuka, Takeuchi, Yashima, and Yoshizawa (2006) compared the psychometric characteristics of three- and four-choice English reading comprehension items using the Rasch model. Thirty-eight multiple-choice items intended to tap reading comprehension as a part of a university entrance exam in Japan formed the four-choice form. Later, the three-choice form was constructed by eliminating the least frequently chosen distracter, and it was linked to the four-choice test form using common-item equating. The two test forms were given to two separate groups. The effects of the number of options per item on the psychometric characteristics of the two test forms were investigated. The results revealed no significant change in the mean item facility. No notable change was observed in test reliability and the number of discriminating distracters. Thus, considering issues like reducing the chances of providing unintended cues by offering more options per item, costs of test development and administration time, three options per item were concluded to be optimal.

Though many researchers have tried to determine the optimal number of options for multiple-choice tests in different fields, few have addressed the issue of the optimal number of options for multiple-choice tests of English as a foreign or second language. This study investigates the issue of optimal number of options for multiple-choice vocabulary tests of English as a foreign language.

3. Method

3.1 Participants and instrument

One-hundred and eighty undergraduate English majors (102 females and 78 males) from two universities in Iran participated in the present study. Their age ranged from 19 to 25.The instrument used was a five-option multiple choice vocabulary test with one correct answer out the five options. To develop the instrument a vocabulary test comprising of 50 completion items was given to 35 undergraduate English major university students.

This was done in an attempt to adhere to a general multiple-choice item-writing guideline suggested by Haladyna (2004) which states: "use typical errors of students when you write distracters" (p. 120). The stems of the items were taken from the British National Corpus; distracters for a 50-item five-choice vocabulary test were constructed based on plausible wrong answers provided by the students as well as the researchers' personal judgment. Great care was taken to write as plausible as possible distracters. The test was reviewed by two native speakers of English, pilot tested and once again reviewed by two other native speakers of English and suggestions as regards the soundness of distracters were implemented.

Twenty four-choice vocabulary items from the Test of English as a Foreign Language (ETS, 1987/89) were added to the 50 five-choice items so as to have some items to serve the linking purpose later on. A pilot test was carried out with 60 undergraduate English majors. The participants were required to choose the best possible answer for each item. Time allowed for answering all the items was 45 minutes though some of the participants finished the test sooner. Out of the 50 items 30 which had acceptable fit indices were selected for the study. Out of the 20 TOEFL items 10 were selected to be used as anchor items in the equating procedure. The procedures of pilot data analysis as well as establishing the validity and reliability of the test are discussed in detail in Baghaei and Amrahi (in press).

The test was administered as part of the biweekly assessments in reading comprehension courses which were meant to motivate English students to expand their reading comprehension skills and vocabulary. The scores on these tests counted towards students' final exam marks.

Research questions

In order to investigate the effects of the number of options for multiple-choice vocabulary test items on distracter performance, item and test level characteristics in a systematic way, the following research questions were formulated:

Q1: Does the number of options for multiple-choice vocabulary test items have any notable impact on item statistics?

Q2: Does the number of options for multiple-choice vocabulary test items have any notable impact on distracter performance indices?

Q3: Does the number of options for multiple-choice vocabulary test items have any notable impact on person statistics?

Q4: Does the number of options for multiple-choice vocabulary test items have any notable impact on the reliability of the test?

3.3 Procedures

The present research employed a common item equating design to compare item statistics, distracter performance indices, person statistics, and test reliabilities across three

similar test forms which were different only in the number of options per item. The three test forms were: Form A, with five options per item, Form B, with four options per item, and Form C, with three options per item.

The validity and reliability of the five-option test (Form A) was established within the Rasch model framework (Baghaei & Amrahi, in press). It contained 30 five-choice items. Form B contained 30 four-choice items formed by omitting randomly one of the distracters from each item in Form A. Form C contained 30 three-choice items formed by omitting randomly one of the distracters from each item in Form B. The three test forms were linked using concurrent common item equating. Ten four-choice vocabulary items from TOEFL past papers (ETS, 1987/89) were used as anchor items and served the linking purpose. That is, 10 four-choice items were common in the three forms which resulted in three 40-item test forms.

Form A, Form B, and Form C were randomly given to 180 graduate and undergraduate English major university students. The data from the three tests were analysed simultaneously in one single analysis in a common item concurrent equating design. The 10 common items linked the three forms making the comparisons across the forms possible. WINSTEPS Rasch software version 3.66.0 (Linacre, 2009) was used for data analysis. Item difficulties and fit statistics, average measures, distracter-measure correlations, person response behaviors, and reliabilities were compared across the three test forms with five, four, and three options per item.

4. Results

To make sure of the quality of the equating procedure the item difficulty estimates of the 10 common items were computed separately for each test form. When the difficulty estimates of these items were cross-plotted, the slopes of the best-fit lines were 1.00, indicating all 10 items could satisfactorily serve as anchor items to bring the three test forms onto a common scale and make comparisons. Thus, anchor items had performed their job and were excluded from later analyses. Item-person map or the Wright map **(**Figure 1), which depicts items and persons jointly on a common scale, shows that the difficulty of the test was matched to the persons' ability. The map shows that the bulk of items on the right are matched to the bulk of persons on the left, indicating the test is appropriately targeted for the participants. In other words, the items are at the ability level of the test-takers.

4.1 Item statistics

Item statistics were calculated for test Forms A, B, and C. The root mean square error (RMSE) value, representing the mean of standard error of item parameter estimates, was 0.34 for all three test forms. Thus, item parameters in all three forms were estimated with the same precision.

In order to compare item difficulty estimates and fit statistics of items across all three test forms, mean values of difficulty estimates and item fit statistics were calculated and compared as indicated in Table 1.

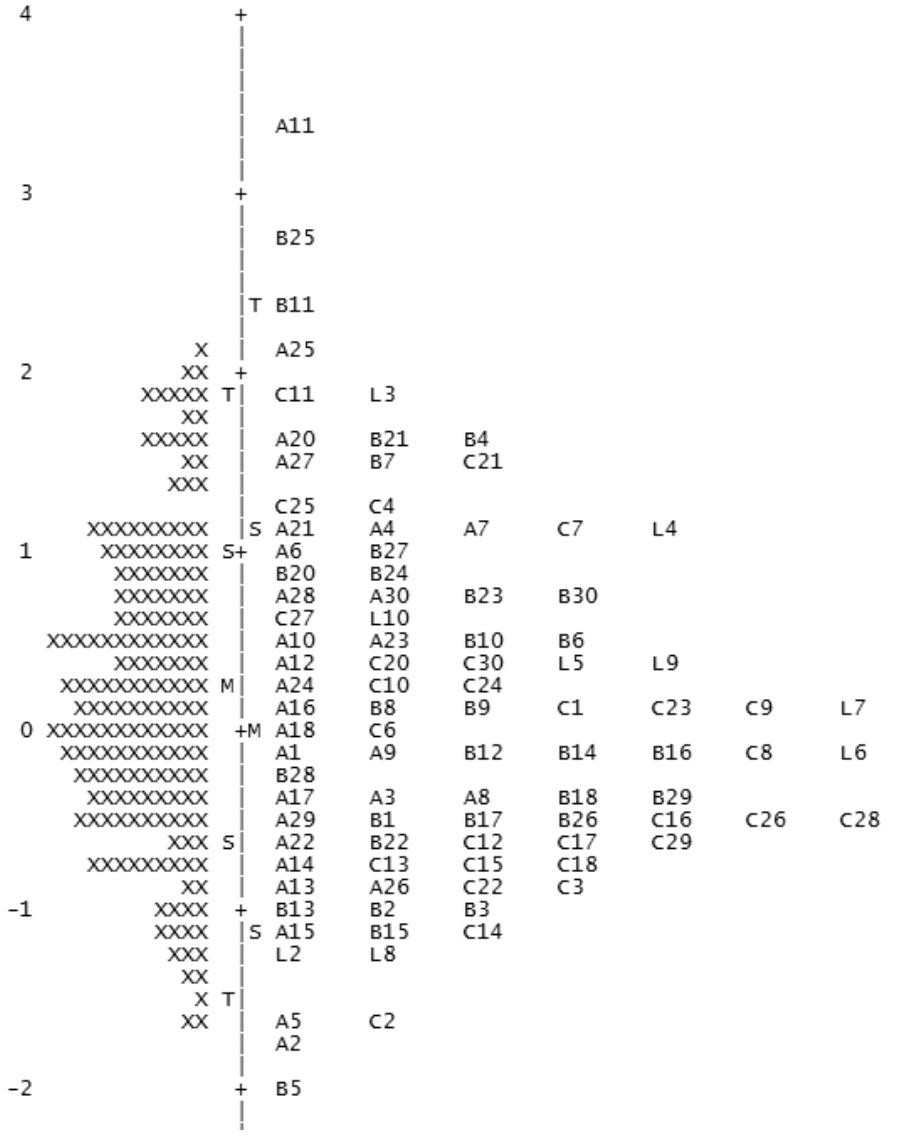


Figure 1:
Wright Map of the order of items and persons

Table 1:
Item statistics for the three forms

	n	Mean item difficulty	Mean Infit Statistics		Mean Outfit Statistics	
			MSQ	ZSTD	MSQ	ZSTD
Form A	30	0.14	0.97	-0.08	0.95	-0.07
Form B	30	0.09	0.98	-0.10	0.95	-0.15
Form C	30	- 0.2	0.98	-0.04	0.96	0.01

n = number of items.

As Table 1 shows, reducing the number of options resulted in a slight decrease in mean item difficulty estimates, suggesting that the number of options had a minimal effect on item difficulty. However, the differences among the mean item difficulty estimates of three test forms were too subtle to conclude that reducing the number of options leads to the reduction of item difficulty estimates. Moreover, a one-way analysis of variance (ANOVA) was used to investigate whether the differences among the mean item difficulty estimates of the three test forms were statistically significant. Results of ANOVA indicated that there was no statistically significant difference at $p < 0.05$ in the means of item measures for the three forms: $F\ (2, 87) = 0.87$, $p = 0.42$. Thus, it was concluded that the number of options for vocabulary test items did not have any significant impact on item difficulty.

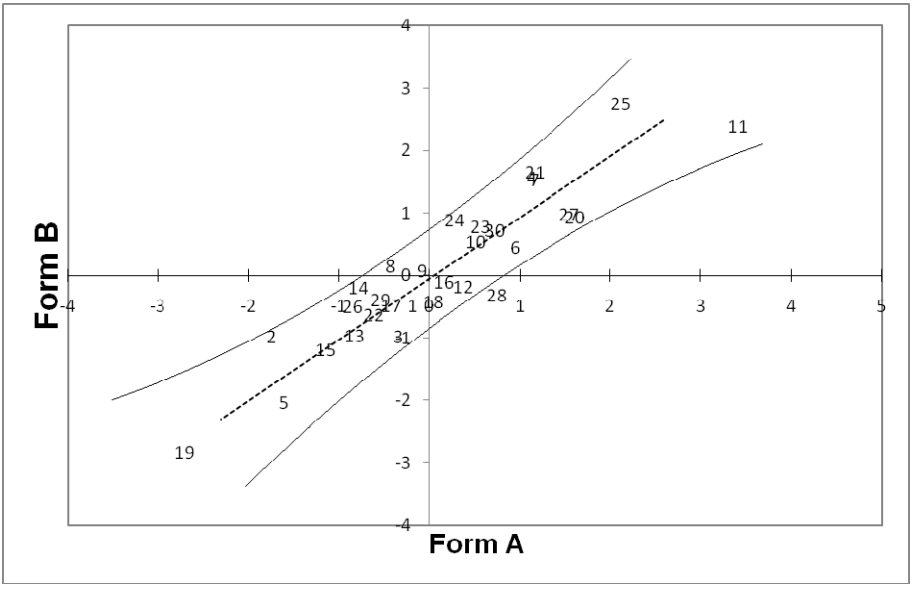
Figure 1 shows the scatterplot of item parameters from the three forms against each other in a pair wise fashion. The figure clearly indicates that the item difficulty estimates do not change significantly when they have different number of response options. All the items fall close to the line of best fit and within the approximate 95 % quality control bands (Wright & Stone, 1979).

Investigation of the residual-based infit and outfit indices revealed that almost all items had acceptable fit. The mean square statistics for the items, which indicate "unmodeled noise or other source of variance in the data" (Linacre, 2009, p. 444), were within the recommended range of 0.7-1.3 and the *z*-standardized (ZSTD) indices, which show the statistical significance of the unexpectedness observed in the responses, were within -2 to 2 (Bond & Fox, 2007). There were one or two items in each form which had indices slightly greater or smaller than these values. However, no specific pattern could be detected as regards the fit of the items in relation to test form.

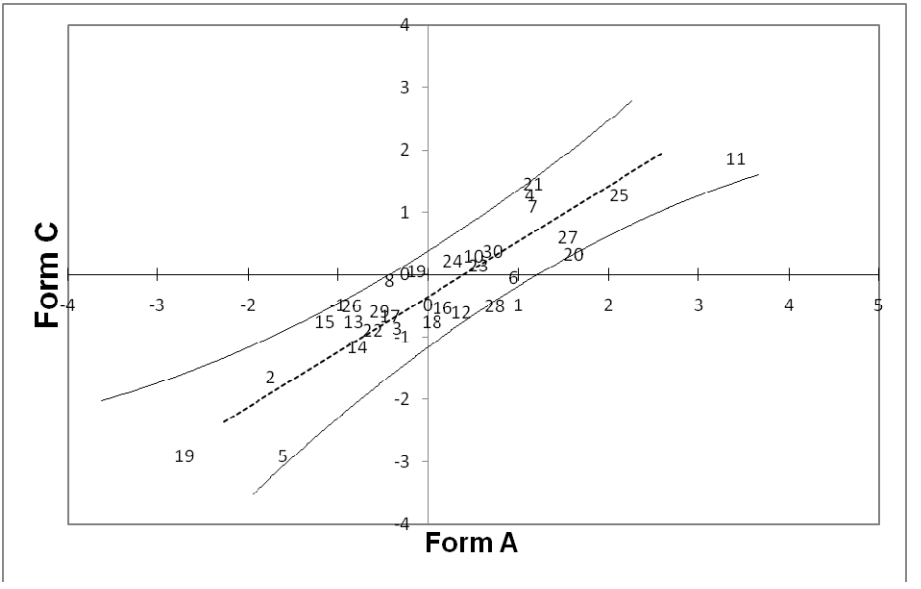
Acceptable infit indices indicated that items in all three test forms were performing well for the test-takers to whom they were targeted. In addition, fairly good outfit indices revealed that all test forms were almost devoid of redundant, dependent or irrelevant items. Thus, it was concluded that the number of options for vocabulary test items did not have any notable impact on the performance of the items. In addition item separation and item reliability indices were calculated for the three test forms. The results are given in Table 2.

Figure 2:
Scatterplot of item parameters.

a) Scatterplot of item parameters from Form A against Form B



b) Scatterplot of item parameters from Form A against Form C



c) Scatterplot of item parameters from Form B against Form C

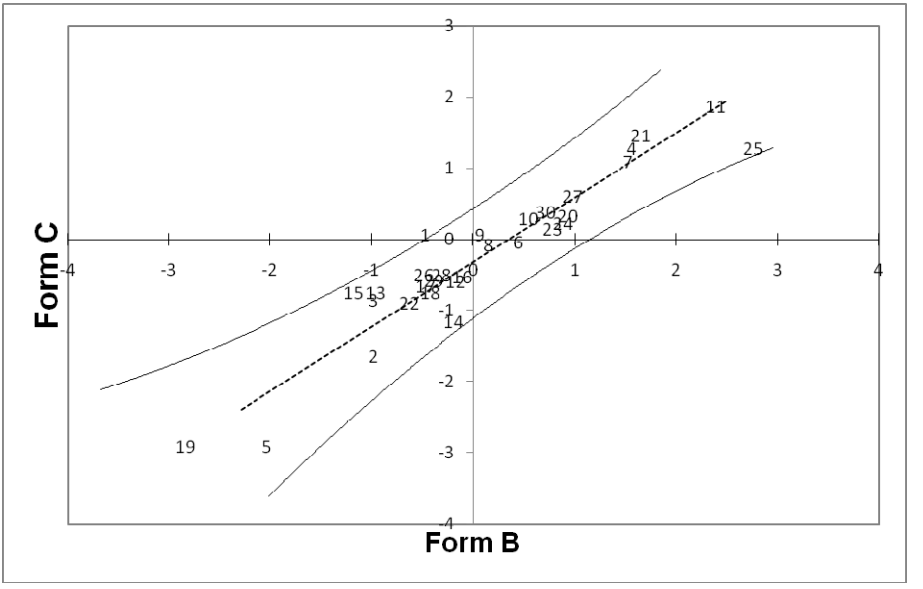


Table 2:
Item separation and reliability for the three forms

	n	Item Separation	Reliability of Items
Form A	30	3.56	0.93
Form B	30	3.50	0.92
Form C	30	3.13	0.91

n = number of items.

Item separation indicates "the number of statistically distinct regions of item difficulty that the persons have distinguished" (Smith, 2001, p. 293). Item reliability, indicating the replicability of item parameter estimates, does not have any traditional equivalent in classical test theory. As indicated in Table 2, there was no notable difference in item separation and item reliability values of the three test forms. Therefore, it was concluded that the number of options for vocabulary test items has no notable impact on item separation and item reliability indices.

Moreover, as indicated in Table 3, a comparison of standard deviations of person ability parameters across the three forms is indicative of no notable difference in the discrimination power of the test forms with different numbers of options per item.

Table 3:
Standard deviations of person ability parameters in the three forms

	n	Standard Deviation
Form A	60	0.80
Form B	60	0.86
Form C	60	0.82

n = number of persons.

4.2 Distracter performance

In order to investigate the performance of the distracters, average measures, outfit mean-squares, and point-measure (PTMEA) correlations were calculated for all the options across the three test forms. Distracter statistics for a handful of the items are given in Table 4.

Table 4:
Distractor statistics for some of the items

ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	%	AVERAGE MEASURE	S.E. MEAN	OUTF MNSQ	PTMEA CORR.	ITEM
20	5	0	1	2	-1.47		.2	-.27	A10
	1	0	22	37	-.17	.16	.9	-.36	
	3	0	4	7	.00	.63	1.6	-.07	
	4	0	7	12	.16	.21	1.1	-.02	
	2	1	26	43	.65	.12	.8	.47	
	MISSING	***	120	67#	.20	.08		-.01	
22	4	0	3	5	-.50	.25	.5	-.21	A12
	1	0	24	41	-.18	.17	.9	-.41	
	5	0	4	7	.39	.40	1.4	.06	
	3	1	27	47	.63	.12	.8	.47	
	MISSING	***	122	68#	.20	.08		-.01	
46	2	0	13	22	-.53	.19	.6	-.39	B6
	1	0	15	25	-.11	.27	1.5	-.14	
	4	0	6	10	.44	.35	1.9	.13	
	3	1	25	42	.47	.12	.9	.37	
	MISSING	***	121	67#	.26	.07		.09	
50	3	0	4	7	-.62	.49	.7	-.22	B10
	1	0	27	45	.05	.13	1.1	-.04	
	4	0	5	8	.18	.53	1.9	.03	
	2	1	24	40	.24	.20	1.4	.14	
	MISSING	***	120	67#	.26	.07		.10	
94	1	0	10	17	-.13	.17	.8	-.24	C24
	3	0	19	32	-.09	.17	.9	-.33	
	2	1	31	52	.70	.14	.9	.49	
	MISSING	***	120	67#	.15	.08		-.09	
95	1	0	17	28	.02	.17	.8	-.22	C25
	2	0	25	42	.35	.15	1.2	.04	
	3	1	18	30	.53	.24	1.4	.17	
	MISSING	***	120	67#	.15	.08		-.09	

Average measures represent the mean of the ability parameter of respondents who choose each distracter. They indicate the discrimination power of distracters. It is expected that distracters be chosen by less able test-takers, thus discriminating between test-takers of high and low ability levels. So the value for average measure should be the highest for the correct option and lower for incorrect options. This was not the case with four items in Form A (five-choice form) and two items in Form B (four-choice form). However, the principle of lower average measure for distracters was not violated in Form C. Therefore, it was concluded that the number of options for vocabulary test items has an inverse effect on the discrimination power of distracters.

Distracter-measure correlations are point-biserial correlations between endorsement (1) and non-endorsement (0) of distracters and test-takers' ability parameters and thus another type of indicators of discrimination power of distracters. Since it is expected that test-takers of low ability choose the distracters (rather than the correct option), negative coefficients of correlation for distracters are desired. The greater the absolute value of negative correlations, the more discriminating the distracter is. Setting -0.2 as criterion, an investigation of point-measure (PTMEA) correlations revealed that 25.33 % of distracters in Form A (five-choice form), 28.33 % of distracters in Form B (four-choice form), and 38.88 % of distracters in Form C (three-choice form) were of quite high discrimination power. So it was concluded that as the number of options per item decreases, the discrimination power of distracters increases. Thus, investigation of PTMEA correlations confirmed the results obtained from the investigation of average measures in that they revealed that the number of options for vocabulary test items has an inverse effect on the discrimination power of distracters.

Finally, outfit mean-square values were investigated. Outfit mean-square values which are greater than 1.3 represent distracters which are ambiguous or written in a different way from the other options. 12.66 % of distracters in Form A, 12.5 % of distracters in Form B, and 13.33 % of distracters in Form C manifested outfit mean-square values greater than 1.3. Since the difference between these values in the three test forms was negligible, it was concluded that the number of options for vocabulary test items has no notable impact on the quality of distracters regarding clarity or relevance in case enough care is taken in the process of test development.

The results revealed that the number of options for multiple choice vocabulary test items has an inverse effect on their discrimination power.

4.3 Person statistics

Investigation of person outfit statistics for the test-takers revealed that out of 180 persons taking the tests, seven of those who had taken Form A, eight of those who had taken Form B, and seven of those who had taken Form C manifested unexpected behavior. That is, they had responded correctly to items with difficulty levels higher than their ability levels, or they had missed easy items with difficulty levels lower than their ability levels. In other words, person outfit indices are indicative of behaviors like guessing and carelessness. Since no trend was found out, it was concluded that the number of options

for multiple-choice vocabulary test items has no notable impact on person response behaviors. Therefore, the number of options for multiple-choice vocabulary test items has no notable impact on person statistics.

4.4 Test reliability

Person separation and Cronbach's Alpha reliability were compared across the three test forms as is indicated in Table 5. Person separation is an indicator of the number of ability strata that a test can distinguish.

As is indicated in Table 5, person separation index of Form B is only slightly higher than the other forms. Therefore, it was concluded that the number of options for multiple-choice vocabulary test items has no notable impact on test reliability.

Table 5:
Reliability and person separation indices for the three forms

	n	Person Separation	Cronbach's Alpha Reliability
Form A	60	1.66	0.77
Form B	60	1.80	0.79
Form C	60	1.67	0.76

n = number of persons.

5. Discussion

The present study sought to determine the impact of the number of options for multiple-choice vocabulary test items on: (a) item statistics, (b) distracter performance indices, (c) person statistics, and (d) reliability of the test.

Three multiple-choice vocabulary test forms were given randomly to 180 participants. Form A was composed of 30 five-choice items, Form B of 30 four-choice items, and Form C of 30 three-choice items. Forms B and C were constructed by randomly deleting one distracter from each item in Form A and Form B respectively. The three test forms were linked using 10 anchor items. The Rasch model was applied to analyze the data concurrently.

No statistically significant difference was observed among the three test forms regarding item difficulty estimates ($p < 0.05$). Moreover, no notable difference was observed among the three test forms regarding the fit indices and reliability of the items. However, the results revealed that the number of options for vocabulary test items has an inverse effect on the discrimination power of distracters while the fit remained unaffected. Surprisingly, person response behaviors were indicated to be identical across the three test forms with different numbers of options per item. This finding contradicts the commonly held idea that fewer numbers of options increase the chance of guessing the correct op-

tion which leads to decreasing the reliability of the test. The point is that test-takers often resort to guessing when either the items are too difficult or they do not have enough time to go through all the items. In addition, the results revealed no notable change in the reliability of three test forms with different numbers of options per item.

The results of the current study revealed no notable difference in the psychometric properties of tests comprised of five-choice, four-choice, and three-choice items, except for a slight increase in the discrimination power of distracters as a result of reducing the number of options. As a result, considering the amount of time and effort needed to develop multiple-choice vocabulary tests with more options per item, three options seem to be optimal for vocabulary test items. Though there exist some differences regarding the effects of the number of options on distracter, item and test characteristics in numerous studies investigating the issue, a wide majority of them recommend the use of three options per item. For example, Straton and Catts (1980), Aamodt and McShane (1992), and Rodriguez (2005) reported a decrease in item difficulty as a result of reducing the number of options per item while Landrum, Cashin, and Thesis (1993) reported either an increase or no change in item difficulty as a result of reducing the number of options per item. Yet, Shizuka, Takeuchi, Yashima, and Yoshizawa (2006), as well as the present study detected no significant difference regarding item difficulty between tests with different numbers of options per item. However, considering a number of psychometric properties together as well as practicality-related issues, all these studies converge on the use of three options per item.

Thus, the findings of the present study regarding the optimal number of options for multiple-choice vocabulary test items corroborate the previous theoretical (e.g., Bruno & Dirkzwager, 1995; Grier, 1975; Lord, 1977; Tversky, 1964) as well as empirical findings (Aamodt & McShane, 1992; Delgado & Prieto, 1998; Haladyna & Downing , 1993; Landrum, Cashin, & Thesis, 1993; Rodriguez, 2005; Shizuka, Takeuchi, Yashima, &Yoshizawa, 2006; Sidick, Barrett, & Doverspike, 1994 to name a few) which recommend the use of three options per item.

Criticism is usually made of using fewer options per item due to enhancing the probability of guessing. However, as the results of the current study revealed, multiple-choice tests, regardless of their number of options per item, would remain almost immune to the effect of guessing factor when the items are appropriately targeted for the group of test-takers, and enough time has been allotted. Furthermore, using three options per item would allow for the inclusion of more items in a fixed period of time, thus increasing the reliability and validity of the test.

The finding that 3, 4 and 5-option multiple-choice items were not significantly different in terms of difficulty is interesting. Although theoretically speaking the chances of getting the items right without being familiar with the construct measured in tests with three, four and five options are 33 %, 25 % and 20 % respectively, in practice we observed that the chance factor had no influence on item difficulties across the tests.

As was mentioned before, the distracters which were omitted from the 5-option test to develop the 3 and 4-option tests were entirely randomly selected. That is, no factor was considered for option deletion. Some of the distracters which were deleted to make the 3

and 4-option items had been chosen frequently and some rarely by the test-takers. In fact, we decided to randomly delete distracters for the construction of 3 and 4-option tests to make sure that any obtained results would be attributable to the effect of only and only the number of options and not the effect of eliminating malfunctioning distractors. Therefore, the claim that 'the 3 and 4-option items were as good as the 5-option items because non-functioning distracters in 4 and 5-option items were deleted' is not justified.

One possible reason for the good performance of the 3-option items could be the care with which the items were written. As was mentioned above, the distracters were chosen out of the wrong answers provided by the test-takers in a free-response format of the test. The preliminary versions of the test were edited and revised by experienced English teachers and educated native speakers several times. Distracters which were flagged as ambiguous or thought might be eliminated easily by the test-takers were replaced by plausible ones. Afterwards, the preliminary version of the test was piloted and 30 Rasch-model-fitting items were selected out of the pool of 50 items for the study.

Moreover, the Wright map showed that items were of appropriate level of difficulty for the test-takers and were very well-targeted. This could be another reason why the 4 and even the 3-option items were as good as the 5-option items. Guessing usually occurs when the items are above the ability level of the test-takers or when the test is administered under time constraints (speed tests). In this study, the time allotted to complete the test was enough. Therefore, one can argue that if 3-option multiple choice test items are carefully constructed, are at the appropriate level of difficulty and are administered as power tests and not speed tests, they can be as difficult as 4 and 5-option tests and as efficient in terms of discrimination, reliability and fit to the Rasch model.

It is important to note that the Rasch model estimation method used to analyse the data was joint maximum likelihood estimation (JMLE) (Wright & Panchapakesan, 1969). JMLE has been criticized for being biased and statistically inconsistent under certain conditions (Jansen, van den Wollenberg & Wierda, 1988; van den Wollenberg, Wierda & Jansen, 1988). One possible reason for the results observed in this study could be the JMLE estimation method. Replications of the present study with more consistent estimation methods such as conditional maximum likelihood estimation method are needed to ascertain that the estimation method was not the reason.

In conclusion, the present study generally recommends the use of three options for multiple-choice vocabulary test items. However, its application to all testing situations should not be taken for granted. Different purposes are followed in different testing situations, and the intended purposes dictate the format of the tests.

The theoretical implications carried by the present study contribute to dispelling the common misconception that multiple-choice tests composed of items which contain less than four or five options per item would be seriously affected by the guessing factor. Moreover, the results of this study confirm Haladyna's (2004) correction of the frequently cited item-writing guideline suggesting test developers to write as many plausible distracters as possible and to "use as many choices as possible, but three seems to be a natural limit" (p. 112).

A number of pedagogical implications can be drawn from the findings of the present study. In case three options are provided for each item, the quality of multiple-choice tests will be considerably improved since the test developers will no longer resort to implausible distractors for the sake of following the custom of providing four or five options per item. Furthermore, the probability of providing clues which would favor test-wise students is extremely decreased. Moreover, less time will be devoted to test construction. Test administration would also take less time (Aamodt & McShane, 1992; Straton & Catts, 1980). Time saved from the processes of test development and administration can be devoted to instruction purposes. In addition, given a fixed period of time, more items can be included in the test which would improve content validity as well as reliability of the test.

Since the psychometric properties of multiple-choice vocabulary tests with five, four, and three options per item were indicated to be almost identical, this study highly recommends the use of three options for multiple-choice vocabulary test items due to practicality-related issues. Considering issues like saving time, money, and energy, reducing the risk of providing implausible distracters, and reducing the probability of providing clues which would favor test-wise test-takers, one may concur with the use of three options for multiple-choice vocabulary test items in most cases.

References

Aamodt, M. G., & McShane, T. (1992). A meta-analytic investigation of the effect of various test item characteristics on test scores and test completion times. *Public Personnel Management, 21*(2), 151-160.

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Baghaei, P., & Amrahi, N. (in press). Validation of a multiple choice English vocabulary test with the Rasch model. *Journal of Language Teaching and Research, 2*(5).

Bruno, J. E., & Dirkzwager, A. (1995). Determining the optimal number of alternatives to a multiple-choice test item: An information theoretic perspective. *Educational and Psychological Measurement, 55*,959-966.

Budescu, D. V., & Nevo, B. (1985). Optimal number of options: An investigation of the assumption of proportionality. *Journal of Educational Measurement, 22,* 183-196.

Crehan, K. D., Haladyna, T. M., & Brewer, B. W. (1993). Use of an inclusive option and the optimal number of options for multiple-choice items. *Educational and Psychological Measurement, 53,* 241-247.

Delgado, A. R., & Prieto, G. (1998). Further evidence favoring three-option items in multiple-choice tests. *European Journal of Psychological Assessment, 14,* 197-201.

Educational Testing Service (1987). *Reading for TOEFL.* Princeton, NJ: Author.

Educational Testing Service (1989). *Listening to TOEFL.* Princeton, NJ: Author.

Grier, J. B. (1975). The number of alternatives for optimum test reliability. *Journal of Educational Measurement, 12*, 109-112.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd Ed.). Mahwah, NJ: Lawrence Erlbaum.

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking.* Boston: Allyn & Bacon.

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

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

Hohensinn, C. H., & Kubinger, K. D. (2009). On varying item difficulty by changing the response format for a mathematical competence test. *Austrian Journal of Statistics, 38*(4), 231-239.

Jansen, P. G. W., van den Wollenberg, A., & Wierda, F. W. (1988). Correcting unconditional parameter estimates in the Rasch model for inconsistency. *Applied Psychological Measurement, 12*, 297-306.

Kubinger, K. D., Holocher-Ertl, S., Reif, M., Hohensinn, C., & Frebort, M. (2010). On minimizing guessing effects on multiple-choice items: Superiority of a two solutions and three distractors item format to a one solution and five distractors item format. *International Journal of Selection and Assessment, 18*(1), 111-115.

Kubinger, K. D., & Gottschall, C. H. (2007). Item difficulty of multiple choice tests dependent on different item response formats – An experiment in fundamental research on psychological assessment. *Psychology Science Quarterly, 49*(4), 361-374.

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

Linacre, J. M. (2008). *A user's guide to WINSTEPS-MINISTEP: Rasch-model computer programs.* Chicago, IL: winsteps.com.

Linacre, J. M. (2008). WINSTEPS® (Version 3.66.0) [Computer Software].Chicago, IL: winsteps.com.

Lord, F. M. (1977). Optimal number of choices per item – a comparison of four approaches. *Journal of Educational Measurement, 14*, 33-38.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Menlo Park, CA: Addison-Wesley.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Expanded Ed.). Chicago: University of Chicago Press.

Rodriguez, M. C. (2002). Choosing an item format. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 213-231). Mahwah, NJ: Lawrence Erlbaum.

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.

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,* 234-247.

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.

Sidick, J. T., Barrett, G. V., & Doverspike, D. (1994). Three-alternative multiple-choice tests: An attractive option. *Personnel Psychology, 47,* 829-835.

Smith, E. V., Jr. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 2,* 281-311.

Straton, R. G., & Catts, R. M. (1980). A comparison of two, three, and four-choice item tests given a fixed total number of choices. *Educational and Psychological Measurement, 40,* 357-365.

Tversky, A. (1964). On the optimal number of alternatives at a choice point. *Journal of Mathematical Psychology, 1,* 386-391.

van den Wollenberg, A., Wierda, F. W., & Jansen, P. G. W. (1988). Consistency of Rasch model parameter estimation: A simulation study. *Applied Psychological Measurement, 12: 307-313.*

Wolfe, E. W., & Smith, E. V., Jr. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II – Validation activities. *Journal of Applied Measurement, 8,* 204-234.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29,* 23-48.