Options in a Multiple-Choice English Test: Quality over Quantity[*]

Sung-Yeon Kim
(Hanyang University)

Kim, Sung-Yeon. 2012. **Options in a multiple-choice English test: Quality over quantity.** *Korean Journal of English Language and Linguistics* 12-1, 19-39. Multiple choice tests are useful for testing a large group of students since scoring is objective and fast. For this reason, they are widely used in various standardized tests. However, it takes time to develop good multiple choice items since it is quite challenging to generate distracters that are attractive and have high discrimination power. The question is how many distracters are optimal in realizing the goal of multiple choice tests. Views on this matter in the literature are mixed. The present study compared the effects of three option formats (3, 4, and 5 options) in terms of item discrimination power. Contrary to the popular beliefs, the study found that the discriminatory power decreased as the number of options increased. In other words, the items with 5 options were found to have lowest item discrimination. Considering that the Korean SAT adopts the 5-option format, the implication of the findings is of great significance to language testing practitioners and researchers alike. It can be inferred from the finding that the quality of each option is much more important than the number of options in designing test items.

Key Words: multiple-choice test, test construction, options, alternatives, distracters, nonfunctioning options

1. Introduction

It is widely held that multiple choice tests are useful and convenient to use. They are also economical and efficient in

terms of test administration and scoring. The efficiency is visible particularly when we measure discrete knowledge of language forms such as grammar and vocabulary. Because of these advantages, multiple choice tests have proliferated in various standardized tests although they "rarely measure communication as such" (Heaton, 1988, p. 27).

Unlike other types of tests such as essays, however, it is both time-consuming and labor-demanding to construct good multiple choice items. This difficulty in item construction has led testing specialists to offer item writing guidelines (Woodford & Bancroft, 2004). The guidelines, while beneficial and useful, remain largely anecdotal because they are not based on empirical research (Rodriguez, 2005).

As one of the factors affecting the validity of multiple choice questions, the number of options has been the focus of debate among advocates of multiple choice questions. For instance, Haladyna and Downing (1989a) recommended as many alternatives as feasible; they then slightly revised their argument in a follow-up study (Haladyna & Downing, 1989b), arguing for "as many functional distractors as are feasible" (p. 59). On the other hand, Cizek and O’Day (1994) supported four options.

This issue of options in a multiple choice test seems particularly important in the Korean context since it bears direct relevance to the Korean scholastic aptitude test (SAT). The Korean SAT as a high-stakes exam adopts five options; consequently, other standardized tests in Korea including primary school exams also adopt the five-option format. However, as Lee and Kwon (2002) cautioned, too many options might make it hard to realize the original purpose of the test. For example, five options in a test of listening comprehension would incur unnecessary cognitive burden on test takers as the examinees are likely to forget the aural input before they have time to review all five options; then the test becomes a test of memory than

[*] 이 논문은 2009년 한양대학교 일반연구비 지원으로 연구되었음 (HY-2009-00000001023)

that of listening comprehension. In addition, all the options might not be functioning and attractive as distracters (Cizek & O'Day, 1994).

Research literature, however, has not reached a consensus on the optimal number of options in multiple-choice tests (Haladyna & Downing, 1989a). Some studies supported a 3-option format (Bruno & Dirkzwager, 1995; Haladyna & Downing, 1993; Owen & Froman, 1987; Sidick, Barrett, & Doverspike, 1994) while others such as Cizek and O'Day (1994) argued for a 4-option format based on their finding that 4 options were as reliable as 5 options.

Recently, there are a few noteworthy studies that have examined the number of options in multiple choice tests. Particularly interesting is a study by Shizuka, Takeuchi, Yashima, and Yoshizawa (2006) and Lee and Kwon (2002). Shizuka and his colleagues wanted to see the effects of eliminating nonfunctional options in multiple choice tests. They administered a 4-option multiple choice reading test to a group of students, removed the least frequently endorsed option in each item, and administered the revised test to a different group of test-takers. Their comparison between the two versions did not result in significant differences in item facility (or item difficulty, the proportion of test-takers who answered correctly) and item discrimination (how well an item can discriminate between those who are good and those who are not).

In contrast, Lee and Kwon's (2002) study compared three groups of test items that were different in the number of options. Their findings showed that five options yielded the highest item discrimination power. Kim (2009) used the same test as Lee and Kwon (2002) and found that the item discrimination power did not increase as a function of the number of options. The items with odd number options (5 options and 3 options) had higher discriminatory power than 4-option items. Interestingly, the items with five options were found to have the highest discrimination power although they had almost the same discriminatory power as 3-option items, which was difficult to account for.

In light of the above findings, the question of ideal number of items still remains unsolved and inconclusive. One way to clarify this matter is to replicate the earlier studies (Kim, 2009; Lee & Kwon, 2002) by examining a different group of examinees. If the effects of the number of options in multiple-choice tests hold true regardless of the variation of participant factors (for instance, with different majors and proficiency levels), conclusions drawn from the study will be valid. The present study, distinct from prior studies in terms of participant characteristics, aims to compare the item discrimination power across the three different option types.

2. Literature Review

Multiple choice items are undoubtedly one of the most widely used test types in standardized testing. They are used to measure a variety of linguistic, pragmatic, and topical content in language teaching. They are particularly useful for testing vocabulary, grammar, listening and reading comprehension (Lee & Kwon, 2002). Some researchers in intercultural pragmatics also used multiple choice tests to measure pragmatic competence (Kasper & Rose, 2002). The widespread use of multiple-choice items is ascribed to the fact that test administration and scoring is efficient and effective. Particularly, scoring is reliable in that it is not influenced by the subjective judgment of the rater (Weir, 1988) unlike integrative tests such as essays or interviews.

Despite these numerous advantages, multiple choice items are not so much used by teachers for classroom quizzes and tests

(Cohen, 1994). This is because it is quite challenging to construct good multiple-choice items. As Ebel (1951) puts it, item writing is not a simple task, but an "art that requires an uncommon combination of special abilities" (p. 185). Particularly challenging is to construct good distracters. Plausible distracters are critical in multiple-choice tests because they determine the quality of the tests by attracting students who do not know the right answer (Cohen, 1994).

An important element for making good multiple choice tests is the number of options. Henning (1987) and Heaton (1988) argued that the optimum number of options or alternatives should be five in most public tests, suggesting that a large number of options would lower the element of chance. By contrast, some researchers have warned against having nonfunctional or dysfunctional options (Cizek & O'Day, 1994; Haladyna & Downing, 1989b). Despite the difference in their view of the number of options, they all acknowledged that more difficulties should be expected as the number of options increases (Cizek & O'Day, 1994 Henning, 1987).

As test writers are in pursuit of developing more distracters, they are likely to construct nonfunctioning distracters, such as nonsense distracters and review distracters (Henning, 1987). Nonsense distracters are options that do not match the stem of the item and are not grammatically acceptable, whereas review distracters oblige test-takers to review the previous choices and infer their interrelatedness.

Those nonsense options and review options, however, should be avoided when constructing many alternatives. Instead, functioning options should be designed even though it takes more time and effort on the part of item writers. For instance, in constructing 5-option items, it is not easy to develop four distracters that are all highly attractive (Delgado & Prieto, 1998). For this reason, as stated by Haladyna (1994), "item writers are

often frustrated in finding a useful fourth and fifth option because they do not exist" (p. 75).

Haladyna and Downing (1989b) advocated the deletion of nonfunctional distracters, arguing that "two-, three-, and four-option test items would perform at least as well as five-option items having one or more dysfunctional distracters" (p. 58). Likewise, Alderson, Clapham, and Wall (1995) recommended that 3 options should be used when it is impossible to think of a third attractive distracter.

Some empirical research confirms this line of argument. For instance, Rodriguez (2005) from a meta-analysis of 27 studies on the number of options concluded that three options were optimal for multiple choice items in most settings. In a more recent study, Berríos, Rojas, Cartaya, and Casart (2005) examined how the quality of English reading comprehension tests changed as a result of reducing the number of options from 4 to 3. Their findings indicated that the mean item difficulty, the mean item discrimination, and the reliability coefficients of the four- and three-option tests were not so much different and that it took less time to administer the 3-option format.

There is a follow-up study that used the same research design as Berríos et al. (2005). Shizuka et al. (2006) examined the effect of the number of options in the Japanese context. In order to investigate the psychometric properties of the reading test, this study removed the least frequently endorsed distracter among 4 options in each item and administered the test to another group of examinees. Their results show that regardless of the number of options, there was not a significant change in item facility. In addition, the new test did not affect the item discrimination and test reliability. In fact, the mean number of functional distracters turned out to be less than 2. In other words, eliminating an ineffective option was found to have little effect on the functions of the remaining options.

Previous studies taken together support the use of three options over four or five options (Berríos et al., 2005; Bruno & Dirkzwager, 1995; Haladyna & Downing, 1993; Owen & Froman, 1987; Rodriguez, 2005; Shizuka et al., 2006; Sidick, Barrett, & Doverspike, 1994; Tversky, 1964). If so, the use of 3 options (or 2 distracters) may be a better choice because it saves printing cost as well as time for item writing and for test administration. This led Haladyna and Downing (1993) to argue for "using fewer options, with three being desirable for most measurement purposes." (p. 11).

Fewer options can contribute to increasing the reliability of a test since the test can contain more items by reducing the number of options (Berríos et al., 2005). In fact, there is a study that has examined this effect quantitatively. Aamodt and McShanes (1992) found that on average 112.4 three-option items can be tested in the same amount of time as 100 four-option items can. Thus, they stated, "other things being equal, giving more items in the same amount of time should result in higher test score reliability" (p. 52).

With respect to the efficiency of the 3-option format, Shizuka et al. (2006) offered the following summary:

- the increased test reliability and item writing efficiency;
- the length of a test booklet is smaller;
- printing costs are reduced;
- the distractors taken as a set should be more plausible;
- students can answer questions with less distraction;
- students will feel less pressure because they can work more slowly or spend time to recheck; and
- the chances of providing unintended cues that profit test-wise students will be decreased. (p. 53)

Despite the advantages of the three-option format, the Korean SAT adopts 5 options rather than 3 or 4 options. The choice of 5 options over 3 or 4 options may be partly due to the generally held belief that more alternatives would reduce the likelihood of getting the correct answer by chance alone (Cohen, 1994; Heaton, 1988). Considering the pedagogical impact of the Korean SAT, it is no wonder that five options are now widely adopted as the norm in many language tests in Korea, such as other school-based exams.

As reviewed above, however, simply increasing the number of options does not always insure a highly reliable and valid test. Prior research has demonstrated how difficult it is to develop as many as five good alternatives because they might contain dysfunctional or nonfunctioning distracters. The following illustrates the case in point. The item is taken from the 1999 Korean SAT cited in Lee (2007).

Item No. 34 from Korean SAT administered in 1999 (Lee, 2007, p. 436)

Stem: Which of the following underlined parts is NOT natural?

It is often believed that the function of school is ① to produce knowledgeable people. If schools ② only provide knowledge, however, they may destroy creativity, ③ producing ordinary people. We often ④ hear stories of ordinary people who, if education had focused on creativity, could have become great artists or scientists. Those victims of education ⑤ should receive training to develop creative talents while in school. It really is a pity that they did not.

Table 1 shows that a small percentage (ranging from 4.48% to 5.48%) of test-takers endorsed the first two options.

Table 1 Item Analysis: Item No. 34 (Lee, 2007, p. 436)

Item	IF	DI	% of test-takers who endorsed the options				
			①	②	③	④	⑤√
34	27.19	0.15	4.48	5.48	46.02	16.73	27.19

*√= correct answer

This means that those two options were not so attractive enough to distract test-takers. Considering that distracters should appear plausible to examinees who do not know the correct answer, they did not function well as distracters.

The presence of many dysfunctional or nonfunctioning distracters found in the Korean SAT therefore makes it difficult to adopt 5-options. Prior research has demonstrated the importance of the number and the quality of options as two important factors for test-takers' performance. Considering possible variation of test takers' performance according to distracters, we need to specify the effects of the number of options with various groups of examinees. The present study used item discrimination (ID) power and item facility (IF) to examine the function of the number of options.

3. Method

3.1. Research Questions

This study was conducted to examine if the 5-option format recommended by earlier studies (Kim, 2009; Lee & Kwon, 2002) would be generalizable to other groups of examinees. While the present study focuses on comparing the effectiveness of three options with that of four or five options, it differs from Lee and Kwon (2002) and Kim (2009) on two accounts.

First, this study presents examples of items with high item discrimination to highlight further the effects of each option. Second, the participants in the study are vastly different from those in prior studies; the study aims to see if the preference for more options can be maintained across different groups of examinees. This result would help us determine if the 5-option format frequently used in many language tests is valid (Kim, 2009; Lee & Kwon, 2002) or if the 3-option format is a better choice. The research questions posed for the study include:

1) How does the number of options affect the item facility index in a multiple-choice English test?
2) How does the number of options affect the item discrimination power in a multiple-choice English test?
3) What are the characteristics of options with high discrimination?

3.2. Participants

For data collection, 337 college freshmen were asked to take a test specially designed for the purpose of the study. They were enrolled in English conversation courses at a university in Seoul. They had not received college-level education since the test was conducted a week after their entrance into college. The participants are quite different from those in prior research (Kim, 2009; Lee & Kwon, 2002). Whereas previous studies examined would-be teachers that were mostly female, the present study examined students in the engineering field of study, who were mostly male. The participants were in the top 3% of all the newly admitted students, as measured by high school grades and the Korean SAT scores.

3.3. Instrument

Three sets of multiple-choice tests were developed to measure students' knowledge of English vocabulary and grammar that differed in the number of options (3, 4 and 5 options). Specifically, the test (see Kim, 2009 for sample items) comprised four different categories of items: two types of vocabulary items (sentence completion and synonym) and two types of grammar items (sentence completion and grammaticality judgment).

The initial version of the test was composed of 60 items, with 15 items for each of 4 types. However, too difficult items (n=22), the items with the IF value of lower than 0.4, were eliminated from data analysis to avoid measurement error because test takers are likely to make a wild guess when solving difficult questions. Therefore, 22 items were excluded from the analysis and the final version of the test contained 38 items: 13 three-option items, 13 four-option items, and 12 five-option items. Table 2 summarizes the number of items for each item category and option type.

Table 2 Composition of the Test

Item Category	3-option	4-option	5-option	Total
Vocabulary (Completion)	2	3	2	7
Vocabulary (Synonyms)	4	2	2	8
Grammar (Completion)	2	3	4	9
Grammar (Judgment)	5	5	4	14
Total	13	13	12	38

3.4. Data Collection and Analysis

For data collection, the English test containing items with different option formats was administered to 337 college freshmen enrolled in English conversation courses. The test was administered in the first week of the spring semester, and the students' course instructors were recruited to administer the test as proctors. It took approximately 40 minutes for students to complete the test.

The study used the item response theory (IRT) instead of the classical test theory (CTT). Item characteristics based on the classical test theory are defined in relation to the test-taker group; therefore, the results from the CTT are likely to change according to the ability of the test-takers. Given that the participants in the study are different from those of prior studies, the IRT was chosen to reduce the intervening effects of the ability of test-takers. The results based on the IRT are produced on the basis of stable population and thus item characteristics (IC) obtained from the IRT do not fluctuate so much due to the ability of test-takers. For this reason, a 3-parameter IRT model was used, and the model produced indexes of item difficulty, item discrimination, and pseudo-guessing.

4. Results and Discussion

This section first reports the item facility index and item discrimination index of the whole test. The next step is to trace how these measures change according to the number of options. For this analysis, the TestAn program for Windows was used.

4.1. Analysis of the Total Items

The item difficulty (b parameter, IF) of the overall test was appropriate at 0.465. The index of the item facility was in the range of -0.5 and +0.5 (Seong, 2001) and thus indicated a moderate level of item difficulty. The Cronbach Alpha or the index of internal consistency was moderate low (r=.68).

It is important to note that the item discrimination index (a parameter, DI = 0.87) was between 0.65 and 1.34. This indicates an appropriate level of discrimination according to Seong (2001, p. 42). Table 3 summarizes the result.

Table 3 Item Analysis for the Entire Test

Items	Subjects	Mean	Maximum	α	*b* parameter (IF)	*a* parameter (DI)
38	337	27.75	38	.68	0.465	0.867

**b* parameter: Item difficulty or Item facility (IF)
a parameter: Item discrimination (DI)

4.2. Analysis by Option Types

Table 4 shows the item facility and item discrimination index of different option types. As in the table, 4- and 5-option items were in the moderate level of difficulty. Henning (1987) suggested that item facility is most appropriate when it is around the middle of the difficulty range. It is interesting to note that the item facility index of 3-option items was over 0.5, which means they were difficult according to the criteria Seong (2001) suggested.

Table 4 Statistics of Item Analysis by Option Types

No. of Options	Items	*b* parameter	*a* parameter	*c* parameter
		(IF)	(DI)	(Guessing)
3	13	0.54	1.053	0.319
4	13	0.39	0.822	0.311
5	12	0.45	0.754	0.221

**b* parameter: Item difficulty or Item facility (IF)
a parameter: Item discrimination (DI)
c parameter: Guessing

The item discrimination power for all the three item groups were found to be positive and appropriate. This finding indicates that regardless of the number of options, all the items were good enough to discriminate between weak and strong students in the ability being tested. What is notable is that the item discrimination power decreased in proportion to the number of options. More specifically, the index of item discrimination (a parameter, DI) was greatest with 3-option items (a=1.05); the item discrimination power was lowest with 5 option items (a=0.75). Table 5 summarizes the statistics of individual item analysis across the three option types.

As shown in the table, none of the items in the 4-option group and in the 5-option group displayed the DI of 1.34 and above, indicating that these items did not have high discriminatory power. From the comparison of DIs for individual items, it can be inferred that the discrimination index was generally higher for 3-option items than for 4- or 5-option items.

Table 5 Statistics of Item Analysis by Option Types(Individual Items)

3-option items				4-option items				5-option items			
Items	*b*	*a*	*c*	Items	*b*	*a*	*c*	Items	*b*	*a*	*c*
4	-0.81	0.73	0.33	7	0.74	0.83	0.28	11	-0.87	0.81	0.22
5	-1.25	1.45	0.30	9	-1.28	1.24	0.28	15	1.22	0.83	0.19
16	0.86	1.34	0.29	10	-0.50	0.98	0.28	27	0.66	0.83	0.20
17	-1.23	0.64	0.34	21	2.62	0.61	0.41	29	-1.18	0.74	0.22
18	0.44	0.96	0.31	23	-1.76	1.03	0.29	41	1.15	0.59	0.24
19	0.44	1.02	0.30	36	-0.78	0.68	0.32	43	0.58	0.85	0.20
32	1.06	1.13	0.33	39	-1.27	1.03	0.30	44	0.37	1.03	0.20
33	0.54	1.09	0.34	40	0.47	0.56	0.31	45	0.29	0.53	0.25
46	1.41	1.05	0.27	51	0.63	0.82	0.30	56	0.11	0.54	0.24
47	1.86	1.17	0.32	52	1.10	0.87	0.29	58	1.15	0.73	0.23
48	0.21	1.28	0.30	53	1.29	0.82	0.33	59	1.45	0.87	0.24
49	3.83	1.03	0.38	54	2.18	0.60	0.33	60	0.48	0.69	0.23
50	-0.39	0.81	0.34	55	1.59	0.62	0.34				

There are two notable observations we can make in the comparison. First, 3-option items seem to be more difficult than 4- or 5- options. Second, 3-option items better discriminate between strong and weak students. For instance, Item 16, 32, 46, 47, and 49 can be classified as difficult, in that they have the IF value of 0.5 and above. Item 5 is one of the 3-option items with high discrimination power.

5. *Choose one that is most appropriate for the blank.*
Since Helen's lived in Korea for more than five years, she is _______ to the Korean way of life.
① due ② related √③ accustomed

Although this item is relatively easy according to Seong's (2001) criteria, it seems to have discriminated well between strong and weak examinees. Notable here is how these three groups of option demonstrate item discrimination power. While only one item in the 3-option group was found to have low discrimination (Item 17), four items in the 4-option group (Item 21, 40, 54, and 55) and 3 items in the 5-option group (Item 41, 45, and 56) displayed low discrimination power.

This might be because each distracter in 3 options is likely to be more distinctive from each other than in 4- or 5- options. The inclusion of more distracters is liable to produce distracters that are similar to one another in form or meaning. Item 21 is an example of 4-option items where two of the distracters are not fully functioning. In contrast to Item 5, where all the three options are different, two of the options (① and ②) in Item 21 are similar in form, i.e., prepositional phrases. This might have affected students' selection of the correct answer. Since the two options give clue to examinees that one of them may be a correct answer, the other two options (③ and ④) may become nonfunctional distracters.

21. *Choose the one that is closest in meaning to the underlined part.*
Convection currents in the air can carry ocean moisture <u>aloft</u>.
① over the hills √② into the air ③ discharges ④ particles.

The problem of nonfunctional distracters is also observed in some items from the 5-option group (Item 41, 45, and 56). The following example illustrates the case in point.

41. *Choose one that is most appropriate for the blank.*
A: That's not the best way to do that.
B: How else ______________?

① can be it done ② it can be done ③ it can be doing
√④ can it be done ⑤ can it be doing

By examining the distracters, we come to see that the question is asking a number of things, such as what "it" refers to, what kind of sentence structure is needed in B, and/or what role the auxiliary verb "can" plays. However, examinees are likely to solve the question using test-taking strategies. They may group the options ①, ②, ④ as one category and the options ③ and ⑤ as another category because of their similarity in form and choose an answer in one of the categories. From the examples of items given above, we can see that distracters in the 3-option format are more distinct from each other whereas some

distracters in the 4- or 5-option format are likely to be nonfunctional or dysfunctional.

Although the difference in their discrimination power was marginal, the finding that 3-option items showed greater discriminatory power than 4- or 5-option items deserves our attention. This finding contradicts with Lee and Kwon's (2002). Lee and Kwon (2002) conducted vocabulary and grammar test items different in the number of options and found that the 5-option items had the greatest discrimination power. In contrast, the findings of the present study seem to be consistent with Kim's observation (2009) that 3 option items displayed high discrimination power. The finding also confirms the results from previous studies that examined non-Korean contexts (Bruno & Dirkzwager, 1995; Costin, 1970; Haladyna & Downing, 1993; Owen & Froman, 1987; Shizuka, Takeuchi, Yashima, & Yoshizawa, 2006).

Notable in the present study is that item discrimination power decreased as the number of options increased. In other words, the 3-options (with 2 distracters) worked better than 4- or 5-options (with 3 or 4 distracters) in discriminating between strong and weak examinees; 4 options functioned better than 5 options in terms of item discrimination. The item quality enhanced as the number of options decreased, with the discriminatory power greatest in the 3-option item group. This finding is different from previous findings by both Lee and Kwon (2002) and Kim (2009) that the 5 option items had the greatest discrimination power.

5. Conclusion

The purpose of the study was to examine if and how the number of options in a multiple choice test incurs changes in the item facility and the item discrimination index. A test was designed with English grammar and vocabulary items that were different in the number of options. A total of 337 college students taking freshman English classes took the test in the first week of the spring semester. These students were different from Kim's (2009) participants because they were all engineering majors studying at a major university in Seoul. For item analysis, the study used the Item Response Theory, from which the item facility and the item discrimination index were obtained. The item discrimination index was then used to examine the effects of the number of options. The study found that the 3-option group demonstrated the strongest item discrimination power. This finding is meaningful in that the effectiveness of 3 options is confirmed with a different group of examinees.

While these findings are beneficial for language testing, they are limited for the following reasons. First, because the difficulty of the items in different option types was not controlled, it might have influenced the item facility (b parameter) and the discrimination index (a parameter). Fortunately, the difference in item difficulty was marginal across the three option groups as reported in Table 4. Second, the stems given were not identical across different option types, and this might have yielded confounding results. Moreover, considering that the 3-PL IRT model requires over 1,000 examinees and at least 60 items, the number of the students (n = 337) was not big enough to use the model.

Although the findings of the study have some limitations, they deserve our attention. First of all, the finding that the 3-option format had the highest discrimination is consistent with earlier findings (Bruno & Dirkzwager, 1995; Costin, 1970; Haladyna & Downing, 1993; Owen & Froman, 1987; Shizuka, Takeuchi, Yashima, & Yoshizawa, 2006). This finding also confirms Kim's (2009) suggestion that 3 properly 'functioning' options can be as good as or even better than to 4- or 5-options; an addition of ineffective options may end up lowering the item discrimination power because the pressure to develop more alternatives might lead to constructing nonfunctioning or dysfunctional distracters (Bruno & Dirkzwager, 1995; Haladyna & Downing, 1993; Owen & Froman, 1987). Therefore, having three options that are fully functioning might be better than having additional options whose distracters do not play functional roles. This would help test-makers to avoid "noise" in their test items (Bruno & Dirkzwager, 1995, p. 962).

The finding about the number of options in multiple choice tests compels us to rethink the nature and the purpose of a test in each context of use. The decision on the number of options has to be based on the cognitive capacity of the test-takers, nature of tests and other potential variables that influence the test validity. The number of options may be irregular, as in the case where the same test contains some items with three options, some with four, and some with five (Henning, 1987).

Therefore, it can be problematic if we blindly follow the format of high-stakes tests, such as the Korean SAT. The Korean SAT has its own distinctive goals and properties, which may not be directly applicable to other tests. Moreover, the fact that the SAT is widely used does not mean that it is valid and reliable for other types of tests. The test item writers should consider the nature and purpose of the test and then determine the number of options accordingly. If one or two among five options are nonfunctional, we can reduce the number of options to 4 or 3. By eliminating nonfunctioning options, we may be able to construct more items, and the test can cover more content. Five options are not desirable for young learners anyway because they might be easily distracted by the test itself.

The prominence of three options found in the study, however, does not necessarily mean that all multiple choice items should adopt the three-option format. Rather, the lesson is to specify the role of all available options while eliminating those distracters that do not seem to work properly since "poorly designed distracters could easily cue a student to the correct answer (Woodford & Bancroft, 2004, p.6). Poorly designed distracters might pull the examinees away from the very construct the test is designed to measure. This argument therefore reiterates the old adage that says quality beats quantity. It is better to have fewer options than having more that are either dysfunctional or nonfunctional. The items with fully functioning options are likely to become more reliable and valid.

In this regard, it is helpful to listen to what Heaton (1988) has to say about designing functional distracters. To summarize Heaton's (1988) suggestion, distracters are plausible when they are "based on (a) mistakes in the students' own written work, (b) their answers in previous tests, (c) the teacher's experience, and (d) a contrastive analysis between the native and target languages" (p. 32). Other suggestions include that distracters should be given in a grammatically correct form, and that they should not be too difficult lest strong examinees should be trapped by such distracters.

References

Aamodt, M.G. and T. McShane. 1992. A meta-analytic investigation of the effect of various test item characteristics on test scores and test completion times. *Public Personnel Management, 21*, 151-60.

Alderson, J.C., C. Clapham, and D. Wall. 1995. *Language test construction and evaluation*. Cambridge, UK: Cambridge University Press.

Berríos, G., C. Rojas, N. Cartaya, and Y. N. Casart. 2005. Effect of the number of options on the quality of EST reading comprehension multiple choice exams. *Paradígna, 26*, 89-116.

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

Cizek, G. J. and D. M. O'Day. 1994. Further investigation of nonfunctioning options in multiple-choice test items. *Educational and Psychological Measurement, 54*, 861-872.

Cohen, A. 1994. *Assessing language ability in the classroom* (2nd ed.). Boston, MA: Heinle & Heinle.

Costin, F. 1970. The optimal number of alternatives in multiple-choice achievement tests: Some empirical evidence for a mathematical proof. *Educational and Psychological Measurement, 30*, 353-358.

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

Ebel, R. L. 1951. Writing the test item. In E.F. Linquist, ed., *Educational Measurement,* 185-249. Washington, DC: American Council on Education.

Haladyna, T.M. and S. M. Downing. 1989a. A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*, 37-50.

Haladyna, T.M. and S. M. Downing. 1989b. The validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*, 51-78.

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

Heaton, J. B. 1988. *Writing English language tests*. New York: Longman.

Henning G. 1987. *A Guide to language testing*. Boston: Heinle and Heinle Publishers.

Kasper, G. and K. R. Rose. 2002. *Pragmatic development in second language*. Mahwah, NJ: Erlbaum.

Kim, S.-Y. 2009. The optimal number of options in a multiple-choice English test. *English Language Teaching, 21*, 69-87.

Lee, W.-K. 2007. *A guide to English language testing*. Seoul: Munjin Media.

Lee, W.-K. and O.-R. Kwon. 2002. *On the number of options in a multiple-choice English test*. The 5th International Conference on Language Testing in Asia. October 4-5, 2002. The Society for Testing English Proficiency, Inc.

Owen, S. V. and R. D. Froman. 1987. What's wrong with three-option multiple choice items? *Educational and Psychological Measurement, 47*, 513-522.

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*, 3-13.

Seong, T. 2001. *Item response theory: Understanding and application*. Seoul: Gyoukgwahaksa.

Shizuka, T., O. Takeuchi, T. Yashima, and K. Yoshizawa. 2006. A comparison of three- and four-option English tests for university entrance selection purposes in Japan. *Language Testing, 23*, 35-57.

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

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

Weir, C. 1988. *Communicative language testing*. Exeter, Great Britain: University of Exeter.

Woodford, K. and P. Bancroft. 2004. *Using multiple choice questions effectively in information technology education*. Retrieved from http://www.ascilite.org.au/conferences/perth04/procs/woodford.html

Sung-Yeon KIM
Dept. of English Education
Hanyang University, Seoul (133-791)
TEL: (02) 2220-1141
E-mail: sungkim@hanyang.ac.kr

Received: 2011.11.02
Revised: 2011.12.05
Accepted: 2012.02.22