Language Testing in Asia
a SpringerOpen Journal

RESEARCH Open Access

The number of options in multiple choice items in language tests: does it make any difference? Evidence from Thailand

Inadaphat Thanyapa[1†] and Michael Currie[2*†]

* Correspondence:
miccurrie@hotmail.com
[†]Equal contributors
[2]AUA Language Center, Platha Road, Muang, Songkhla 90000, Thailand
Full list of author information is available at the end of the article

Abstract

The overall study from which findings are presented gave stem-equivalent short answer and multiple choice tests of English structure and reading to students at Prince of Songkla University, Thailand. A comparison of scores, facility, discrimination, reliability and validity from 3-, 4- and 5-option versions of the multiple choice test found little variation in the performance of the three option formats although higher numbers of options tended to increase difficulty and to spread responses between the options. Analysis of the responses offered by the participants in the short answer and multiple choice tests highlighted the areas where the variation in the responses arose and also pointed to some differences in the way that the multiple choice format affected the structure and reading sections. Major differences were also noted between the answers offered by the participants to the short answer test and their option selections in the multiple choice test, with at least 60% of the responses changing between the two tests. The researchers conclude that option number format is of little importance in the measurement of language ability compared with the much greater construct variation apparently induced by the use of multiple choice items as against constructed response formats.

Keywords: English structure test; English reading test; Multiple choice; Number of options; Constructed response; Short answer

Background

The findings described in this paper were derived from a wider study conducted in Thailand, examining the performance and validity of different item formats in tests of English structure and reading ability and how item format affects the measurement of language based constructs. The study looked particularly at differences between the performance of stem-equivalent items in the constructed response, short answer format and multiple choice items with 3- 4- or 5-options.

Despite the widespread use of multiple choice items as a method of measuring language learning, there has been relatively little research into their validity for that purpose and justifications for their use tend to concentrate on their reliability and practicality in large-scale testing situations.

One area of their operation which has however, received regular attention has been the question of the optimal number of options in multiple choice items, and there have

been a number of theoretical and empirical studies which have sought to justify the use of 3-option items both based on their psychometric efficiency as well as on the grounds of being easier and less time consuming to construct because of the relative difficulty of writing large numbers of plausible distractors.

The section of the study described here addressed the effect of using 3-, 4- and 5-option items and investigated if there were significant differences in the validity, reliability, difficulty, and discrimination ability of stem-equivalent items associated with differing numbers of options. The study also considered how the effectiveness of the distractors included in the multiple choice items was affected by the numbers of options in the items, as well as examining whether any effects noted differed between structure and reading items.

The main study also investigated how the test takers' responses varied between the two tests they took, the first in the constructed response, short answer format and the second using stem-equivalent multiple choice items. This paper also reports on how those changes were affected by the number of options in the multiple choice items as well as comparing the effect on validity and performance noted among the multiple choice items with differing numbers of options, as against the effects noted between items in multiple choice and the constructed response, short answer format.

These research topics were addressed using the following analysis methods:

- Concurrent validity: Correlation with criterion reference data.
- Reliability: Comparison of Fisher z coefficient confidence intervals
- Differences in item difficulty between items in different multiple choice option formats: Facility indices compared using ANOVA
- Differences in Discrimination ability: Discrimination indices calculated using the 27% method and R_{pbis} coefficients compared using ANOVA or Friedman tests and a Wilcoxon signed rank test.
- Effect on distractor performance: various methods of distractor analysis detailed in Haladyna and Downing (1993) plus the AENO coefficient to compare the spread of option selections.
- Consistency and changes in option selections between constructed response and multiple choice tests: Percentage based data classified according to a taxonomy of possible patterns and compared using ANOVA and *t* tests.

Research context and importance

Although there is widespread agreement that it is desirable for Thais at all levels of society to acquire useable English language skills, it has often been noted that Thai students generally do not learn to use English as a means of communication. Many studies into aspects of why this is so (see for instance Musigrungsi, 2002; Prapaisit, 2003 and Thongsri, 2005) have linked this fact with the way in which English is taught and tested in formal education. Researchers have particularly noted a concentration by teachers on teaching knowledge and abilities, for instance grammar and reading skills, viewed as being useful for passing high stakes tests such as the university entrance (O-NET) examinations, rather than attempting to teach communicative skills.

Despite attempts to broaden the spectrum of testing methods used in Thailand the multiple choice item continues to represent the most frequently used format (Prapphal,

2008; Watson Todd, 2008). In a nationwide survey, (Piboonkanarax cited in Watson Todd, 2008) multiple choice items were found to account for around 50% of school grades in English and as of 2014, the O-NET examinations continue to be entirely based on multiple choice items.

The findings of this study are therefore of considerable importance since they shed light on the effectiveness of the multiple choice format as a means of measuring educational achievement as well as suggesting that the number of options in multiple choice items is a matter of small importance when judged against far greater effect on measurement accuracy of using multiple choice items as against items in a constructed response format.

Previous research into the optimal number of options in a multiple choice item

Research into the optimal number of options dates back over 80 years and Rodriguez (2005) in his meta-analysis, identified 48 studies, the earliest of which was published in 1925. On the theoretical level, Tversky (1964), (Grier 1975 & 1976), Lord (1977) and Bruno & Dirkzwager (1995) all used mathematical formulations to conclude that 3-option tests have superior psychometric properties although Lord also suggested that for high ability test takers, two options would work best with 5-option items being more efficient for low ability candidates. An underlying assumption in these studies was that the time available would remain proportional to the total number of options and Ebel (1969) hypothesized that the reliability of a test will increase with the number of options per item where the number of items remains constant.

However, Budescu and Nevo (1985), in a partly theoretical and partly empirical study, attacked the assumption of proportionality based on recording the time taken to complete tests using different option format items. From a 5-option university entrance examination of English vocabulary and reading, verbal reasoning and mathematics, they created 4-, 3- and 2-option versions by successively deleting options. Using a group of 1000 Israeli subjects they found that the 2-option version of the test was easiest and least reliable with the difficulty of the formats increasing with the number of options, and concluded that the optimum number of options was greater than three.

Although many studies have considered the optimal number of options, there have only been a small number of empirical studies relating to option format in tests of language learning. Ramos & Stern (1973) compared 4- and 5-option tests of reading in Spanish and French as foreign languages, finding a small but significant reduction in reliability in reducing items from five to four options as well as a small decrease in discrimination. There was also a reduction in 'formula scores' (i.e. scores adjusted to eliminate the effect of guessing) associated with the move from 5- to 4-option items.

Green, Sax and Michael (1982) created a 5-option multiple choice test for French beginners which was then used to create 3- and 4-option tests by eliminating distractors. The study's research hypotheses were based on Lord's (1977) prediction concerning the performance of items based on test takers' ability levels. However the findings in the main did not support the superiority of lower numbers of options for higher ability test takers, nor of higher option numbers for lower ability levels, and there was little difference in difficulty or reliability across the three formats. The researchers were therefore unable to support Lord's prediction, suggesting that this was because the abilities of the

subjects were too close and that the items in the tests may have been too easy, a situation they felt was not untypical of classroom tests.

The issue was also dealt with in a study in Japan by Shizuka, Takeuchi, Yashima and Yoshizawa (2006) which compared the performance of two groups of test takers in two tests based on stem-equivalent 3- and 4-option items derived from the English section of a university entrance examination. Because the groups they compared were found to be of unequal ability, the researchers used a one parameter Rasch model analysis in which the results of the two groups were equated using a number of common items, and their conclusion that there was no difference in the difficulty of the items in 3- and 4-option format was based on a comparison of the mean item difficulty measures disregarding the common items, although their logic in applying this comparison appears doubtful, based as it was on two norm referenced sets of scores centered on a mean logit item difficulty of zero. However, their conclusion was that 3-option tests are not significantly different in either their difficulty or their discrimination ability to the 4-option equivalents nor was the reliability of the test markedly decreased by deleting one option from the original 4-option test.

More recently the effect of varying the number of options in practice CSAT English listening tests in Korea was considered by Lee and Winke (2012). They found significant differences in difficulty between 3- and 4-, and between 3- and 5-option versions of the three tests adapted to produce the different option versions, but found no differences between the 4- and 5-option versions. They also found no significant differences in the discrimination coefficients of the different option versions, but for two of the three tests used they found the 3-option version to be significantly more reliable than the 4-option version, while for the third test they found the 5-option test to be more reliable than the 4-option test. Interestingly they concluded that the KICE who administer the CSAT tests were probably correct in adopting the 5-option format in the high stakes CSAT test but that they should not overlook the test-takers' perspective that 3-option tests were easier and less stressful, and that items with 3-options took less time to answer.

The most thorough review to date of empirical studies into the optimal number of options in multiple choice items in all domains was the meta-analysis carried out by Rodriguez (2005). He identified 48 studies relating to the performance of items with different numbers of options of which 27 were selected to be part of the meta-analysis, producing a total of 56 comparisons between two tests consisting of items with different numbers of options. The findings were that all reductions in option numbers resulted in significant decreases in item difficulty (i.e. the items were made easier) and except in the case of a reduction from five to three options, resulted in significant changes in item discrimination ability (both increases and decreases). The greatest reductions in discrimination were noted in reducing to two options and in one case (four to three options) there was a significant increase in discrimination.

Rodriguez' conclusion was that generally the meta-analysis supported three as being the optimal number of options. Nevertheless, to support his endorsement of 3-option items, he felt it necessary to review the practical arguments in their favour to support their worth, notably that less time is needed to prepare two plausible distractors than three or four distractors, and that more 3-option items can be administered per unit of time than 4- or 5-option items, potentially improving content coverage. Generally in studies supporting the use of 3-option items, there is a marked tendency for researchers to cite the savings and other benefits which can be achieved by reducing tests

to 3-option items. As Sidick, Barrett & Doverspike (1994) estimate, there would be a saving of 5 minutes per test item in item-writing time by reducing tests from four to three options and as Shizuka et al. (2006) note, concomitant savings in stationery, printing and test administration costs.

However, as Rodriguez (2005) points out, few of the studies have addressed the issue of validity and a far more realistic conclusion from many of the studies is that their findings tend to show that reducing the number of options makes very little difference to the quality of the items as tools for measuring educational achievement beyond predictably and favorably affecting the test takers' ability to answer items where their knowledge is less than complete, with the expected upward effect on item facility.

Methods

The study from which the results reported herein were drawn, was conducted using groups of first and third year undergraduate students from the Faculty of Liberal Arts at Prince of Songkla University, Thailand. Although in accordance with the University's policy on the ethical conduct of research, approval for the use of the participants was sought from and granted by the Associate Dean for Academic Affairs. Although participation was voluntary, the overall sample was selected purposively with the aim of achieving a normal distribution of scores across the main study instrument, a test of English structure and reading ability administered in four different formats; firstly as a constructed response, short answer test and later as three multiple choice tests with items in 3-, 4- and 5-options. In each test the item stems were identical, the only difference being in the number of options.

The items were written specially for the test, based on language and reading texts appropriate for students at or slightly above the level of English generally found in students taking the O-NET examination. The 52 items used in the study and the two texts on which the reading section was based, were drawn from a larger pool of 172 items and 8 texts which were piloted using 178 first year medical students from the same university. Following pilot testing 40 structure and 12 reading items based on 2 texts (one linear, one non-linear) were selected which appeared in all forms of the experimental test used in the study. (See Additional file 1: Appendix A for samples of the item types used.)

The test was first administered in July 2007 to the 216 subjects detailed below in Table 1a, in the constructed response, short answer format, in which the participants were asked to record their own answer to each of the items. The answers offered by the participants were then recorded and the number of times a response was offered in an item was established. Following this, three multiple choice tests in 3-, 4- and 5-option format were constructed. For the 3-option test, the options consisted of the expected response identified during the item writing process together with two distractors, being the two incorrect responses which had been offered the most number of times in the short answer test. The 4-option test was constructed by adding to each 3-option item, the third most popular incorrect response and the 5-option item also included the fourth most popular incorrect response from the short answer test.

To aid comparisons between the short answer and multiple choice tests (not dealt with in detail in this paper) eight structure items and four reading items were designated as control items and appeared in all forms of the multiple choice tests (3-, 4- and 5-option) in 4-option format.

Table 1 Research participants

	Major/course	English for business	Thai/English language	Chinese	Community development	Total
	Year	Third	First	First	First	
a.	Short answer	21	73	62	60	216
b.	Multiple choice: 3-option	7	16	20	9	52
	4-option	7	17	22	9	55
	5-option	7	16	17	5	45
	Total multiple choice	21	49	59	23	152
c.	Mean scores	Structure		Reading	Overall	
	3 option group ($N = 52$)	7.077		4.096	11.173	
	4-option group ($N = 45$)	7.364		4.000	11.364	
	5-option group ($N = 55$)	7.578		4.244	11.822	

a. Subjects who sat 1st test (short answer) in July 2007.
b. Subjects who sat 2nd tests (multiple choice) in August and September 2007.
c. Mean scores of the multiple choice option groups in the short answer test.

Then, in August/September 2007, the three multiple choice tests in 3-, 4- and 5-option format were administered to groups of respectively 52, 45 and 55 experimental participants drawn from the original 216 who had sat the first test, detailed above in Table 1b. The different numbers of participants in the groups resulted from the subjects from the first round of testing being hypothecated in advance to different forms of the multiple choice test, in an effort to achieve equivalent ability groups. However, all participation in the study was voluntary and not all of the students who participated in the first round of testing participated in the multiple choice tests. Nevertheless, three one-way ANOVAs on the structure section, reading section and the overall scores of the three groups in the short answer test found no significant differences between the groups, establishing that they were of comparable ability (structure: $F = 0.085,\ p > 0.05$; reading: $F = 0.140,\ p > 0.05$; overall: $F = 0.087,\ p > 0.05;\ df = 2,\ 149$)

Following the marking of the multiple choice tests, the participants' option selections were analysed and comparisons were conducted of the validity, reliability and difficulty of the three different option format tests as well as the discrimination ability of the items in their those option formats.

Further, since to a large extent the operation of a multiple choice item depends upon the effectiveness of its distractors, their performance was subjected to a number of different analyses following the methods suggested in Haladyna and Downing (1993) to investigate whether distractor performance was influenced by the number of options per item. In addition the overall spread of selections of the item options was analysed using the AENO coefficient (Sato and Morimoto, cited in Shizuka et al., 2006).

Analysis was also conducted of the effect which varying the number of options per item had as between the answers offered in the short answer test against the options selected in the three multiple choice tests.

All the comparisons were based on 31 stem-equivalent structure items and 8 stem-equivalent reading items answered by the three groups. The outcome of the control items, which were all in 4-option format, was disregarded as was that of one item in the structure section (item #6), which following adjudication of the answers given in

the short answer test by a panel of three umpires, was found to have been constructed in multiple choice format with two acceptable answers among the options.

Results

Concurrent validity

Consideration of the validity of the three multiple choice item formats detailed in this paper was primarily aimed at showing that the construct being measured by the experimental tests was broadly similar to that measured by the O-NET entrance examination and was therefore confined to a comparison of the correlations between the overall scores of the 131 first year subjects among the three groups in the O-NET examination and their experimental multiple choice test scores. Thus the criterion reference data against which comparison of the experimental tests were made was also a multiple choice test (in 4-option format) whose scope was somewhat wider than the experimental test, incorporating three sections, Language Use and Usage, Writing and Reading. The broader question as to whether the multiple choice tests were a valid measure of the language ability demonstrated by the answers to the items in their short answer format is not dealt with in detail in this paper but the interested reader is referred to Currie (2008); Currie and Chiramanee (2010) for consideration of this issue.

As can be observed from Table 2 below, the correlation coefficients between the first year participants' O-NET scores and their respective scores in the three multiple choice experimental tests were all significant at the 0.001 level. That for the 5-option test was lowest with the 3-option group producing the highest figure despite the criterion reference test itself being in 4-option format.

Comparison of confidence intervals calculated based on the Fisher z conversion established that none of the differences in the three coefficients were however significant at the 0.05 level.

Reliability

The Cronbach alpha coefficients for the structure and reading sections separately, as well as for the overall scores of the two sections combined are shown in Table 3. The coefficients for the structure sections of the three tests and also overall were very close, lying between 0.87 and 0.89. The coefficients for the reading section were somewhat lower and more widely separated largely due to the small number of items in the reading sections. No one test produced consistently superior coefficients. Overall, the 5-option test produced the highest coefficient; for the structure section the highest

Table 2 Concurrent validity of experimental tests based on correlation between first year participants' O-NET scores and scores in multiple choice tests

	3-option test	4-option test	5-option test
n (1st year students only)	45	48	38
Mean O-NET score (%)	44.29	44.52	45.97
Mean multiple choice test score (%)	47.98	41.77	39.14
Correlation coefficient (*r*)	0.790	0.759	0.634
significance ($p <$)	0.001	0.001	0.001
$P < 0.05$ confidence interval of correlation coefficients	0.746 – 0.879	0.606 – 0.858	0.394 – 0.793

Table 3 Cronbach alpha reliability coefficients of the experimental tests

	No. of items	3-option test	4-option test	5-option test
Structure section	31	0.872	0.890	0.885
Reading section	8	0.532	0.375	0.444
Overall	39	0.887	0.887	0.890
$p < 0.05$ Confidence intervals:				
Structure section		0.786 – 0.925	0.818 – 0.935	0.799 – 0.935
Reading section		0.303 – 0.703	0.121 – 0.583	0.173 – 0.652
Overall		0.810 – 0.934	0.813 – 0.933	0.807 – 0.938

coefficient was found for the 4-option test with the 3-option test producing the highest reading section coefficient.

However, comparison of confidence intervals calculated based on the Fisher z conversion established that none of the differences in the three sets of coefficients were significant at the 0.05 level.

Overall difficulty based on participants' scores and item facility

Table 4 includes details of the mean scores achieved by the three groups in respectively the structure and reading sections. Visually, the scores achieved by the 3-option group were higher than those of the 4-option group which were in turn greater than those of the 5-option group. However, one-way ANOVAs showed that there were no significant differences between the scores of the three groups (structure: $F = 1.472$, $p > 0.05$; reading: $F = 2.098$, $p > 0.05$; $df = 2, 149$).

Table 5 shows the facility index values based on the proportion of the number of correct option selections in each group against the total number of participants in that group, for each of the 31 structure and 8 reading items individually. A one-way ANOVA conducted on the three sets of individual structure item values found no significant difference between them ($F = 1.360$, $p > 0.05$, $df = 2, 90$). A one-way ANOVA on the reading item facility index values also produced no indications of significant differences ($F = 0.258$, $p > 0.05$ $df = 2, 21$) although it was notable that the facility index values for all eight of the 3-option reading items exceeded those of the 5-option items

Table 4 Mean scores and discrimination indices of the experimental tests

		3-option test	4-option test	5-option test
Mean score	Structure (/31)	16.19	14.40	13.98
	Reading (/8)	4.19	3.85	3.51
	Overall (/39)	20.38	18.25	17.49
SD of mean scores	Structure	6.57	7.12	6.89
	Reading	1.77	1.51	1.62
	Overall	7.87	7.94	7.97
Mean	Structure			
Discrimination	R_{pbis}	0.45	0.48	0.54
Coefficients	DI	0.53	0.56	0.47
	Reading			
	R_{pbis}	0.47	0.43	0.47
	DI	0.54	0.45	0.42

Table 5 Facility index values for the structure (1 – 40) and reading (41 – 51) experimental items

Item #	3-option test	4-option test	5-option test	Item #	3-option test	4-option test	5-option test
	Structure items			**Structure items (cont.)**			
1	0.63	0.69	0.67	28	0.44	0.27	0.27
2	0.35	0.36	0.38	29	0.10	0.18	0.16
4	0.79	0.67	0.64	30	0.46	0.38	0.38
5	0.44	0.38	0.27	31	0.63	0.55	0.51
7	0.27	0.20	0.20	32	0.56	0.44	0.47
8	0.85	0.87	0.78	34	0.38	0.29	0.33
10	0.40	0.47	0.42	35	0.52	0.44	0.42
11	0.25	0.20	0.13	37	0.35	0.44	0.47
13	0.50	0.33	0.56	38	0.62	0.42	0.38
14	0.50	0.53	0.44	39	0.60	0.58	0.62
15	0.81	0.67	0.62	40	0.29	0.36	0.18
16	0.75	0.60	0.62	**Reading Items**			
17	0.77	0.73	0.73	41	0.58	0.56	0.53
18	0.58	0.60	0.73	43	0.48	0.33	0.42
19	0.40	0.45	0.47	45	0.54	0.49	0.44
21	0.75	0.55	0.60	46	0.35	0.36	0.24
22	0.48	0.33	0.22	48	0.92	0.98	0.84
23	0.73	0.55	0.36	49	0.42	0.24	0.22
25	0.75	0.51	0.60	50	0.71	0.71	0.67
26	0.25	0.36	0.36	51	0.19	0.18	0.13
Structure section				**Reading section**			
Mean *p*	0.52	0.46	0.45		0.52	0.48	0.44
S. Dev.	0.19	0.17	0.18		0.22	0.27	0.24

Note: Items, 3, 9, 12, 20, 24, 27, 33, 36, 42, 44, 47 and 52 were control items which did not form part of the experimental test. Item 6 was excluded from analysis as it was found to be faulty. Similar comments apply to Tables 6 and 7 below.

indicating a tendency for the 3-option reading items to be easier than their 5-option counterparts.

Discrimination ability

Two methods of assessing discrimination ability were adopted. Firstly, point biserial (R_{pbis}) coefficients were calculated for each item based on the scores achieved by the groups in their respective sections (i.e. separately for the structure and reading items). Secondly, the members of each group were divided into higher and lower ability groups of 27% each, separately for the structure and reading sections, and a discrimination index (DI) for each item was established based on the number of correct answers in the upper group, less the number correct in the lower group, the product being divided by the participants in each of the higher and lower groups. Table 4 includes the mean R_{pbis} and DI coefficients for the structure and reading sections.

Table 6 shows the individual item DI and untransformed R_{pbis} values. Prior to analysis, the R_{pbis} coefficients were standardized using the Fisher z transformation. The two sets of coefficients (DI and z-transformed R_{pbis}) were then checked for the normality of their distribution. It was found that while the reading item coefficients appeared

Table 6 Discrimination indices for the structure experimental items

Structure	3-option		4-option		5-option	
Item	**DI**	**R_{pbis}**	**DI**	**R_{pbis}**	**DI**	**R_{pbis}**
1	0.71	0.48	0.33	0.24	0.37	0.46
2	0.64	0.53	0.47	0.50	0.67	0.85
4	0.29	0.25	0.73	0.58	0.37	0.46
5	0.57	0.51	0.80	0.65	0.08	0.00
7	0.36	0.30	0.40	0.58	0.74	0.69
8	0.36	0.36	0.40	0.43	0.38	0.54
10	0.43	0.47	0.67	0.59	0.68	0.77
11	0.64	0.60	0.40	0.49	0.58	0.46
13	0.79	0.58	0.60	0.58	0.52	0.62
14	0.79	0.64	0.80	0.63	0.67	0.77
15	0.36	0.43	0.67	0.54	0.49	0.69
16	0.50	0.42	0.53	0.45	0.52	0.62
17	0.71	0.63	0.73	0.59	0.51	0.69
18	0.29	0.26	0.27	0.18	0.32	0.31
19	0.79	0.58	0.87	0.67	0.63	0.77
21	0.64	0.57	0.87	0.63	0.60	0.77
22	0.86	0.62	0.60	0.54	0.61	0.62
23	0.43	0.44	0.33	0.27	0.30	0.23
25	0.50	0.52	0.80	0.53	0.40	0.38
26	0.64	0.58	0.73	0.62	0.57	0.62
28	0.14	0.13	0.40	0.46	0.26	0.31
29	0.21	0.27	0.33	0.43	0.38	0.31
30	0.71	0.58	0.80	0.61	0.56	0.69
31	0.64	0.55	0.80	0.62	0.54	0.69
32	0.50	0.42	0.40	0.31	0.50	0.54
34	0.71	0.57	0.40	0.47	0.63	0.69
35	0.71	0.53	0.67	0.53	0.54	0.62
37	0.29	0.21	0.27	0.30	0.47	0.62
38	0.50	0.42	0.60	0.53	0.36	0.46
39	0.71	0.63	0.60	0.47	0.47	0.62
40	−0.07	−0.06	0.00	−0.03	−0.12	−0.15
Mean	0.53	0.45	0.56	0.48	0.47	0.54
Reading	**3-option**		**4-option**		**5-option**	
Item	**DI**	**R_{pbis}**	**DI**	**Item**	**DI**	**R_{pbis}**
41	0.57	0.52	0.87	0.67	0.57	0.69
43	0.79	0.64	0.67	0.51	0.55	0.69
45	0.93	0.72	0.67	0.51	0.53	0.62
46	0.50	0.45	0.47	0.38	0.34	0.46
48	0.29	0.32	0.07	0.26	0.45	0.31
49	0.50	0.42	0.33	0.43	0.34	0.46
50	0.79	0.60	0.27	0.26	0.49	0.54
51	0.00	0.11	0.27	0.39	0.10	0.00
Mean	0.54	0.47	0.45	0.43	0.42	0.47

to be normally distributed, some of those for the structure items indicated non-normal distribution. In particular, the z coefficient calculated by dividing the skewness and kurtosis statistics by their respective standard errors produced figures of 3.26 and 3.44 for the 5-option structure test DI coefficients (normally distributed values should produce z coefficients between −1.96 and +1.96), and for the R_{pbis} coefficients in the 3- and 4-option tests, the skewness z statistics were respectively 2.273 and 2.69. Further, all 3 sets of coefficients produced Shapiro Wilks test statistics significant at the 0.05 level (DI 5-option, $p = 0.009$; R_{pbis} 3-option, $p = 0.022$; R_{pbis} 4-option, $p = 0.015$; $df = 31$. Accordingly while the reading section discrimination coefficients were analysed using parametric one-way ANOVAs, the structure section coefficients were analysed using non-parametric Friedman tests.

For the reading section discrimination coefficients, the one way ANOVAs conducted separately on the z-transformed R_{pbis} and the DI coefficients established that there were no significant differences between the mean values of each of the sets of coefficients (z-transformed R_{pbis}: $F = 0.241$, $p > 0.05$; DI: $F = 0.545$, $p > 0.05$; $df = 2, 21$).

For the structure sections, the non-parametric Friedman tests conducted on the two sets of coefficients produced no evidence of significant differences between the DI's ($\chi^2 = 4.439$, $df = 2$) but for the R_{pbis} coefficients there was evidence of differences, which three Wilcoxon signed ranks tests suggested related to comparisons between the 5-option group values and both the 3- and 4-option groups (5-/3-option: $z = -2.768$, $p < 0.01$; 5-/4-option: $z = -2.295$, $p < 0.05$), indicating that generally, the 5-option test had produced higher individual R_{pbis} coefficients.

Option and distractor performance

A number of different methods of assessing option and distractor performance were adopted in this study based on the methods recommended in Haladyna and Downing (1993). However, the analysis of the performance of the options commenced with establishing how well the sets of options had distributed the test takers' selections, using the AENO coefficient.

AENO

The AENO (actual effective number of options; Sato and Morimoto, cited in Shizuka et al., 2006) is a measure of the distribution of option selections so is not necessarily indicative of individual option performance. AENO coefficients range from a value of 1 up to the number of options in the item (*k)*. A value of 1 would indicate that all selections were of the same option; a coefficient equal to *k* (3 for a 3-option item, 4 for a 4-option item etc.) would indicate an equal distribution of selections between the available options. Shizuka et al. suggest that ideally the AENO value should fall between *k*-0.5 and *k*.

The AENO values for the 31 structure and 8 reading items are shown in Table 7. The 3-option items produced the highest number of AENO values between *k*-0.5 and *k*, 15 among the structure items and five among the reading items, with the 4- and 5-option items producing only five and three respectively for the structure items and one each for the reading items. Because the scale of values derived from each option format test differed, a comparison of mean values was not appropriate. Therefore, in order to compare the performance of the different option formats in spreading the

Table 7 AENO values of items in the three multiple choice tests

Structure items							
Item	**3-option**	**4-option**	**5-option**	**Item**	**3-option**	**4-option**	**5-option**
1	2.472	2.411	2.544	29	2.417	3.553	4.674
2	2.585	2.790	3.989	30	2.896	3.759	4.417
4	1.896	2.467	2.640	31	2.472	2.891	3.726
5	2.567	3.431	3.158	32	2.561	3.271	4.076
7	2.923	3.468	4.238	34	2.806	3.455	4.382
8	1.709	1.597	1.958	35	2.788	3.147	3.424
10	2.759	2.915	3.459	37	2.945	3.377	3.604
11	2.916	3.870	4.106	38	2.360	3.492	3.598
13	2.812	3.823	3.198	39	2.317	2.731	2.996
14	2.555	2.578	3.563	40	2.575	2.648	4.271
15	1.857	2.633	2.672				
16	1.954	2.582	2.838	***Reading items***			
17	1.904	2.416	2.516	Item	3-option	4-option	5-option
18	2.479	2.603	2.077	41	2.608	3.045	3.515
19	2.264	2.922	3.500	43	2.855	2.814	3.479
21	2.073	2.845	2.940	45	2.719	3.441	4.183
22	2.690	3.806	4.763	46	2.999	3.764	4.712
23	2.059	2.923	4.554	48	1.383	1.095	1.713
25	2.048	3.307	3.218	49	2.944	3.341	4.388
26	2.689	2.931	3.697	50	2.156	2.496	2.667
28	2.392	3.396	3.429	51	2.418	3.317	3.311
Structure section				***Reading section***			
Mean	2.44	3.03	3.49	Mean	2.51	2.91	3.50
S. Dev.	0.36	0.52	0.75	S. Dev.	0.54	0.83	0.97

test taker's option selections, the AENO values were converted to a common scale based on the 4-option values by deducting the mean value of each of the 3- and 5-option sets from each value in those sets then adding back the mean value of the 4-option set. The resulting standardized values were then compared by deriving correlation coefficients. The structure items produced greater variation as judged by those correlation coefficients between the sets of values shown below in Table 8, although all the correlation coefficients were moderately high and significant at $p < 0.001$. The reading coefficients were more highly correlated and all were significant at or above the 0.01 level.

As would be expected, the mean AENO values for the structure and reading items increased with the number of options (structure: 3-option: 2.44; 4-option: 3.03; 5-option: 3.49; reading: 3-option: 2.51; 4-option: 2.91; 5-option: 3.50). Generally the mean AENO values increased by approximately 0.5 for each added option, suggesting that adding discriminators tends to spread the selection across the available options, but that this tendency is lessened with increasing numbers of options. This is as would be expected given that the distractors in this study were selected based on having been offered as answers by the participants in the short answer test, with the third and fourth

Table 8 Correlation coefficients from the comparison of AENO values (based on values standardised according to the 4-option scale)

	3-option test	4-option test	
Structure items			
4-option test	0.675***		
5-option test	0.580***	0.679***	***significant at $p < 0.001$; $df = 29$
Reading items			
4-option test	0.914**		**significant at $p < 0.01$
5-option test	0.943***	0.932***	***significant at $p < 0.001$; $df = 6$

distractors having in most cases been offered by relatively few participants.

Performing and discriminating distractors per item

Each of the distractors was analysed in each of the items in which it appeared and was classified as performing, if it produced an increasing number of selections from within three equivalent-sized ability groups (high, middle and low) drawn from within each experimental group (i.e. the high groups made the least selections and the low group produced the highest number with the middle group in between), or discriminating if it produced a significant negative R_{pbis} correlation with the overall performance of the group in the section (structure or reading) concerned (Haladyna and Downing, 1993). Table 9 shows the mean numbers of performing and discriminating options per item.

As can be seen, the mean values were very close and one-way ANOVAs confirmed that there were no significant differences in the numbers of performing or discriminating distractors in either the structure or reading sections (structure: performing: $F = 0.529$, $p > 0.05$; discriminating: $F = 1.679$, $p > 0.05$; $df = 2, 90$; reading: performing: $F = 0.717$, $p > 0.05$; discriminating: $F = 0.797$, $p > 0.05$; $df = 2, 21$). Table 9 also includes details of how many items produced 3, 2, 1 and 0 performing and discriminating distractors. Only one item produced three performing or discriminating distractors, with no items producing four. The 3-option test produced more items with two performing or discriminating options than both the 4- and 5-option tests in all sections and generally the fewest items with no performing or discriminating options (although not in the structure section in respect of performing options, where the 5-option test produced the fewest), tending to confirm that option selection spreads across the available options thus being likely to increase the number of non-performing/discriminating distractors in items with higher numbers of options.

Performing sets of distractors

The performance of the whole set of distractors in each of the items was analysed based on the same equivalent-sized ability groups (high, middle and low) used in ii above, to judge if the whole set of distractors in an item had performed correctly, (i.e. that the high group had produced the fewest selections of any distractor, and the low group had produced the highest number with the middle group in between). The total numbers of performing sets for the structure and reading items are shown below in Table 10. As can be seen the figures were very close for the structure items and a one-way ANOVA detected no significant differences ($F = 0.614$ $p > 0.05$; $df = 2, 90$). For the reading section the numbers of performing sets were identical. Notably however the distributions of performing sets between the items

Table 9 Summary of performing and discriminating distractors

	Structure section (31 items)			*Reading section (8 items)*		
	3-option test	**4-option test**	**5-option test**	**3-option test**	**4-option test**	**5-option test**
Mean performing distractors per item	0.87	0.77	0.94	1.25	0.75	0.88
Items with:						
3 performing distractors	n/a	0	0	n/a	1	0
2 performing distractors	5	3	4	3	0	2
1 performing distractor	17	18	21	4	3	3
Non performing distractors	9	10	6	1	4	3
Mean discriminating distractors per item	1.23	1.07	0.94	1.13	0.75	0.75
Items with:						
3 discriminating distractors	n/a	0	0	n/a	0	0
2 discriminating distractors	9	8	5	2	1	1
1 discriminating distractor	20	17	19	5	4	4
Non discriminating distractors	2	6	7	1	3	3

n/a: not applicable.

was not consistent as reflected by the correlation coefficients between the data sets shown in Table 10, none of which were significant and for the reading section, were in two out of three cases, negative suggesting that it was not the same items which were producing the performing sets in the different option format groups.

Individual performing distractors

The distractors were compared based on their content (as opposed to their letter denomination in the list of options, which was allocated randomly between the items in different option format), their inclusion in the items, and the number of times that an individual distractor was judged to be performing or discriminating according to the definitions employed in section ii above. This produced the numbers of performing and discriminating options shown in Table 11 below.

A clear pattern emerges, with more positive findings overall for the first and second distractors present in all three option format tests with the 3-option items producing more positive findings than the 4- and 5-option tests among the common distractors. For the structure items the third distractors were more successful than the fourth distractor present only in the 5-option items, although for the reading items, the fourth distractors were rather more successful than the third.

Table 10 Performing sets of distractors

Number and mean per item

	Structure items (31)			*Reading items (8)*		
	3-option test	**4-option test**	**5-option test**	**3-option test**	**4-option test**	**5-option test**
Number of performing sets	24	20	22	6	6	6
Mean performing sets per item	0.77	0.65	0.71	0.75	0.75	0.75

Correlation coefficients between distribution of performing distractor sets

	3/4-option	3/5-option	4/5-option
Structure section	0.244	0.164	0.268
Reading section	0.333	−0.333	−0.333

Table 11 Numbers of individual performing or discriminating options

Structure experimental	3-option test		4-option test		5-option test	
	Performing	Neg. R_{pbis}	Performing	Neg. R_{pbis}	Performing	Neg. R_{pbis}
1st and 2nd options in all 3 tests (62)	27	38	18	23	20	20
3rd option in 4/5-option test (31)	n/a	n/a	6	10	6	7
4th options in 5-option test (31)	n/a	n/a	n/a	n/a	3	2
All distractors (124)	27	38	24	33	29	29
Reading experimental						
1st and 2nd Options in all 3 tests (16)	10	9	4	5	3	2
3rd option in 4/5-option test (8)	n/a	n/a	2	1	1	1
4th options in 5-option test (8)	n/a	n/a	n/a	n/a	3	3
All distractors (32)	10	9	6	6	7	6
Control items combined (36): Total	14	14	6	11	7	14

n/a: not applicable.

The effect of option numbers on response changes between the short answer and multiple choice tests

The overall study looked in some detail at the changes in the participants' responses between the two stem-equivalent tests which they took, in short answer and multiple choice format, and although this paper will not present the full findings from this aspect of the study (for which see Currie and Chiramanee 2010) the findings insofar as they shed light on the performance of the items in their three multiple choice option formats are detailed below.

In order to study the changes in responses between tests, the participants' answers from the short answer test were directly compared with their option selection in the multiple choice test, and the patterns of variation and consistency were classified according to the taxonomy shown in Table 12 below, and analysed based on the test option formats, and the test sections (structure and reading). The results of the analysis appear in Table 13.

One-way ANOVAs conducted on the individual response classifications across the three option number format groups, found significant differences for the structure items in only two patterns; firstly in pattern C ($F = 13.078$, $p < 0.001$, $df = 2, 149$) which covered instances where the participant changed from an incorrect answer given in the first test to a different incorrect option in the multiple choice test even though their original answer was offered as a distractor in that test. A Tukey HSD post hoc analysis established that the difference was between the 3-option group and both the 4- and 5-option groups (no significant difference was found between the 4- and 5-option groups), with the 3-option group making such a change about half as often as did either the 4- or 5-option groups. Secondly for pattern J, where the participant had been able to select the correct option in the multiple choice test when her/his incorrect answer from the first test was not offered as a distractor, the 3-option group were found to have been significantly more successful than either the 4- or 5-option groups ($F = 5.757$, $p < 0.01$, $df = 2, 149$).

For the reading items, pattern C again produced a significant difference ($F = 6.554$, $p < 0.01$, $df = 2, 149$) and the reading items also produced a significant difference in the occurrence of pattern F ($F = 4.905$, $p < 0.01$, $df = 2, 149$) where participants had abandoned a correct answer offered in the short answer test in favor of an incorrect option in the multiple choice test. For both patterns, a post hoc analysis pointed to a

Table 12 Taxonomy of response changes/consistency between the short answer and multiple choice tests (first published in Currie and Chiramanee 2010)

Code	1st test response	Available or not available option in 2nd test	Response in 2nd test
A	No answer		Incorrect option
B	No answer		Correct option
C	Incorrect	Available	Incorrect option – different option from answer in 1st test
D	Incorrect	Available	Incorrect option same answer as given in 1st test
E	Incorrect	Available	Correct option
F	Correct*	Available	Incorrect option
G	Correct*	Available	Correct option
H	Incorrect	Not available	Incorrect option
J	Incorrect	Not available	Correct option
K	Acceptable**	Not available	Correct option
L	Acceptable**	Not available	Incorrect option
M	No answer		No answer
N	Incorrect	Available	No answer
O	Incorrect	Not available	No answer
P	Correct*	Available	No answer
Q	Acceptable**	Not available	No answer

Notes: *Expected response **Non-expected but correct response.

Table 13 Response patterns between the short answer and multiple choice tests

Pattern	Structure items			Reading items		
	3-option group	4-option group	5-option group	3-option group	4-option group	5-option group
A	2.85%	2.58%	2.94%	4.57%	2.27%	2.91%
B	1.99%	1.88%	1.36%	2.16%	2.27%	2.03%
C	5.89%[a]	11.61%[a]	13.05%[a]	4.09%[c]	8.41%[c]	14.83%[c]
D	13.40%	11.67%	12.69%	12.50%	14.32%	9.88%
E	11.66%	11.03%	10.97%	6.97%	7.73%	7.56%
F	2.23%	2.76%	2.72%	4.81%[d]	7.05%[d]	10.17%[d]
G	14.45%	14.43%	14.19%	30.53%	27.50%	25.58%
H	22.52%	24.16%	22.65%	21.39%	19.32%	19.19%
J	22.21%[b]	17.07%[b]	17.06%[b]	12.26%	10.68%	7.27%
K	1.67%	1.99%	1.22%	0.24%	0.00%	0.00%
L	1.12%	0.82%	1.15%	0.24%	0.23%	0.00%
N	0.00%	0.00%	0.00%	0.00%	0.23%	0.00%
O	0.00%	0.00%	0.00%	0.24%	0.00%	0.29%
P	0.00%	0.00%	0.00%	0.00%	0.00%	0.29%

Letter superscripts indicate patterns producing significant differences:
[a] $F = 13.078$, $p < 0.001$, $df = 2, 149$.
[b] $F = 5.757$, $p < 0.01$, $df = 2, 149$.
[c] $F = 6.554$, $p < 0.01$, $df = 2, 149$.
[d] $F = 4.905$, $p < 0.01$, $df = 2, 149$.

significant difference between the 3- and 5-option groups although it is notable that there was a clear rising trend in both patterns of behavior across the 3-, 4- and 5-option groups, with the 5-option group three times as likely as the 3-option group to change between incorrect options when their original incorrect option was among the distractors and more than twice as likely to abandon a correct answer in the multiple choice test.

Discussion

Generally the comparisons of concurrent validity, reliability, facility, discrimination ability and distractor performance produced only minor indications of differences in the performance of the multiple choice tests in the three different option number formats.

The comparison of concurrent validities based on the first year participants' O-NET scores found no significant differences between the correlations with the experimental test scores derived from the multiple choice tests although there was a higher correlation for the 3-option test than the 4-option test, despite the O-NET examination itself being composed of 4-option multiple choice items, with the 5-option test producing the lowest correlation. There were no significant differences in the Cronbach alpha reliability coefficients of the three tests and those for the structure sections and also overall were very closely aligned, with the coefficients for the reading sections somewhat lower as well as being more widely spread due to the lower number of items on which the calculation of the coefficients was based.

The analysis of discrimination found no significant difference in discrimination ability overall in either the structure or reading sections although individually the R_{pbis} coefficients for the 5-option structure sections were found to be significantly higher than those in both the 3- and 4-option tests. The analysis of option and distractor performance also found little indication of any significant effect as between the three different formats although the AENO analysis as well as the analysis of performing and discriminating distractors highlighted a tendency for the greater number of options in the 4- and 5-option items to spread the distribution of option selections more widely. However, given the findings of no significant differences either in the numbers of performing or discriminating distractors per item or in the performance of the sets of distractors over the three different option format tests, the change in the distribution of option selection seems to have had little or no effect on the overall item performance, suggesting that the number of options in a multiple choice item is not a major factor in influencing option selection.

Comparisons of the mean item facility in both the structure and reading sections produced no significant differences although a general tendency for the difficulty of both sections to increase with the change from three to four to five options was observable. There was some support for this tendency from the comparison of individual item facility values in the reading section where the 5-option items were noted to have been consistently more difficult than the 3-option items, although this was not found to be statistically significant.

The findings from the comparison of responses between the short answer and multiple choice tests provided significant clues as to why the 3-option items produced somewhat lower facility indices in this study, with the 3-option group being generally more successful in selecting the correct option in the multiple choice test when they were unable to offer a correct response in the short answer test (behavior suggestive of

guessing, or cued recall) and being less likely to be misled by the smaller number of distractors in the 3-option format items Currie and Chiramanee (2010). Overall however it is notable that the analysis of change and consistency between the two tests produced three highly consistent data sets across the three different option number format groups, suggesting once again that the number of options in a multiple choice item is a matter of little significance in the measurement of language based constructs.

It is also worth commenting that the study found much greater variation in response behavior by comparing the distribution of patterns between the structure and reading sections where *t* tests found significant differences in the distribution of six patterns covering around 50% of the overall patterns recorded (patterns E, F, G, J, K & L – see Table 14), and that a comparison of patterns based upon high middle and low ability groups conducted in the main study, found differences in 10 of the 14 patterns detected, accounting for more than 98% of the patterns noted Currie (2008).

Finally, although the three different option number format tests produced few significant differences in performance throughout this study, the comparison of the participants' responses in the two tests they took, produced an indication of major differences in how test takers answer language tests in constructed response and multiple choice formats, with only 28% of the answers from the short answer test structure section being maintained in the multiple choice test (see Table 13 – groups D & G) and 43% being maintained in the reading section, pointing to a fundamental difference in the constructs being measured by the two formats.

Conclusions

Overall therefore the findings of this study suggest that despite the considerable theoretical and research interest which has been devoted to the issue of the optimal number of options in multiple choice items, for the measurement of language

Table 14 Overall response patterns by section and comparison of means

	Structure	Reading	*t* stat
A	2.58%	3.67%	−1.549
B	1.60%	1.92%	−0.661
C	9.80%	9.54%	0.329
D	12.62%	12.83%	−0.244
E	12.31%	7.02%	6.914***
F	2.28%	5.98%	−6.227***
G	13.46%	26.75%	−13.093***
H	22.99%	21.22%	1.605
J	18.76%	9.38%	9.927***
K	2.13%	0.66%	5.187***
L	1.47%	0.82%	2.542*
N	0.00%	0.05%	−1.000
O	0.00%	0.11%	−1.419
P	0.00%	0.05%	−1.000

t test: two-tailed, paired sample.
*significant at $p < 0.05$
***significant at $p < 0.001$;
$df = 151$ (*t* test based on comparison of individual subjects' response patterns. Figures for reading adjusted by 39/12 to allow for number of items in each section).

structure and reading ability, the issue may be of minor importance in considering the performance of the multiple choice format.

There was some evidence that higher numbers of distractors may have induced some test takers to select incorrect options even though in the short answer test they were able to offer a correct response, a tendency which was more marked in the reading items. Additionally, in the structure section particularly, reducing the number of options to three seems to have had the entirely predictable effect of making it easier for the test taker to select the expected response option when he/she either had no idea of the correct answer (i.e. guessing) or had only a vague or partial knowledge, which was assisted by the presence of the correct answer among the options (i.e. cued recall). The only other effect of decreasing the number of options to three was a marked tendency to reduce the number of changes between incorrect answers given in the first test and incorrect options selected in the second test. Although this tendency had no effect overall on the number of correct answers achieved by the participants, it does predictably suggest that when the test taker is ignorant of the correct answer, higher numbers of options are likely to make it more difficult for them to arrive at the correct response by guessing.

In contrast to the limited variations in performance arising from the different numbers of options per item, the study found more substantial differences between the effect of the use of multiple choice items as between the structure and reading sections suggesting that the skill area being tested may be a much more significant factor in considering the effectiveness of multiple choice items, and a larger effect still associated with the ability of the tests takers Currie (2008). But by far the greatest differences were found in comparing the test takers' response as between the short answer and the three multiple choice tests, differences so great as to suggest that the constructs being measured by the two types of test are substantially different and that the results of the tests in the multiple choice format may have owed at least 60% to factors associated with the test takers' ability to deal with the test format such as guessing, cued recall and the use of test taking strategies, which are entirely irrelevant to the language based construct which the test was set to measure.

The implication of these findings is that the use of the multiple choice format in language tests, often justified on the grounds of its reliability and practicality, may not actually represent a valid measurement of language ability at all, and that measurement accuracy is being sacrificed in the cause of economy and efficiency. In fact arguably the results of multiple choice tests are likely to represent distorted measurements owing more to the test takers' ability to manipulate the testing medium than to their language knowledge and ability. Clearly this is an implication with far-reaching consequences in view of the noted tendency for test format to influence the content of teaching, and may go a long way towards explaining the noted lack of communicative English language skills acquired by Thai students during their formal education.

Limitations and recommendations for further studies

This study was conducted using subjects at or above the level of English ability generally found in Thai students at the time of university entrance and was based

on a test of English structure and reading ability. Its findings cannot be generalized outside of the testing of language and its applicability to other levels of English learning and ability and to learners in other educational contexts would have to be confirmed by further studies.

The main recommendation of this study is therefore, that rather than searching for apparently illusory psychometric benefits from varying option numbers in multiple choice items in order to justify the practical and economic advantages to be gained by a reduction to 3-options, future research effort should be concentrated on studying whether the differences noted between the outcomes of tests using constructed response formats (such as the short answer format used in this study) and selection based formats of the multiple choice type, are a general feature of language tests, and also whether the effects occur in educational domains outside of language.

Additional file

Additional file 1: Appendix A. Sample of item types used in the experimental tests.

Competing interest
The authors declare that they have no competing interests.

Authors' contributions
Both authors read and approved the final manuscript.

Author details
[1]Faculty of Liberal Arts, Prince of Songkla University, Hatyai 90110, Thailand. [2]AUA Language Center, Platha Road, Muang, Songkhla 90000, Thailand.

Received: 1 October 2014 Accepted: 24 October 2014
Published online: 07 December 2014

References
Baines, S. (2000). Selling ivory. *New English Digest, 3*(2), 36–37.
Bruno, JE, & Dirkswager, A. (1995). Determining the optimal number of alternatives to a multiple choice test item: an information theoretic perspective. *Educ Psychol Meas, 55*, 959–966.
Budescu, DV, & Nevo, B. (1985). Optimal number of options: an investigation of the assumption of proportionality. *J Educ Meas, 22*, 183–196.
Currie, M. (2008) Measuring language learning or distorting the construct? A study of multiple choice items in a test of English structure and reading. Unpublished MA thesis, Prince of Songkla University, Hatyai Campus, Thailand.
Currie, M., & Chiramanee, T. (2010) The effect of the multiple choice item format on the measurement of knowledge of language structure. *Language Testing*, 27, 471–492.
Ebel, RL. (1969). Expected reliability as a function of choices per item. *Educ Psychol Meas, 29*, 565–570.
Green, K, Sax, G, & Michael, WB. (1982). Validity and reliability of tests having different numbers of options for students of differing levels of ability. *Educ Psychol Meas, 42*, 239–245.
Grier, JB. (1975). The number of alternatives for optimum test reliability. *J Educ Meas, 12*(2), 109–113.
Grier, JB. (1976). The optimal number of alternatives at a choice point with travel time considered. *J Math Psychol, 14*, 91–97.
Haladyna, TM, & Downing, SM. (1993). How many options is enough for a multiple choice test item? *Educ Psychol Meas, 53*, 999–1010.
Lee, HS, & Winke, P. (2012). The differences among three-, four, and five-option-item formats in the context of a high stakes English-language listening test. *Lang Test, 1*(30), 99–123.
Lord, F. (1977). Optimal number of choices per item: a comparison of four approaches. *J Educ Meas, 14*, 33–38.
Musigrungsi, S. (2002). *An Investigation of English Grammar Teaching in Government Secondary Schools in Educational Region II* (Unpublished MA Thesis). Hatyai campus, Thailand: Prince of Songkla University.
Prapaisit, L. (2003). *Changes in Education After the Educational Reform in Thailand* (Unpublished Doctoral Thesis). Michigan, USA: Michigan State University.
Prapphal, K. (2008). Issues and trends in language testing and assessment in Thailand. *Lang Test, 25*(1), 127–143.
Ramos, RA, & Stern, J. (1973). Item behaviour associated with changes in the number of alternatives in multiple choice items. *J Educ Meas, 10*, 305–310.
Rodriguez, MC (2005). *Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. Educational Measurement: Issues and Practice*, Summer, 3–13.
Shizuka, T, Takeuchi, O, Yashima, T, & Yoshizawa, Y. (2006). A comparison of 3 and 4 option English tests for university entrance selection purposes in Japan. *Lang Test, 23*(1), 35–57.

Sidick, JT, Barrett, GV, & Doverspike, D. (1994). Three alternative multiple choice tests: an attractive option. *Pers Psychol, 47*, 829–835.

Thongsri, M. (2005). *An Investigation into the Implementation of 2001 English Language Curriculum in Government Secondary Schools in Songkhla* (Unpublished M.A. Thesis). Hatyai Campus, Thailand: Prince of Songkla University.

Tversky, A. (1964). On the optimal number of alternatives at a choice point. *J Math Psychol, 1*, 386–389.

Watson-Todd, R. (2008). *The Impact of Evaluation on Thai ELT. Proceedings of the 12th English in South East Asia Conference, Bangkok*. Bangkok: KMUTT. December 2007.

doi:10.1186/s40468-014-0008-7
Cite this article as: Inadaphat and Currie: **The number of options in multiple choice items in language tests: does it make any difference? Evidence from Thailand.** *Language Testing in Asia* 2014 **4**:8.