THE DIFFERENCES AMONG THREE-, FOUR-, AND FIVE-OPTION-ITEM FORMATS ON A HIGH-STAKES ENGLISH LISTENING TEST

By

HyeSun Lee

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

MASTER OF ARTS

Teaching English to Speakers of Other Languages

2011

ABSTRACT

THE DIFFERENCES AMONG THREE-, FOUR-, AND FIVE-OPTION-ITEM FORMATS ON A HIGH-STAKES ENGLISH LISTENING TEST

By

HyeSun Lee

The aim of this research is to investigate the differential effects of multiple-choice items with three, four, and five options on a high-stakes, English language listening test. Three-option multiple-choice items on an English listening test were compared with items with four and five options in terms of average total score assignments, average item facility, average item discrimination, overall test reliability, and processing time. Three prep English listening tests for the CSAT (College Scholastic Aptitude Test), each with five-option items, were adapted into parallel forms with four- and three-option items by eliminating the least plausible options, as selected by 73 Koreans. A total of 264 Korean EFL learners, divided into three groups, participated in the study. Each group took tests with five-, four-, and three-option items. The test administrations were based on Latin squares to control for order and practice effects. Results indicated that the average scores between tests with three- and five-option items differed significantly. However, there was no significant difference in the average item facility, average item discrimination, and overall reliability among the tests with the different number of options. Regarding time on exam, the three-option-item test took 11% less time than the five-option-item test. Survey data from the 265 test takers revealed that 54% preferred the three-option-item format. Also, 68% agreed administering the CSAT with three-option items would be preferable. Results are discussed from the perspective of statistical, cognitive, emotional and contextual factors in determining the optimal number of options.

To my parents, HK, and HJ whom I always love and miss so much.

Acknowledgements

I would like to thank the principal and teachers for allowing me to collect data in their school; students for their voluntary participations; my classmates for their generous assistance in revising tests. Finally, I wish to express my deep gratitude to Dr. Paula Winke and Dr. Charlene Polio who provided insightful comments throughout my thesis writing.

Table of Contents

List of Tables

List of Figures

Introduction

Haladyna (2004) stated that multiple-choice is preferred as a test item format due to its scoring objectivity, higher reliability, and efficient administration even though many test writers have difficulty in writing multiple-choice items. It is challenging for test developers to word the stem, to make a single, clear correct answer, and to write plausible distracters. On the other hand, test takers report that multiple-choice items are much less demanding and make them less anxious, so this format is preferred compared with other test item formats (Haladyna, 2004). In addition to the advantages of the automaticity of scoring and less demanding workload for the test takers, the multiple-choice format produces relatively higher reliability than other essay formats, as Haladyna claimed. Reliability is an important factor to be considered in testing. Brown (2005) defined reliability as “the extent to which the result can be considered consistent or stable” (p. 175). Also, Haladyna argued that multiple-choice items make it possible to cover content knowledge and a wide range of skills, as other types of items such as short-answer or essays do. Thus, the multiple-choice item format has been widely used in many language tests. Also, a lot of studies have been conducted regarding the best type of multiple-choice format. One area of contention is the optimal number of options in a multiple-choice item, and this argument has lasted more than 80 years (Rodriguez, 2005).

Before proceeding with the exact details of this study, I first discuss prior empirical research that has been conducted regarding the number of options that multiple-choice items should have. This is tied in with researchers’ discussions on guessing that results from the multiple-choice item format, thus I review this area of research as well. I also discuss the College Scholastic Aptitude Test (CSAT) in Korea, a high-stakes, five-option multiple-choice test, which provides the general context for this study on the optimal number of options for multiple-choice items. Then, I introduce the research questions and research method of this study. In the result section, analysis of quantitative and survey data will be presented with an outline of the major findings. Finally, I conclude with a discussion, research implications, and limitations and future directions.

Background

Optimal Number of Options

As mentioned above, the construction of multiple-choice items, especially the construction of options, takes item writers a lot of time. The optimal number of options for items has been debated considerably (Haladyna & Downing, 1993, 2002; Lord, 1977; Trevisan, Sax, & Michael, 1991). In the review of multiple-choice item-writing guidelines, Haladyna, Downing, and Rodriguez (2002) reported 70 percent of 31 guidelines of multiple-choice item-writing cited the rules of “write as many plausible distracters as you can” and only four percent of the guidelines were against this. Plausible distracters should be chosen by low performers and eliminated by high performers. In other words, a plausible distracter should be a good, functional distracter that looks like a correct answer to test takers who do not have enough knowledge or skill concerning what an item intends to measure (Haladyna, 2004). However, it is hard to write many plausible options. Thus, the ultimate goal is often to construct test items with an optimal number of plausible distracters. Since Ruch and Stoddard (1925) claimed three-option items do not have detrimental effects on test outcomes, compared with four- or five-option items, many empirical studies have been conducted that investigate the optimal number of functional options a multiple-choice item should have (e.g., Haladyna, Downing, & Rodriguez, 2002; Rodriguez, 2005).

In the early of 1940s, Lord conducted a study on the number of options in multiple-choice items, claiming that three options were optimal. In a mathematical demonstration, Tversky (1964) argued that items with three options maximize the power of a test, test discrimination, and information in a test, provided that the total number of options would be fixed. Costin (1970) revised four-option-item tests in introductory psychology for the Air Force by random elimination of one option from each item. Costin reported that mean item discrimination and reliability of three-option items were better than those of four-option items. Straton and Catts (1980) compared two-, three-, and four-option-item tests in a fixed number of options in Tversky's condition— the total exam time is in proportion to the number of options in a test (1964). Straton and Catts developed four different versions of a trial test for a college entrance exam in economics: one version of 60 items with two options, two versions of 40 items with three options. The only difference between these two versions in the sets of 40-items was in the way the options were discarded—random discarding of options in one 40 items set and eliminating the worst option in the other 40 items set. Finally, one version of 30 items with four options was adopted for the study of Stranton and Catts. Administering four tests to 260 students at Sydney technical college, the results showed an increase in item difficulty and item discrimination when there were fewer numbers of options. Also, reliability and the Standard Error of Measurement of items with three options were equal to or better than the items with two or four options.

Owen and Froman (1987) conducted an experiment with two parallel psychology final comprehensive exams, consisting of a pretest and posttest with three- and five-option items. Both of the tests were composed of 100 items, half of which were three-option items and the other 50 items with five options. In this research, each of 114 undergraduate students enrolled in an educational psychology class took both three-option items and five-option items as a pretest and a posttest over a ten-day period—if a student took a pretest which consisted of first 50 items as five-option questions and the other 50 items as three-option questions, then he or she would take a posttest in which 50 three-option items came up first, then 50 five-option items were followed. Correlated *t*-tests were used to estimate item difficulty and item discrimination. The results indicated that there were no significant differences between the three-option test and the five-option test in mean item difficulty and mean discrimination. Also, total test scores between the two tests were not so different. After finishing the posttest, the students voted for the preferred form. The result showed that 111 students preferred three-option items to five-option items. Also, Owen and Froman reported that response time per an item were significantly reduced with fewer options. Owen and Froman claimed that more options were likely to give more unintended clues that may help test-wise students. According to Owen and Froman, if three-option items provide equal statistical outcomes to four-or five-option items, testwiseness may not be an important issue in three-option items. Also, they claimed that content validity and reliability could be improved by adding more items into a three-option-item test due to the saved testing time.

Trevisan, Sax, and Michael (1991) investigated the interaction between student ability and the number of options in items, and reliability and validity affected by this interaction. From a verbal test in the Washington Pre-College Admission Test Battery (University of Washington, 1983) which originally consisted of items with five options, Trevisan et al. created three- and four- option items by using the point biserial coefficients from the standardization data and by discarding the least discriminating options. The students were divided into high, medium, and low ability groups based on their GPAs. A total of 435 junior class parochial high school students took only one of three tests—the five-option test, four-option test, or three-option test. The result showed that there were significant differences in reliability on the three tests for low-ability group. Validity was not significantly different across groups. Trevisan et al. claimed that the optimal number of options was three when the groups were combined.

Also, Haladyna and Downing (1993) reported the summary of empirical and theoretical studies about three-option items. Also, they investigated the frequency of effective and ineffective distracters by analyzing three standardized multiple-choice tests: a test of a graduate medical education program for physicians which had 200 five-option items, ACT Assessment which consisted of total 127 four-option items, and a state certification examination in the health sciences in which there were 150 four-option items. In analysis of the 477 items including 2,108 options from the frequency of distribution of options, Haladyna and Downing claimed that items with three options may be a “natural limit for multiple-choice item writers in most circumstances” (p.1008), reporting that only one or two distracter(s) were effective among two thirds of the items, and that items with three effective distracters were between 1.1 percent and 8.4 percent of the 477 items, and all 200 items with five options did not have four effective distracters. Haladyna and Downing also found that there was no relation between the number of effective distracters and item difficulty, whereas items with more plausible distracters could affect items’ discrimination. Furthermore, Haladyna and Downing stated that more options per item could contribute to better reliability if the distracters performed adequately, which would be rare. Different from Haladyna and Downing’s result, Crehan et al. (1993) reported that items with three options had a little bit higher item difficulty than four-option items and no difference in item discrimination among the tests—220 university students from a psychology class took one of the manipulated tests. Crehan et al. concluded that three-option items had three benefits: they were easier to write, had more effective distracters, and took less time to administer.

Cizek and O’Day (1994) examined item difficulty and item discrimination, comparing 32 five-option items with four-option items in a test for certification in a medical specialty with 700 participants. The result showed that deleting nonfunctioning options made a slight increase (but not a significant a difference) in item difficulty. Based on their report, item discrimination indexes of both tests were not so different and the reliability of the four-option test was slightly better than the five-option test. Bruno and Dirkzwager (1995) investigated the optimal number of options, comparing the results with those from statistical perspectives. According to Bruno and Dirkzwager, information from the item could increase with more options. However, too many options per item would produce less information because the amount of information per option has a maximum. Bruno and Dirkzwager stated that when each option was equally probable, items with three options would give test takers an opportunity to obtain the maximum information per option. Bruno and Dirkzwager stated that the result from the information theory approach was in line with the previous findings from statistical and empirical research (e.g. Costin, 1970; Cox, 1980; Ebel, 1969; Grier, 1976; Lord, 1944; Rush & Stoddard, 1927; Torabi-Parizi & Campbell, 1982; Tversky, 1964).

Rogers and Harley (1999) investigated the susceptibility to testwiseness in a high-stake school-leaving mathematics examination with 158 senior high school students, comparing four-option items with three-option items. According to Millman, Bishop, and Ebel (1965), testwiseness means “a subject’s capacity to utilize the characteristics and formats of the test and /or test-taking situation to receive a high score. Testwiseness is logically independent of the subject matter for which the items are supposedly measure” (p. 707). They investigated differences in item difficulty, item discrimination, and reliability. A total of 31 multiple-choice items with four options in the 1992 version of the mathematics 30 test was used. Thirteen items were testwiseness susceptible and 18 items were not testwiseness susceptible. Based on the original four-option test, a three-option test was developed by eliminating a distracter in five different ways. Both of three-and four-option tests were administered in the same class or section. Odd numbered students took the original four-option test while even numbered students took the three-option test. Comparing item difficulty, item discrimination, and reliability between four-option items and three-option items, Rogers and Harley reported that testwiseness was less affected when three-option items were administered. Also, the result showed that item difficulty in fewer-option items was increased and reliabilities of both four-and three-option items were almost equal. Rogers and Harley also reported that when asked if they would develop a three-option test or four-option test, the teachers preferred a three-option test because of the difficulty in writing three functional distracters.

Recently, in their review of multiple-choice item-writing guidelines for classroom assessment, Haladyna et al. (2002) summarized the results from empirical research regarding item difficulty, item discrimination and reliability. Haladyna et al. stated that different number of options did not affect item discrimination while it changed item facility. Haladyna et al. reported that according to five studies (Landrum, Cashin, & Theis, 1993; Rogers & Harley, 1999; Sidick, Barrett, & Doverspike, 1994; Trevisan, Sax, & Michael, 1991, 1994) tests with fewer numbers of options could lead to a decrease in item difficulty. However, two studies (Cizek & Rachor, 1995; Crehan, Haladyna, & Brewer, 1993) indicated that fewer-option tests resulted in higher item difficulty. Haladyna et al. stated that, citing Cizek and Rachor’s result, the fewer options on a test, the more the item discrimination. However, research from Crehan et al. showed no significant difference in item discrimination when the number of options varied. When it comes to reliability, a study from Trevisan et al. (1994) and Sidick et al. indicated that fewer options increased reliability. However, Trevisan et al. found no change in reliability. Haladyna et al. also mentioned that it did not seem that the item writers could write three plausible distracters with consistency, claiming that four or five options were not worthy of the extra effort. Instead, three options could be enough in most testing situations because the most important thing is how functional the item distracters are, not how many numbers of options the item has.

Rodriguez (2005) reported on a meta-analysis of 27 studies that investigated the number of options multiple-choice items should have. Rodriguez reported significant changes in item facility were found when the number of options was reduced, especially research in which the number of options was reduced to two options. Even though item discrimination was reported to show significant changes, there was one study in which no change was reported by reducing from five options to three options and a slight increase of item discrimination was found in research when the number of options was changed from four to three. Furthermore, from the perspective of test reliability, it was stated that in most cases the reduced options resulted in a decrease of reliability. However, there was a case that reliability was not affected significantly by reducing options from five to three in the meta-analysis or that reliability increased by changing a four-option test to a three-option test.

Shizuka, Takeuchi, Yashima, and Yoshizawa (2006) investigated the effects of different numbers of options in an English reading test for a university entrance exam. They changed an original four-option reading test to a three-option test by discarding the least chosen option. A total of 38 five-option items were revised into ten five-option items and 28 three-option items. The original five-option test was administered to all applicants to the university and 1000 Japanese applicants were randomly selected to be analyzed. In this study, 192 Japanese participants who did not take the original test took the three-option test in about nine weeks.

Before analyzing data, the comparability of two groups was run on the ten common items. However, a *t*-test result showed that four-option group was significantly higher in their reading ability. Thus, it was mentioned that common item equating in the Rasch measurement was operated by using FACET v.3.0 software. Researchers stated that for a distracter analysis, the "actual equivalent number of options (AENOs) defined by Sato and Morimoto (1976)" (p. 46) was computed. The results indicated that the average item facility and average item discrimination between the four-option test and the three-option test were not significantly different. Also, test reliability in the test with three-option items was not significantly lower than in the test with four-option items. Furthermore, in their analysis of distracters, they found that the average number of actual functional distracters was less than two. Thus, researchers claimed that three-option multiple-choice items would be optimal, considering item facility, item discrimination, test reliability and efficiency of information availability in tests.

Guessing

The issue of guessing can be raised if multiple-choice items with fewer options are adopted. According to Haladyna (2004), the effects of guessing are quite overestimated. Haladyna stated that the chance of guessing correctly on ten items is about .00000001% when the items have four options. Thus, the probability of getting a high score on a test of 50 items, due to guessing, is quite remote. Also, Costin and Kolstad et al. (Costin, 1976; Kolstad, Briggs, & Kolstad, 1985) claimed that test takers were not likely to randomly guess. Instead, test takers seemed to delete the least attractive distracters, resulting in only two or three remaining options. According to Costin and Kolstad et al. plausible distracters would prevent test takers from getting correct answers through blind guessing. In other words, they argued that the number of options would not create an effective alternative for an undeserved lucky chance.

CSAT

As one of the high-stakes and norm-referenced tests, the College Scholastic Aptitude Test (CSAT) is used in Korea to screen out applicants for entrance into universities. The CSAT consists of multiple-choice items with five options because it is believed that such items are better at discriminating among higher and lower ability students than four-option multiple-choice items and result in fewer passing test candidates. The Korea Institute of Curriculum and Examination (KICE) is responsible for developing test items, administering the CSAT, reporting the CSAT results, and analyzing the CSAT data. KICE has implemented five-option items in the CSAT since 1993—before that, all items on the test had four options. Also, each office of education, including Seoul metropolitan office of education, has been developing the CSAT prep-tests and administering them five times a year. All prep-CSAT materials have five-option multiple-choice items, as in the CSAT.

According to Choi (2008), it is not an exaggeration to say that the CSAT results determine the future social status of each student in Korea. Thus, the national obsession over the CSAT in Korea is unimaginable as *Asia Times* described. On the test day, under the government policy, employees in government and companies arrive to work one hour later than usual to avoid traffic jams. The Korean stock market also opens at ten in the morning, which is one hour later than usual. Especially during the listening test, domestic and international flights are not allowed to take off and land, as the noise may influence the test results. All cars are not allowed to honk horns near the testing site on that day (*Asia times*, 2005).

As noted in the language testing literature, the development of five-option items takes a lot of effort, time, and money (Budescu & Nevo, 1985; Delgado & Prieto, 1998: Haladyna, 2004; Haladyna & Downing 1993; Owen & Froman, 1987; Rogers & Harley, 1999; Straton & Catts 1980). This is especially true in the case of the CSAT, which is administered once a year. In developing items of the CSAT, more than 650 personnel including professors and teachers are isolated in a secure place for 33 days because of the high-level of test security that is needed (*The Hankyoreh*, 2005).

Research Questions

Considering that many empirical studies have been supporting that items with three options are optimal, it is highly worthwhile investigating the claims that fewer options do not affect test statistics and outcomes, especially in the high-stakes CSAT in Korea. Also, given that studies have rarely been conducted regarding an English listening test in a high-stakes test, this study will contribute to the future development of a high-stakes listening test item.

An additional claim in the language testing literature is that fewer options in multiple-choice items result in less processing time and thus can lead to shorter exam times that can be more effective in terms of less anxiety for test takers and less money to administer the test (Budescu & Nevo, 1985; Delgado & Prieto, 1998; Haladyna, 2004; Haladyna & Downing 1993; Owen & Froman, 1987; Rogers & Harley, 1999; Straton & Catts 1980). On the other hand, Owen and Froman (1987) claimed that test developers could, consequently, add more items to a test in place of the saved time, which would increase content validity and reliability of a test. In the case of the CSAT English language test, the listening section takes almost 20 minutes in total as part of the 70-minute test. Therefore, it will be worth investigating the claim that three-option items in a multiple-choice format can shorten the total testing time due to the reduced processing time per item.

Using a mixed-method research design, I adopted a survey questionnaire to triangulate quantitative data (Mackey & Gass, 2005). The questionnaire data provide the perspectives of examinees on the different item formats. The survey questionnaire explored the examinees' opinions about the three-option-item format and the administration of this format on the CSAT. The survey data supplemented the quantitative data, contributing to a better understanding of the results (Brown, 2001).

The context for this study is in South Korea with EFL learners taking the prep-CSAT English language listening tests. The research questions are as follows:

1. In terms of the average total score assignment, average item facility, average item discrimination, and overall reliability, are there significant differences among three-, four-, and five-option multiple-choice item formats on a high-stakes English listening test?
2. Can a listening test with three-option items function as selectively as those with five-option items in screening applicants for university entrance?
3. Can a three-option multiple-choice item format significantly reduce the exam time?
4. Do test takers prefer a three-option-item format on a high-stakes test?

Methods

Participants

A total of 264 Korean high school students (40 males and 224 females) from six intact, tenth- grade English classes preparing to take the CSAT to enter a university participated in this study. Each class size was composed of 45 to 50 students. Initially, I started with 300 participants. However, some of the participants could not take all three tests due to absence from school or other reasons. As a result, 36 participants were removed from the initial data set since they did not complete all three tests. Located in Seoul, this high school, in which I have taught for six years, has students who are approximately eighty percent female and twenty percent male. The principal of the school generously approved of this research experiment. The students in this school participated in this experiment voluntarily. At this school, students are taught English four hours per week in Korean by Korean teachers. The participants had been learning English for seven years through classroom-oriented instructions. The instruction is intensively focused on reading comprehension and grammar in preparation for the CSAT. They are usually exposed to written materials in class. Like those in other EFL learning environments, they rarely have opportunities to use English outside the classroom. In the next section, I will review the overall structure of the CSAT English test and introduce the listening tests adopted in this study.

Materials

The CSAT English Test

The English test in the CSAT is composed of 50 multiple-choice items in total: 17 listening items, two or three grammar and vocabulary items, and 30 or 31 reading comprehension items. Even though the portion of listening items in the CSAT is quite large, listening skills are not often dealt with in class. Instead, reading and grammar are the main focus of instruction. Thus, the listening section in the CSAT English language test is prepared for by self-learning or private tutoring.

In the CSAT English listening test, all 17 multiple-choice listening items with five options are printed in the test booklet with 33 reading items which also have five options. Test takers can look over the listening items before listening to the audio files. Each item has one audio file. That is, test takers listen to 17 different audio files, one for each item. First, test takers listen to the direction for an item. Following the direction, an audio file is played through audio speakers at a testing site. Then, test takers mark the correct answer based on the given five options which are printed in the test booklet. For example, the direction such as *listen to the conversation and choose what the man will do next* is given. Then the audio file of a conversation between a man and a woman is played and test takers choose their answer from five written options. After ten seconds, automatically another direction is played for the next listening item.

The Prep-CSAT Listening Tests Adopted in Research: Nine Listening Tests

In this research, I adopted three English listening prep-tests for the CSAT. The prep-tests for the CSAT are administered to high school students seven times a year (In March, May, June, July, September, October, and November): The offices of educational districts develop and administer the prep-tests. The three tests that this study used (versions I, II, and III) were actually administered in November, 2007 (version I), November, 2008 (version II), and November, 2009 (version III). All versions had the same format (five-option items) but differed only in content (see Appendix A). To investigate the main research effect, as a pre-research task, these three original tests (I5i, II5i, and III5i) were adapted into two parallel forms with four-option and three-option items by deleting the least plausible option, as selected by 73 native speakers of Korean. These new forms are Test I4i, II4i, and III4i (four-option items), and I3i, II3i, and III3i (three-option items). As a result, nine different tests were used in investigating the main research questions.

Survey Questionnaire

To triangulate the quantitative data, a survey questionnaire was used. This questionnaire asked all participants their preferred number of options (three, four, or five options) and their opinions about changing the option format in the CSAT. The survey questionnaire was written in the native language of participants, Korean, to avoid the concern that English proficiency may affect the quality of response as Mackey and Gass (2005) mentioned. The questionnaire consisted of seven items: Two closed-end questions, one Likert-scale question, and four open-ended questions (see Appendix B). The Likert-scale question (item three) was added to triangulate participant’s preference of item format (item one). This questionnaire was distributed to participants if they took at least one of the three test sessions (see Appendix B).

1. Closed-ended item: The preference among three-option, four-option, and five-option-item tests
2. Open-ended item: The reasons why a participant preferred three-, four-, or five-option-item tests
3. Likert-scale item about the three-option-item test: from 1(most dislike) to 7 (most like)
4. Open-ended item: Advantage of three-option-item tests
5. Open-ended item: Disadvantage of three-option-item tests
6. Closed-ended item: Agreement or disagreement with the administration of the three-option item format on the CSAT

7. Open-ended item: The reason why a participant agreed or disagreed with the administration of the three-option-item format on the CSAT

Procedures

Pre-Research: Revising the Three English Listening Prep-Test for the CSAT

The pre-research task to revise a test format was conducted in March, 2010. It was a part of the final project in a testing course that I took during the spring semester of 2010. A total of 73 Korean native speakers in Korea and U.S. participated in revising the three original tests with five-option items into four-option items tests (phase one) and three-option-item tests (phase two). In phase one, three groups of the 43 participants took two original listening tests with five-option items among the three original tests—I5i, II5i, and III5i (see Table 1). While taking tests, they were required to choose the least plausible options. Based on the wrong answers and the least plausible options selected by them, I deleted the least plausible options and revised the five option tests to four-option items (I4i, II4i, and III4i). In phase two, another 30 participants, who did not participate in phase one, took two of the revised tests with four-option items among three test (see Table 2). Like in phase one, they selected the least plausible option while taking the tests with four-option items. Finally, I revised the four-option-item tests to three-option-item tests (I3i, II3i, and III3i) by deleting the least plausible option. In revising them, Tversky's (1964) condition—the total exam time is in proportion to the number of options in a test— was not considered as in the study of Owen and Froman (1987). As a result, nine different tests were generated to investigate the research questions.

Table 1
Revision of the Three Original Prep-Tests to the Four-option-item Tests

Group	Tests taken		Tests created
1 (n=15)	I5i	II5i	I4i/II4i
2 (n=15)	II5i	III5i	II4i/III4i
3 (n=13)	III5i	I5i	III4i/I4i

Note: n = number of test takers.

Table 2
Revision of the Four-option-item Tests to the Three-option-item Tests

Group	Tests taken		Tests created
1 (n=10)	I4i	II4i	I3i / II3i
2 (n=10)	II4i	III4i	II3i/III3i
3 (n=10)	III4i	I4i	III3i/I3i

Note. n = number of test takers.

Main Research

The main data collection sessions were conducted from the second week to the fourth week of May, 2010 with the consent of participants from the high school in which I have taught in Korea. As discussed earlier, six intact English classes were divided into three groups to take nine different tests, administered to the three groups based on Latin squares, which controlled for order effects and practice effects. The three different groups took the three tests with a one-week interval between each administration (see Table 3). As a result, all participants took three-, four-, and five-option tests in different versions (version I, II, and III).

Table 3
Test Administration Schedule

		Session			
Group	Male/ Female	1	2	3	4
1 (n=86)	11/75	II4i (four-option)	I3i (three-option)	III5i (five-option)	Survey session
2 (n=89)	15/74	I5i (five-option)	III4i (four-option)	II3i (three-option)	Survey session
3 (n=89)	14/75	III3i (three-option)	II5i (five-option)	I4i (four-option)	Survey session

Note. n = number of test takers.

This study additionally investigated the claim that fewer-option items in multiple-choice questions result in less processing time and thus can lead to shorter exam times (Crehan, Haladyna, & Brewer, 1993; Owen & Froman, 1987). In the CSAT English language test, the listening section takes 20 minutes out of the 70-minute test. If a test with fewer-option items in the CSAT English listening test can significantly reduce the exam time—as mentioned, during this 20-minute CSAT English listening test administration, no plane is officially allowed to take off or land—it may contribute to the test efficiency, content validity, and reliability. Thus, this claim was examined by recording response time per item. In each test taking session, three participants (a total of 37 participants) recorded their response time per item by using a stop watch.

Finally, to explore test takers' opinions about the three-option-item format, I conducted a survey session with the participants after completing the third session. They took 10 to 15 minutes to answer the seven questions and the survey was done anonymously to protect the individual test takers' confidentiality (Dörnyei, 2003).

Scoring

Listening test.

A scoring system where some items are weighted with more or less points based on a pre-expected level of difficulty has been adopted on the CSAT to increase item discrimination. In a prep-English listening test for the CSAT, the most difficult item out of 17 items is given three points and the easiest item has one point. The other 15 items are given two points. However, in this study, each item was equally scored as one point for a correct answer and zero for an incorrect one, unlike the original test, with the rationale that the pre-expected level of difficulty can be different from the actual level of difficulty that will be identified only after the administration of a test.

Survey questionnaire.

Regarding the two closed-ended items, item one was coded three, four, or five based on the preferred number of options. Item six was coded dichotomously: one for agreement and two for disagreement. Item three, a Likert-scale item, was coded from one (most dislike) to seven (most like). Three open-ended items was processed through open coding. I decided initial categories through the first reading. Based on the initial categories, I coded item two, four, five, and seven. Then, I finalized these categories into detailed subcategories by investigating connected patterns and removing overlapping themes. Once final categories were listed, I coded 10 percent of the data. (Ten percent was based on the previous literature (Brown, 2001; Chandler, 2003). Then, to confirm the consistency of coding procedure, another Korean rater coded these 10 percent randomly selected data based on the final coding list. By comparing the coding results from another coder and me, the intercoder agreement was checked. Differences in coding were resolved through discussion.

Data Analysis

Quantitative Data

Preliminary Data Analysis

Before analyzing main effects, the comparability of three groups was checked by investigating achievement test scores. The achievement test was administered one month before the 1st session (the second week of May) of research. The achievement test consisted of listening items and reading comprehension items: Forty percent of the items were measuring listening skills and the rest of the items were for reading comprehension. As listening skills had significant correlation with reading skills (Hedrick & Cunningham, 1995, 2002), the achievement test scores were used to confirm the compatibility of the three groups in this study. The mean scores of three groups are as followings: The mean score of group 1 was 70.74, that of group 2 was 69. 98, and for group 3, was 70. 12 (Maximum score of the achievement test is 100).

Second, to ensure whether the three original tests with five-option items (I5i, II5i, and III5i) were of equal difficulty, total scores of each five-option-item test were analyzed by using Kruskal-Wallis test. Because Kolmogorov-Smirnov statistics showed that the distribution was not normal ($p < .05$, Table 4) nonparametric testing was used to check the equal difficulty of the tests. Kruskal-Wallis test indicated a significant difference across the three versions, $H(2) = 29.89$, $p = .000$. To explore significant differences, Mann-Whitney test was used with the adjustment of Bonferroni significance level as .017. The result revealed that test version II and version III were not significantly different. However, version I was significantly different from version II and version III (Table 5). The mean of three original tests with five-option items indicated that version I was the hardest version and version II and III were easier than version I. Also, skewness showed that the hardest test, version I, was closer to having a normal distribution than other two versions were (Table 4). Thus, I will first analyze data without considering the level of difficulty. Then, I will separately analyze the hardest test, version I, and version II and III, the easier versions, to explore the differences among the tests with the different numbers of options.

Table 4
Descriptive Statistics of Three Original Tests with Five Options (I5i, II5i, and III5i)

Test Version	n	Mean (Std. Error)	SD	Kolmogorov-Smirnov (Sig.)	Skewness
I5i	89	12.65 (.34)	3.21	.000	-0.74
II5i	86	14.31 (.32)	3.04	.000	-1.91
III5i	86	14.71 (.33)	3.07	.000	-2.01

Note. n = number of test takers. Maximum score =17.

Table 5
Significance in terms of the Average Scores in Five Original Tests (I5i, II5i, and III5i)

Version pair	*U*	*z*	*p*	*r*
I X II	2563.50	-4.10	.000*	-.31
II X III	3294.00	-1.62	.11	
III X I	2158.50	-5.03	.000*	-.38

Note. Bonferroni adjustment (*p* < .017).

Nine English Listening Prep-Tests for the CSAT

This research was designed to be robust across the different level of difficulty in the tests, because all participants took the difficult test once based on Latin-square combination that could contribute to the counterbalancing of different level of difficulty. Thus, without considering the level of difficulty, the differences in the average total score assignment among three-, four-, and five-option-item tests were investigated.

Average total score assignment.

As Kolmogorov-Smirnov statistics showed, the distribution was not normal (p < .05). Therefore, Friedman test, the nonparametric counterpart of one-way ANOVA for repeated measures, was used. As expected, the mean for the three-option-item tests was the highest and the mean for five-option-item tests was the lowest. The mean of four-option-item tests ranked between those of three- and five-option-item tests (Table 6). The result revealed a significant difference among three-, four-, and five-option-item tests, x^2 (2) = 11.40, p = .003 (Table 7). To detect the differences, Wilcox tests were adopted with the Bonferroni adjustment as .017 significance level. The tests reported that the three-option-item tests were significantly different from the five-option-item tests T = 7464.00, p = .000, r = -.16 (or z = -3.58, p = .000, r = -.16). There was no significant difference between the other pairs of item formats (Table 8).

Table 6

Descriptive Statistics of Three-, Four-, and Five-option-item tests: All Nine Tests

Test	n	Mean (Std. Error)	SD	Kolmogorov-Smirnov (Sig.)	Skewness
3-option (I3i/II3i/III3i)	264	14.48 (.17)	2.72	.000	-1.80
4-option (I4i/II4i/III4i)	264	14.03 (.18)	2.89	.000	-1.15
5-option (I5i/II5i/III5i)	264	13.88 (.20)	3.22	.000	-1.37

Note. n = number of test takers. Maximum score = 17.

Table 7

Significance in terms of the Average Total Scores in All Nine Tests

Test	Mean rank	Chi-square(x^2)	*df*	*p*
3-option (I3i/II3i/III3i)	2.15			
4-option (I4i/II4i/III4i)	1.97			
5-option (I5i/II5i/III5i)	1.89			
		11.40	2	.003*

p < .05.

Table 8

Paired Comparisons: Differences in terms of the Average Total Score Assignment in All Nine Test

Test pair	*T*	*z*	*p*	*r*
3opt. X 4opt.	9114.50	-2.35	.019	
4opt. X 5opt.	9664.00	- .71	.48	
5opt. X 3opt.	7464.00	-3.58	.000*	-.16

Note. Bonferroni adjustment (*p* < .017). Opt. = option.

Item facility and item discrimination.

In line with the expectation, the tests with five-option items were slightly more difficult than those with the three- and four-option items. The homogeneity of variances was verified. Therefore, the parametric test, one-way ANOVA was used to explore any significant difference. The results revealed that the average item facility was not significantly different across the three different formats (Table 12).

Aligned with results from the previous literature in testing, the average item discrimination in tests with five-option items was slightly higher than tests with the other two formats. The result of one-way ANOVA indicated the average item discrimination was not significantly different across the three different item formats (Table 13).

Correlation.

As the CSAT is a norm-referenced test, it is worthwhile investigating about the changes in test takers' ranks across the three different formats of tests. Spearman's correlation revealed positive correlations among three different item formats (Table 9). However, the correlation coefficients were rather low (.444 to .541), indicating the three different item formats construed exams that perhaps tapped into different underlying test constructs. This will be discussed further in the discussion section.

Table 9

Correlations of Test Scores among Three Different Item formats

		3-option	4-option	5-option
3-option	Spearman's rho	1	.444**	.541**
	Sig. (2-tailed)		.000	.000
4-option	Spearman's rho		1	.501**
	Sig. (2-tailed)			.000
5-option	Spearman's rho			1
	Sig. (2-tailed)			
n		264	264	264

Note. n = number of test takers. Correlation is significant at the 0.01 level (2-tailed).

Harder Version (version I) vs. Easier Versions (versions II &III)

Based on the preliminary analysis regarding the level of difficulty of three test versions, I separately analyzed version I (the hardest) from version II and III to investigate the differences among three-, four-, and five-option item formats in terms of the average total score assignment.

Average total score assignment: harder version (version I).

Kruskal –Wallis test was used due to the skewed data distribution. As shown in table 10, the three-option-item test (I3i) was the easiest while five-option-item test (I5i) was the hardest. The result revealed a significant difference among the three-, four-, and five-option-item test groups, $H(2) = 8.29$, $p = .016$. To explore significant differences, Mann-Whitney test was used with the adjustment of Bonferroni significance level as .017. Same as when nine tests were considered, the results indicated that only the three-option-item test was significantly different from the five-option-item test ($U = 2850.00$, $z = -2.94$, $p = .003$, $r = -.23$).

Table 10

Descriptive Statistics of Three-, Four-, and Five-option-item Tests: Harder Version Tests

Test	n	Mean (Std. Error)	SD	Kolmogorov-Smirnov (Sig.)	Skewness
3-option (I3i)	86	14.03 (.26)	2.43	.000	-1.01
4-option (I4i)	89	13.25 (.34)	3.24	.000	-. 95
5-option (I5i)	89	12.65 (.34)	3.20	.000	-. 74

Note. n = number of test takers.

Table 11

Paired Comparisons: Differences in terms of the Average Total Score Assignment in Harder Version Tests

Test pair	*U*	*z*	*p*	*r*
I3i X I4i	3408.50	-1.26	.21	
I4i X I5i	3447.00	-1.50	.13	
I5i X I3i	2851.00	-2.94	.003*	-.23

Note. Bonferroni adjustment (*p* < .017).

Item facility and item discrimination: harder version (version I).

The average item facility was as follows: the three-option item format (*M* = .83, *SD* = .12), the four-option item format (*M* = .78, *SD* = .12), and the five-option item format (*M* = .74, *SD* = .14). As expected, the tests with five-option items were slightly difficult than those with the three- and four-option items. The one-way ANOVA result indicated that the average item facility was not significantly different across the three different formats, *F*(2,48) = 1.78, *p* = .18 (see Table 12).

The average item discrimination in the three-option item format was *M* = .30 (*SD* = .20), in the four-option item format, *M* = .40 (*SD* = .17), and in the five-option-item test, *M* = .40 (*SD* = .17). The average item discrimination in tests with five-option items was slightly higher than tests with three-option items. However, the average item discrimination was not significantly different across the three different item formats, *F*(2,48) = 1.81, *p* = .18 (see Table 13).

Average total score assignment: easier versions (version II & III).

Aligned with the expectation, the mean for the thee-option-item format was *M* = 14.66 (*SD* = 2.85), that for the four-option-item format was *M* = 14.33 (*SD* = 2.62), and for five-option item format was *M* = 14.51 (*SD* = 3.05). However, different from the result with all nine tests and that with harder tests (version I), Kruskal–Wallis test indicated no significant difference across the three different item formats in terms of the average total score assignment.

Item facility and item discrimination: easier versions (version II & III).

The average item facility was as follows: the three-option item format (*M* = .86, *SD* = .08), the four-option item format (*M* = .85, *SD* = .10), and the five-option item formats (*M* = .85, *SD* = .10). The one-way ANOVA revealed that the average item facility was not significantly different across the three different item formats, *F*(2,48) = .09, *p* = .92, ω^2 = -.04 (see Table 12).

The average item discrimination in the three-option item format was *M* = .25 (*SD* = .12), in the four-option item format, *M* = .32 (*SD* = .17), and in the five-option item format, *M* = .33 (*SD* = .18). There was no significant different across the three different item formats, *F*(2,48) = 1.32, *p* = .28, ω^2 = .01 (see Table 13).

Table 12

Average Item Facilities: All Nine Tests, Harder Version (I), and Easier Versions (II & III) Tests

Test	N	Mean (Std. Error)	SD	*df*	*F*	*p*	ω^2
All 9 tests							
3-option	17	.85 (.02)	.07				
4-option	17	.83 (.03)	.08				
5-option	17	.82 (.02)	.09				
				2	.78	.47	.26
Harder version (I)							
3-option	17	.83 (.03)	.12				
4-option	17	.78 (.03)	.12				
5-option	17	.74 (.03)	.14				
				2	1.78	.18	.26
Easier versions (II & III)							
3-option	17	.86 (.02)	.08				
4-option	17	.85 (.02)	.10				
5-option	17	.85 (.03)	.10				
				2	.09	.92	-.04

p < .05.

Table 13
Average Item Discriminations: All Nine Tests, Harder Version (I), and Easier Versions (II & III) Tests

Test	N	Mean (Std. Error)	SD	*df*	*F*	*p*	*ω*^{2}
All 9 tests							
3-option	17	.31 (.03)	.12				
4-option	17	.35 (.04)	.14				
5-option	17	.38 (.04)	.15				
				2	1.17	.32	.06
Harder version (I)							
3-option	17	.30 (.05)	.20				
4-option	17	.40 (.04)	.17				
5-option	17	.40 (.04)	.17				
				2	1.81	.18	.03
Easier versions (II & III)							
3-option	17	.25 (.29)	.12				
4-option	17	.32 (.40)	.17				
5-option	17	.33 (.43)	.18				
				2	1.32	.28	.01

$p < .05$

Overall Test Reliability

The reliabilities of each listening test were examined through the Cronbach alpha coefficient as seen the table 14. The average reliability of the three-option-item tests was *M* = .71, in the four-option-item tests *M* = .77, and in the five-option-item tests *M* = .82, and the reliability increased as the number of options were added. However, the results from a one-way ANOVA revealed no significant difference among the reliabilities of the three different item formats.

Table 14
Overall Reliability

	Item format		
Version	3-option	4-option	5-option
I	.66	.79	.76
II	.85	.74	.85
II	.63	.78	.86
Overall	.71	.77	.82

Processing Time

A total of 37 participants recorded their processing time while taking a test: For three-option-item tests, 12 participants recorded their time, 13 for four-option-item-formats, and 12 for five-option-item tests. The average exam time including audio file playing time was 667.92 seconds for three-option-item tests, 696.00 seconds for four-option-item tests, and 736.08 seconds for five-option-item tests. The average processing time per item was 41 seconds. This indicated that the administration of the three-option-item tests could save 68.16 seconds, in which one or two more items (11.8% of the total items) might be added.

Survey Data

If participants took at least one of the three tests, they were asked to fill out a survey at the end of the study. In analyzing the survey data, missing values were deleted pairwise. Due to the absence on the third session day or the rejection to respond to survey questions, 35 were missing from 300 initial participants. As a result, 265 participants' data from the questionnaire were analyzed. Another Korean rater coded 10% of the data (randomly selected) to ensure the reliability of coding. The agreement between the raters on that 10% of the data was 87%. The disagreements were resolved through an in-depth discussion.

Closed-Ended and Likert-scale Items

Item one, three, and six.

Among 265 participants, 143 participants preferred the three-option-item test, 75 participants showed preference of the four-option-item test, and 45 respondents chose the five-option-item test (Figure 1).

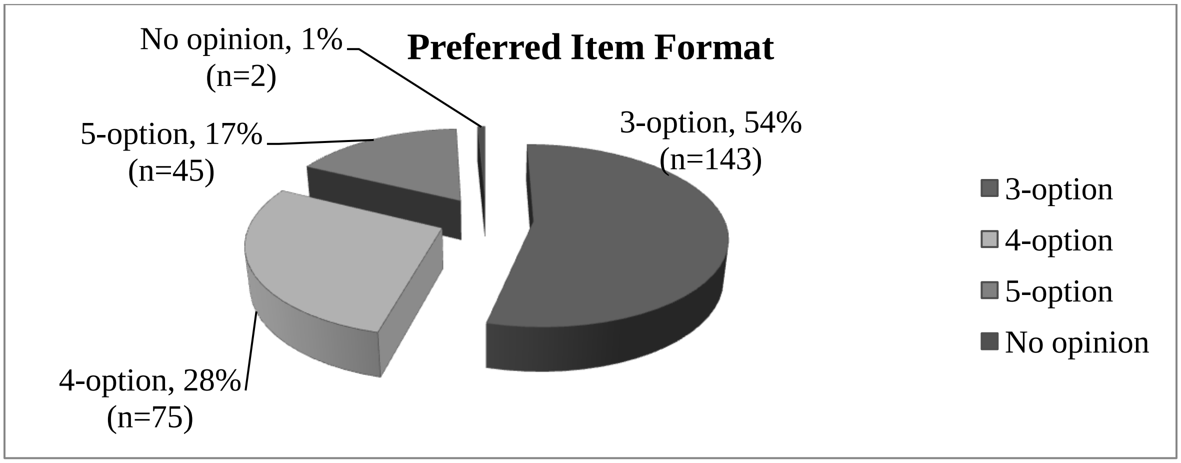


Figure 1. Preferred item format among three-, four-, and five-item tests

The Likert-scale question on the test takers' preference for a three-option-item test (item three on the questionnaire; one was "most dislike" and seven was "most like"), triangulated the result from the first item. Spearman's correlation revealed a significant negative correlation between the preferred number of options and the Likert-scale points, r = -.61, p < .001, R^2 = .37: The group preferring the three-option-item test showed the highest average points (M = 5.94, *SD*

= 1.28), the average points of those preferring the four-option item format was in-between (*M* = 4.36, *SD* = 1.11), and those preferring the five-option-item test had the lowest average points in the Likert-scale (*M* = 3.60 , *SD* = 1.51).

Regarding question six—whether test takers agree or disagree that an administration of the CSAT with a three-option item format would be fair—164 respondent agreed and 86 respondents disagreed (Figure 2). A two-variable, Pearson's chi-square test was used to explore the relationship between the preferred number of options and the test takers' agreements or disagreements with the three-option-item administration on the CSAT. The test indicated that there was a significant association between the preferred number of options and the agreement with the administration of the three-option item format, X^2 (2) = 52.71, p < .001, Φ = .46. Among the respondents preferring three options, 85% agreed with the administration of the three-option items on the CSAT. Among the respondents preferring the four-option item format, 47% agreed with it. Only 34% agreed among the respondents preferring five options (Figure 3, Table 15).

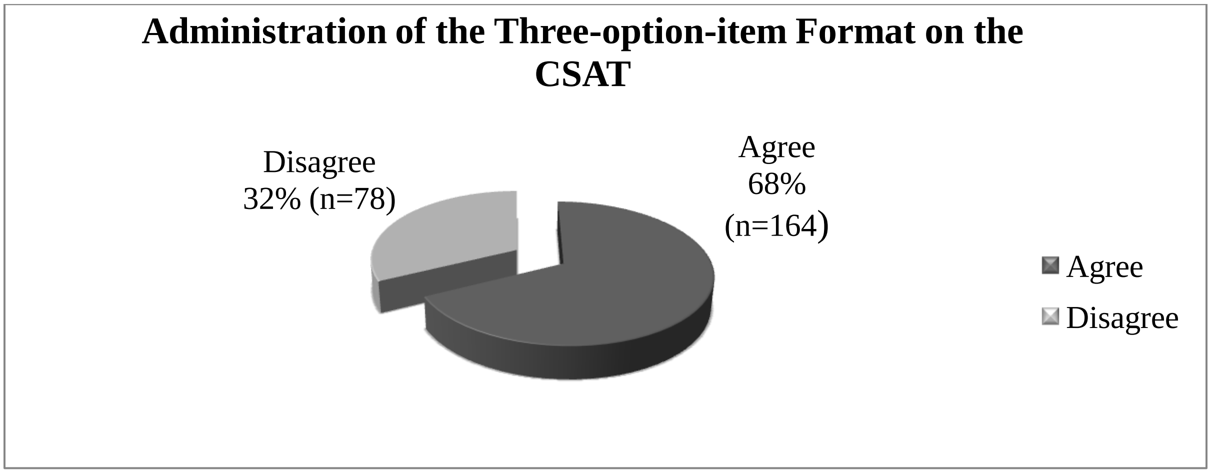


Figure 2. Administration of the three-option-item format on the CSAT

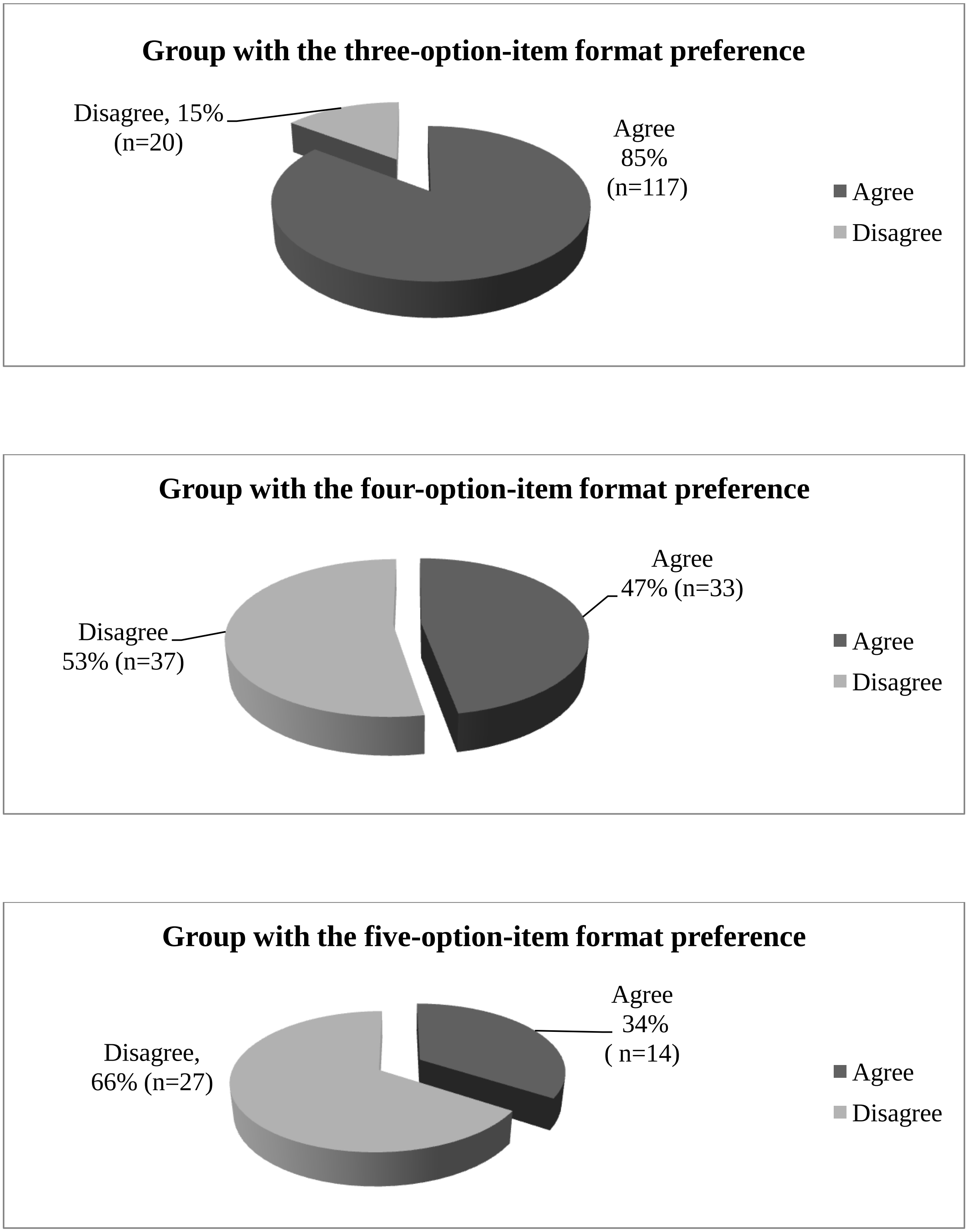


Figure 3. Agreement with the administration of the three-option-item format based on the preferred number of options

Table 15

Agreement vs. Disagreement based on the Preferred Number of Options

Preferred options	(n)	Agree (n)	Disagree (n)
3 options	(137)	117	20
4 options	(70)	33	37
5 options	(41)	14	27
Total	(248)	164	84

Note. n = number of respondents.

Open-ended Items

Item two: why do you prefer the three-, four-, or five-option-item test?

Among the 143 respondents who preferred a three-option-item test, 48.3% liked the item format because it was easy to choose a correct answer. A second group of 12.6% responded that they liked three options due to time saving and the user-friendly format it presents. Among 75 respondents preferring the four-option-item test, 61.8% chose this format because they thought that four would be the proper number of options and 10.7% liked it because they were used to the four-option-item test. Among 45 respondents, the five-option preferring group, 62.2% said that they were used the five-option-item test and 15.5% chose it because the five-option-item test would, they said (in their own words), increase item discrimination (Table 16).

Table 16

Item Two: Why do you prefer a three-, four-, or five-option-item test?

3 options		4 options		5 options	
Reason	n	Reason	n	Reason	n
- I can easily choose an answer.	70	- I think that the number of options is proper.	46	- It is a familiar format because I have been taking the five-option items.	28
- I can save time.	18	- It is a familiar format because I have been taking the four-option items.	8	- I think that five-option items will have better item discrimination.	7
- It is a user-friendly format	18	- I can easily choose an answer.	4	- I can easily choose an answer.	1
- I think that the number of options is proper.	9	- I feel less anxiety with three-option items.	4	- I think that the number of options is proper.	1
- I feel less anxiety with three-option items.	7	- I think that four-option items will have better item discrimination.	4		
I can concentrate more on the test.	1				
- I do not think there is a difference among three-, four-, and five-option items.					11

Note. n = number of respondents.

Item four: the advantage and disadvantage of the three-option-item tests.

The first advantages of the three-option item format was the higher possibility of correct answers (N=119) and the second was time saving (N=35) among 263 respondents. The disadvantage of the three-option item format was the lower item discrimination (N=27). Also, the respondents (N=22) were concerned that the level of difficulty would be increased due to the reduced number of options. Interestingly, among those preferring the three-option item format (N=143), 57.3% (N=82) of the participants did not respond with anything as a disadvantage of the three-option item format (Table 17).

Table 17

Item Four: Advantages and Disadvantages of the Three-option-item Tests

Advantage		Disadvantage	
Reason	n	Reason	n
- It has the higher possibility to choose correct answer than four- or five-option items (e.g., I can easily choose a correct answer. A correct answer can be easily guessed due to reduced number of options)	35	- Item discrimination will be decreased.	27
- I can save time.	10	- Item facility will be decreased. (Items will be more difficult.)	22
- It is a user-friendly format.	7	- There is the higher possibility to have a correct answer by guessing because of reduced number of options.	7
- I can concentrate more on the test.	2	- Test validity will be decreased. (e.g. Face validity will be decreased. The test with three-option items does not seem to screen out applicants for universities.)	3
- I feel less anxiety.	2	- The test format (three-option items format) is so unfamiliar.	2
- Item discrimination and test validity will be increased.	1		

Note. n = number of respondents.

Item seven: why do you agree or disagree with the administration of the three-option item tests?

Among those who agreed with the administration of the three-option item format on the CSAT (N=164), 34.8% (N=57) stated the higher possibility of correct answers as the reason why they agreed and the time-saving (N=22) was the next reason. Respondents who disagreed (N=86) pointed the lower item discrimination as the main reason (N=36). Also, 16 respondents disagreed because they thought the items on the CSAT would be more difficult if the number of options were to be reduced (Table 18).

Table 18
Item Seven: Why do you agree or disagree with the administration of the three-option-item tests?

Agree		Disagree	
Reason	n	Reason	n
- It has the higher possibility to choose correct answer than four- or five-option items (e.g. I can easily choose a correct answer. A correct answer can be easily guessed due to reduced number of options)	57	- Item discrimination will be decreased.	36
- I can save time.	22	- Item facility will be decreased. (Items will be more difficult.)	16
- I do not think there is a difference among three-, four-, and five-option items.	17	- The test format (three-option items format) is so unfamiliar.	14
- It is a user-friendly format.	13	- There is a higher possibility to have a correct answer by guessing because of reduced number of options.	3
- I feel less anxiety.	7	- Test validity will be decreased. (e.g., Face validity will be decreased. The test with three-option items does not seem to screen out applicants for universities.)	3
- Item discrimination and test validity will be increased.	4	- I do not think there is a difference among three-, four-, and five-option items.	3
- I can concentrate more on the test.	3		

Note. n = number of respondents.

Results

In the following section, I sum up the results based on the research questions.

(1) In terms of the average total score assignment, average item facility, average item discrimination and overall reliability, are there significant differences among three-, four-, and five-option multiple-choice item formats on a high-stakes English listening test?

With respect to the average total score assignments in all nine tests, the significance was revealed between three-option-item tests and five-option-item tests (T = 7464.00, p = .000, r = -.16 or z = -3.58, p = .000, r = -.16 with Bonferroni adjustment p < .017). When the hardest tests (version I) were separately investigated from versions II and III, Mann-Whitney test revealed the significant difference between the three-option-item tests and the five-option-item tests (U = 2850.00, z = -2.94, p = .003, r = -.23 with Bonferroni adjustment p < .017), like the result when all nine tests were included. When the easier tests (version II and version III), which were equal of difficulty, were analyzed, no significant difference was indicated across the three-, four-, and five-option-item tests.

In terms of the average item facility, average item discrimination, and overall test reliability, a one-way ANOVA reported no significant difference, when considering all nine tests and the level of difficulty. Additionally, Spearman's correlation revealed positive correlations among three different item formats. However, the correlation coefficients were rather low (.444 to .541), which could mean that the ranks of test takers were a bit changed across three different formats of tests.

(2) Can a listening test with three-option items function as selectively as those with five-option items in screening applicants for university entrance?

Considering the normality of the data distribution, the five-option-item tests were closer to the normal distribution than other two test formats: The mean for the three-option-item tests was *M* = 14.48 (*SD* = 2.72), for the four-option-item tests was *M* = 14.03 (*SD* = 2.89), and for the five-option-item tests was *M* = 13.88 (*SD* = 3.22). From the perspective of the level of difficulty in a test, the hardest version was slightly closer to the normal distribution: The mean for version I was *M* = 12.65 (*SD* = 3.20), for version II was *M* = 14.31(*SD* = 3.04), and for version III was *M* = 14.71, (*SD* = 3.06). However, as indicated by the average item discrimination, there was no significant difference among the item formats in terms of how they screen examinees. That is, the listening tests with three-option items could function selectively as four- and five-option-item tests in screening out applicants.

(3) Can a three-option multiple-choice item format significantly reduce the exam time?

The three-option-item format could reduce the total exam time by 10.8%, compared with the time for five-option items (From 736.08 second to 667.92 seconds). Considering the average processing time per item was 41 seconds, the reduction in time by 10.8% could allow for the possibility of adding one or two more items (11.8% of total 17 items in the CSAT listening test), which may increase the validity and reliability of the test.

(4) Do test takers prefer a three-option-item format on a high-stakes test?

As the survey data indicated, 54% of the respondents preferred the three-option-item test. Also, 68% of the respondents agreed that the administration of the three-option-item test on the CSAT would be fair. In exploring the relationship between the preferred number of options and the agreement or disagreement with the administration of the three-option items on the CSAT, results from a two-variable, Pearson's chi-square test indicated that there was a significant association between the preferred number of options and the agreement with the administration of the three-option-item format ($X^2(2) = 52.71$, $p = .000$, $\Phi = .46$): The group that prefer the three-option-item format more tends to agree with the administration of the three-option-item format on the CSAT than those preferring the four- or five-option-item format. However, interestingly, 6.5% (N=22) of the participants among those who preferred the three-option-item format disagreed with the administration of three-option items on the CSAT. The main reason for their disagreement was a perceived increased possibility of guessing, which they indicated would unfairly benefit those who did not study as hard as they did.

Discussion

This study found that in terms of the average total score assignment, the five-option-item tests could spread out examinees' scores more closely along a normal distribution than the three-option-item tests. However, the average item facility, average item discrimination, and overall reliability did not indicate any significant differences among the three different item formats. Thus, results from this study support previous research (e.g., Costin, 1972; Delgado & Prieto, 1998; Green, Sax, & Michael, 1982; Owen & Froman, 1987; Sidick, Barrett, & Doverspike, 1994; Trevisan, Sax, & Michael, 1991; Trevisan, Sax, & Michael, 1994), stating that tests with different numbers of options do not indicate any statistically significant differences.

However, in this study I also ran correlations among the three different item formats of tests, which previous studies on the optimal number of options did not investigate before. While Spearman's correlations were positive and statistically significant across the three different item formats of the tests, the strength of the correlations were slightly lower than what I expected. The result from the correlation suggests that factors external to the listening construct might be differentially tapped into depending on the number of options presented.

Regarding the correlation result, I speculate that testwiseness strategies were an additional construct that affected test takers' scores on the five-option-item format tests while testwiseness strategies probably affected the scores least when three-option items were presented. Rogers and Harley (1999) also reported that items with fewer options were less affected by testwiseness. According to them, 13 items were susceptible to testwiseness in their four-option-item test while only four items were testwise susceptible in their three-option-item test. My speculation can be explained by the survey data in the present study: Some of participants reported that they felt the three-option-item test was more difficult than the five-option-item test because they could not delete any non-plausible options before listening to audio files as they usually did on listening tests with more options. Also, test anxiety might be another factor affecting the scores. According to the survey data, some test takers (n=7) had less anxiety with a three-option-item format whereas there were test takers (n=8) who felt less anxiety with the four-option-item format, over the three-option item format. Additionally, test takers' familiarity with the five-option-item format possibly affected the scores: Most of test takers were trained with five-option-item formats since all official middle/high school tests in Korea are administered with the five-option-item format. (Test takers were quite surprised when they were given the three-option-item test. They told me that the test seemed to be so awkward, like it was missing something.) Finally, the cognitive reading load might be considered as one of the factors: the five-option-item tests required reading options faster than the three-option-item tests did. I will discuss in more depth the possible effects of testwiseness strategies and the cognitive reading load on outcomes and perceptions later in this section.

In a high-stakes test, item discrimination is critical for screening and segregating test takers. In the CSAT, test developers increased the number of options (from four to five options) and adopted a weighted point system to increase item discrimination in 1993. However, with respect to efficiency in developing items, this study suggests that the three-option-item format is worth considering even for high-stakes tests like the CSAT. This also has been supported by the claims that most multiple-choice items have only one or two functional distracters (Haladyna, Downing, & Rodriguez, 2002) and the increased number of options are not related to better testing statistics, rather the extra options provide unintended clues to other items (Rodriguez, 2005). A three-option-item format can contribute not only to saving time but also to covering more content, if additional items are inserted (Rodriguez, 2005). This results in accurate testing outcomes with increased test reliability (Rodriguez, 2005; Shizuka et al., 2006).

From the correlation analysis and the survey data, this study also suggests that the three-option-item format can function well to prevent the need for testwiseness strategies, as Rogers and Harley (1999) claimed. When asked about disadvantages of the three-option-item format, students (n=22) stated that the three-option-item format would make a test more difficult. When all distracters are equally plausible, the item itself is more difficult—or so it appears to be—because there are no obvious distracters that can be immediately eliminated. Therefore, some stated they felt like the five-option items were better, in that they provided test takers with an initial satisfaction of being able to immediately recognize non-plausible distracters and eliminate them: This nice, psychologically pleasing effect (being able to easily eliminate distracters) was missing from the three-option-item format. This relates to the concept of "testwiseness" (Rogers & Harley, 1999)—these students have been highly trained in test taking, and one of their strategies, most likely, is to read quickly through the options and eliminate non-plausible distracters, even before listening to an audio file of a listening item. Some student who disagreed with the three-option-item format on the CSAT said that the test items with three options made them so confused: All options seemed to be correct answers. They stated that they could not use the satisfying strategy to increase the possibility of a correct answer by deleting two or more options before audio files play.

When the five-option-item format in a listening test is used to make tests more difficult—based on the expectation that test takers will get fewer correct answers and, therefore, a much wider spread of scores is obtained—test validity needs to be considered. Buck (2001) claimed that the factors affecting the level of difficulty in a listening test are “linguistic characteristics, organization, familiarity and explicitness” of audio content (p. 150). Thus, more options do not equate with a better measure of listening skills. Rather, adding more options to items presents test takers with an additional reading cognitive burden: Test takers are forced to read options as fast as possible within the limited time and choose a correct answer. Therefore, instead of pouring efforts into writing more options per item, spending the same amount of time to develop more three-option items that are carefully constructed for measuring listening ability may contribute to the enhancement of content and construct validity.

In alignment with the study by Owen and Froman (1987), this research found that the three-option-item format could save 10.8% of the total exam time, which enables to increase 11.8% of total number of items on the CSAT English listening test. Therefore, better test reliability can be attained through the added items, because longer tests (with more items) have better reliability than shorter tests: With five options, more time is needed to answer the questions, so the test must necessarily have fewer items, which limits reliability.

Even though prior literature has stated that blind guessing should not be an issue (Costin, 1976; Haladyna, 2004; Kolstad, Briggs, & Kolstad, 1985), in the context of this study, it may be a critical part to be considered: According to Ebel (1965), blind guessing is “selecting the answer at random without considering the content of the options” (p. 715). In a time-limited test, as Rodriguez (2005) stated, blind guessing can threaten test validity. This claim is applicable to a lower-ability group of examinees on the CSAT English listening test. By blind guessing, lower-ability test takers can get 33.3% of the answers correct on a three-option-item test, whereas those taking a five-option-item test have only a 20% chance of correct answers. Considering the unique situation in Korea—the CSAT is a very high-stakes test that can determine the future social status of each student in Korea (Choi, 2008). For instance, during the listening test, domestic and international flights are not allowed to take off and land, as the noise may affect the test results (*Asia times*, 2005)—the 13.3% score difference is most likely to be viewed as extremely detrimental to test reliability. Thus, the effect of blind guessing and the number of items in a high-stakes test needs to be examined robustly.

In sum, the results of the present study, supporting those from the testing literature, suggest that the three-option-item format contributes to developing test items more efficiently, enhancing test validity and reliability, and preventing testwiseness strategies.

Conclusion

This study suggests that the three-option-format would be worthy of consideration in a high-stakes test. However, context dependent factors need to be considered to answer the argument of how many options are optimal that has been debated for decades.

Statistically, the three-option-item format can be recommended in developing test items more efficiently while screening out test takers selectively, when the results in this study are considered—no significant differences across the three different item formats in terms of the average item facility, average item discrimination, and overall reliability. Also, the three-option-item format can increase the test reliability by adding more items due to the saved exam time. Additionally, from the cognitive perspective, the three-option-item format in the CSAT listening test can enhance test validity (content and construct validity) by reducing reading loads in a limited time—when the listening test is supposed to measure mostly listening skills, not combined with cognitive skills.

Emotionally, the three-option-item format is strongly preferred among test takers because fewer options make it easier for them to choose correct the answers. Interestingly, among the test takers with testwiseness strategies, the five-option-item format is preferred to the three-option-item format because testwiseness strategies to remove non-plausible options can be easily applied. Thus, the three-option-item format could be recommended to prevent the need for testwiseness strategies.

In addition, based on the correlation analysis that the tests with different item formats did not correlate highly, it is possible that the inclusion of more options could potentially increase variability in the underlying construct being measured. That is, more options could force the need for testwiseness strategies and other cognitive capacities irrelevant to listening skills. Thus, it is also essential to consider what needs to be measured by the listening test: mostly listening skills alone or listening skills combined with other academic and cognitive capacities that are perhaps irrelevant to the underlying listening skill construct.

However, contextually, test developers of the CSAT may be reluctant to adopt the three-option-item format for two reasons: (1) The CSAT is such a high-stakes test that the need for even non-statistical differences—very minute, non-statistically significant differences in scores—may be still needed for very finite discrimination among test takers; and (2) The effect of blind guessing in a time-limited test is extremely detrimental to reliability. Considering Koreans' unique obsession with the CSAT scores, it might be hard to suggest to the administration of the CSAT to adopt and employ a three-option-item format. Therefore, there may be no right answer to the optimal number of options. Test developers and researchers will have to consider the statistical, cognitive, emotional, and contextual factors when determining the optimal number of options for their test.

Limitations and Future Research

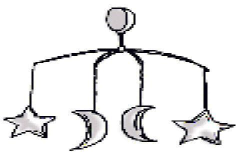
When it comes to the limitations in the present study, I adopted the previously-administered CSAT-prep tests that students can download via the Internet. This may have the possibility that participants were exposed to the original three versions of the five-option-item tests. Thus future research with new tests would be worthwhile. In addition, the effect of the CSAT's weighted point system on item discrimination was not investigated here. Therefore it is worthy to consider the extent to which a weighted points (assigned based on the pre-expected item facility) could affect item discrimination and overall test reliability. Finally, the effect of blind guessing and the number of items, which was not investigated in the present study, would be an interesting topic for future research.

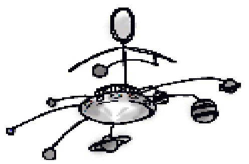
Appendices

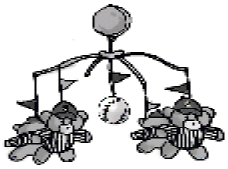
Appendix A-1

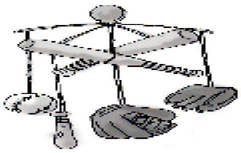
Test Version: I5i

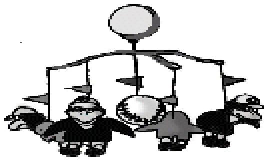
1. 대화를 듣고, 남자가 구입할 물건을 고르시오.

①


②


③


④


⑤


2. 대화를 듣고, 남자의 심경으로 가장 적절한 것을 고르시오.

① bored ② relieved ③ ashamed

④ frustrated ⑤ impressed

3. 대화를 듣고, 두 사람이 대화하고 있는 장소로 가장 적절한 곳을 고르시오.

① park ② hotel ③ library

④ museum ⑤ cafeteria

4. 대화를 듣고, Jenny 가 남자에게 부탁한 일을 고르시오.

① 새로운 소식 알려주기 ② 친구 대신 전화 걸어주기

③ Jenny 의 남자친구 만나보기　　　　④ 전화를 끊지 않고 기다리기

⑤ David 의 남동생 소개시켜 주기

5. 대화를 듣고, 두 사람의 관계를 가장 잘 나타낸 것을 고르시오.

① coach – athlete　　② Father – daughter　　③ teacher – student

④librarian-student　　⑤ salesclerk - customer

6. 대화를 듣고, 남자가 할 일로 가장 적절한 것을 고르시오.

① to see a doctor　　② to buy some tea　　③ to drop by a drugstore

④ to go to Carla’s home　　⑤ to visit a herbal garden

7. 다음을 듣고, 남자가 하는 말의 목적으로 가장 적절한 것을 고르시오.

① 적절한 칫솔 선택법을 알리려고　　② 치의학의 발전을 소개하려고

③ 양치질의 중요성을 알리려고　　④ 칫솔질하는 방법을 소개하려고

⑤ 전동칫솔의 위험성을 경고하려고

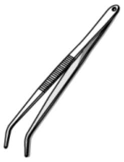
8. 대화를 듣고, 남자가 지불한 금액을 고르시오.

① ₩6,000　　② ₩6,500　　③ ₩7,000

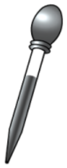
④ ₩7,500　　⑤ ₩8,000

9. 다음을 듣고, 여자가 설명하고 있는 실험도구를 고르시오.

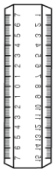
①



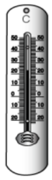
②



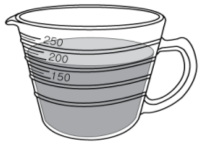
③



④



⑤

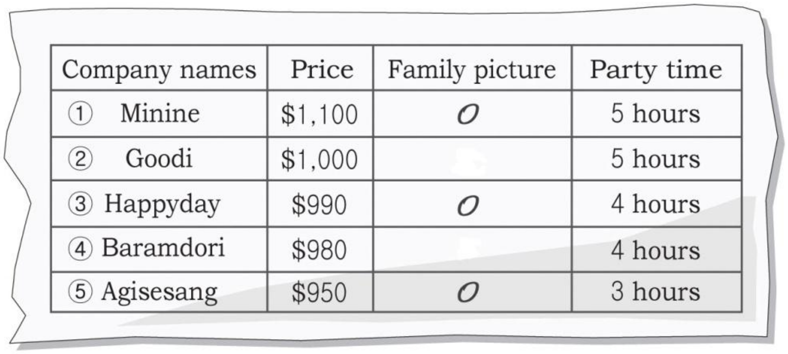


10. 대화를 듣고, 남자가 여자를 위해 할 일로 가장 적절한 것을 고르시오.

① 학교 안내하기 ② 일자리 구해주기 ③ 아파트 청소하기

④ 컴퓨터 구입하기 ⑤ 부동산 중개업소 들르기

11. 다음 자료를 보면서 대화를 듣고, 여자가 선택할 회사를 고르시오.



Company names	Price	Family picture	Party time
① Minine	$1,100	O	5 hours
② Goodi	$1,000		5 hours
③ Happyday	$990	O	4 hours
④ Baramdori	$980		4 hours
⑤ Agisesang	$950	O	3 hours

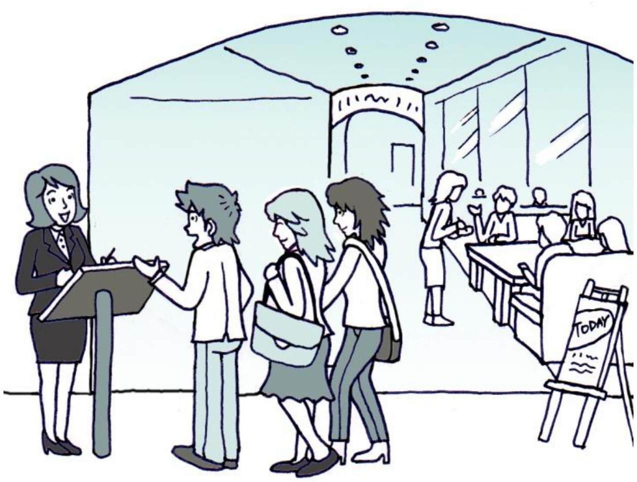
12. 다음을 듣고, 방송에서 언급한 내용을 고르시오. [3 점]

① 오존층 파괴의 새로운 원인 발견 ② 콩고의 고릴라 출산율 증가

③ 영국 남부 지역의 식량 생산량 감소 ④ 영국의 홍수 피해 복구 완료

⑤ 인도 내 휴대전화 이용자 증가

13. 다음 그림의 상황에 가장 적절한 대화를 고르시오. [1 점]



①　　　　②　　　　③　　　　④　　　　⑤

14. 대화를 듣고, 남자의 마지막 말에 대한 여자의 응답으로 가장 적절한 것을 고르시오.

Woman: ____

① Don’t worry. No news is good news.

② You know, hunger is the best sauce.

③ Cheer up! Every dog has his day.

④ Well, strike while the iron is hot.

⑤ Come on! Haste makes waste.

15. 대화를 듣고, 여자의 마지막 말에 대한 남자의 응답으로 가장 적절한 것을 고르시오.

Man: ____

① Thanks. How nice you are!

② Turn right at the second corner.

③ What a fancy restaurant this is!

④ I am going to reserve two seats.

⑤ I'm sorry that you made a mistake again.

16. 대화를 듣고, 남자의 마지막 말에 대한 여자의 응답으로 가장 적절한 것을 고르시오.

Woman: ____

① Keeping promise is important.

② I haven't visited my elementary school.

③ What do you say to meeting her tonight?

④ We email each other several times a week.

⑤ We've seen each other somewhere before, right?

17. 다음 상황을 듣고, Jim 이 할 말로 가장 적절한 것을 고르시오.

Jim: Excuse me, ____

① are you in line?

② show me your ticket.

③ where is the entrance?

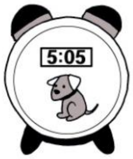
④ two tickets for 2:30, please.

⑤ could you tell me which is better?

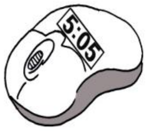
Appendix A-2

Test Version: II5i

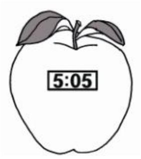
1. 대화를 듣고, 남자가 구입할 물건을 고르시오. [1 점]



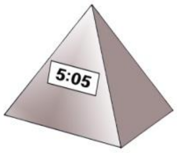
①



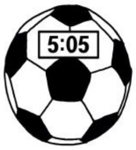
②



③



④



⑤

2. 대화를 듣고, 여자의 심정으로 가장 적절한 것을 고르시오.

① bored ② angry ③ excited ④ scared ⑤ jealous

3. 다음을 듣고, 무엇에 관한 설명인지 고르시오.

① ruler ② scale ③ camera ④ whistle ⑤ stopwatch

4. 대화를 듣고, 남자가 할 일로 가장 적절한 것을 고르시오.

① 과제 제출하기 ② 만화책 반납하기 ③ 부모님께 전화 드리기

④ 선생님께 사과 드리기 ⑤ 친구의 생일 선물 사기

5. 대화를 듣고, 남자가 지불할 금액을 고르시오.

① $8 ② $10 ③ $16 ④ $20 ⑤ $30

6. 다음을 듣고, 여자가 하는 말의 목적으로 가장 적절한 것을 고르시오.

① 상담 교사를 소개하려고 ② 온라인 상담을 권장하려고

③ 식당 이용 시간을 알리려고 ④ 상담 신청 방법을 안내하려고

⑤ 수강 신청 기간을 공지하려고

7. 대화를 듣고, 여자가 남자에게 부탁한 일로 가장 적절한 것을 고르시오.

① to buy coffee for her ② to treat her to dinner

③ to clean the refrigerator ④ to wake her up in the morning

⑤ to study together for an exam

8. 대화를 듣고, 두 사람의 관계를 가장 잘 나타낸 것을 고르시오.

① 은행 경비원 - 고객 ② 교통 경찰관 – 운전자 ③ 우편배달부 - 거주자

④ 기숙사 관리인 - 학생 ⑤ 부동산 중개인 - 세입자

9. 대화를 듣고, 두 사람이 대화하고 있는 장소로 가장 적절한 곳을 고르시오.

① in a car ② in a subway ③ in a theater

④ in an elevator ⑤ in a parking lot

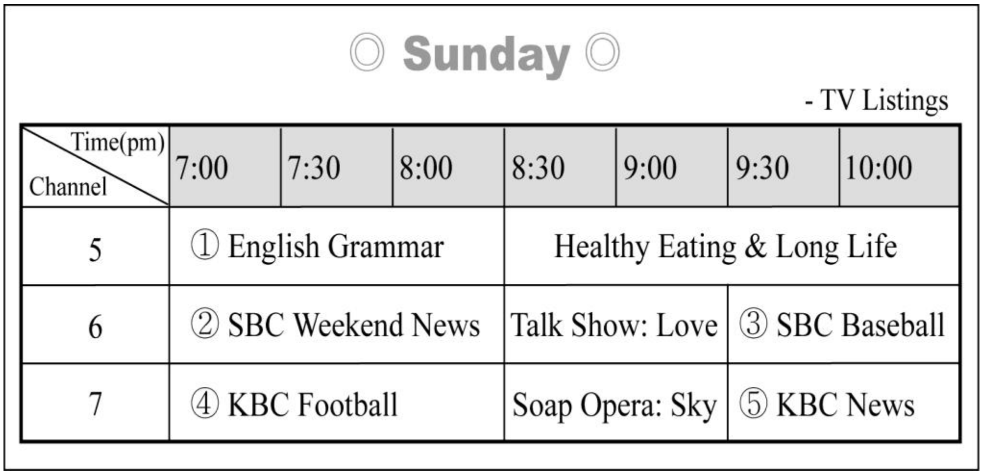
10. 대화를 듣고, 여자가 남자를 위해 할 일로 가장 적절한 것을 고르시오.

① CD 구입해 주기　　② 보고서 작성 도와주기

③ 노트북 컴퓨터 빌려주기　　④ 컴퓨터 프로그램 설치해 주기

⑤ CD 에 자료 저장 방법 가르쳐주기

11. 다음 TV 편성표를 보면서 대화를 듣고, 남자가 시청하게 될 프로그램을 고르시오. [3 점]



◎ Sunday ◎

- TV Listings

Time(pm) / Channel	7:00	7:30	8:00	8:30	9:00	9:30	10:00
5	① English Grammar			Healthy Eating & Long Life			
6	② SBC Weekend News			Talk Show: Love		③ SBC Baseball	
7	④ KBC Football			Soap Opera: Sky		⑤ KBC News	

12. Gloria Hotel 에 관한 다음 내용을 듣고, 일치하지 않는 것을 고르시오.

① 시내 중심에 위치해 있다.

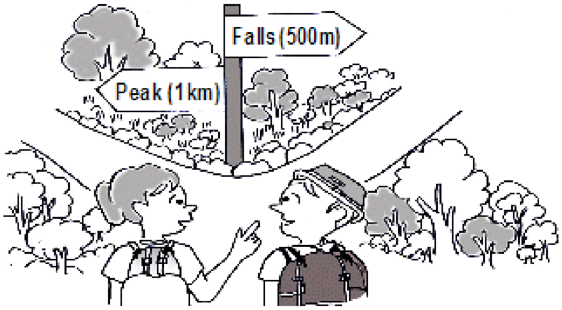
② 전망 좋은 방이 구비되어 있다.

③ 디럭스 룸에는 새 가구가 비치되어 있다.

④스낵바에서는 음료수를 무료로 제공한다.

⑤ 실내 수영장은 밤 10 시까지 개방한다.

13. 다음 그림의 상황에 가장 적절한 대화를 고르시오.



① ② ③ ④ ⑤

14. 대화를 듣고, 여자의 마지막 말에 대한 남자의 응답으로 가장 적절한 것을 고르시오.

Man:____

① I hope no one was hurt.

② Really? I’ll try it right now.

③ I don’t like to eat cucumber.

④ Then, apply this cream to your face.

⑤ You shouldn’t scrub your face with a sponge.

15. 대화를 듣고, 남자의 마지막 말에 대한 여자의 응답으로 가장 적절한 것을 고르시오.

Woman:____

① Yes. If it doesn't rain, it will be serious.

② Right. I hope that the rain will stop soon.

③ I agree. These plants grow well without rain.

④ Don't worry. We have enough rainfall this year.

⑤ Sounds good. I enjoy the sunny days these days.

16. 대화를 듣고, 여자의 마지막 말에 대한 남자의 응답으로 가장 적절한 것을 고르시오.

Man: ____

① Right, I had better see a doctor now.

② Thanks, but you'd better take it easy.

③ Wow, that's wonderful news for me.

④ Well, don't overdo it from the beginning.

⑤ Really? It's been a week since I started dieting.

17. 다음 상황 설명을 듣고, Judy가 아버지에게 할 말로 가장 적절한 것을 고르시오.

Judy: ____

① What scholarship am I going to win?

② Why don’t you get advice from mother?

③ How can I get admitted to the university?

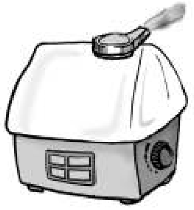
④ You look a bit down today. What’s the matter?

⑤ Which university do you think is better for me?

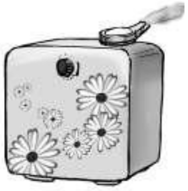
Appendix A-3

Test Version: III5i

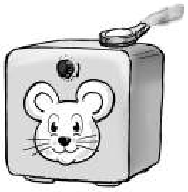
1. 대화를 듣고, 남자가 구입할 가습기를 고르시오. [1점]



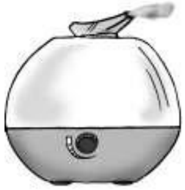
①



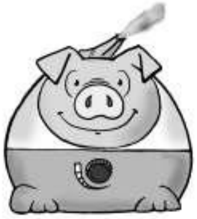
②



③



④



⑤

2. 대화를 듣고, 여자의 심경 변화로 가장 적절한 것을 고르시오.

① upset → relieved ② nervous → disappointed ③ amused → frightened

④ indifferent → interested ⑤ thankful → dissatisfied

3. 다음을 듣고, 무엇에 관한 설명인지 고르시오.

① 전광판 ② 신호등 ③ 가로등 ④ 횡단보도 ⑤ 도로 표지판

4. 대화를 듣고, 여자가 남자를 위해 할 일로 가장 적절한 것을 고르시오.

① 옷 수선하기 ② 셔츠 찾아오기 ③ 친구 마중 나가기

④ 세탁물 맡기기　　　　　⑤ 출장 일정 조정하기

5. 대화를 듣고, 남자가 여자에게 지불할 금액을 고르시오.

① $20　　② $60　　③ $80　　④ $100　　⑤ $120

6. 다음을 듣고, 여자가 하는 말의 목적으로 가장 적절한 것을 고르시오.

① 자연보호 활동 참여를 권장하려고

② 동굴 탐사 장비 사용방법을 설명하려고

③ 비상상황 시 응급조치 방법을 소개하려고

④ 동굴 탐사를 위한 준비 사항을 안내하려고

⑤ 비상용품 구입 시 고려할 사항을 알려주려고

7. 대화를 듣고, 여자가 남자에게 부탁한 일을 고르시오.

① to print out pamphlets　　　② to schedule the events

③ to make the guest list　　　④to set the festival stage

⑤ to design the invitation card

8. 대화를 듣고, 두 사람의 관계를 가장 잘 나타낸 것을 고르시오.

① 승무원 - 승객　　② 수리공 – 고객　　③ 관리인 - 입주민

④ 안내원 – 관람객　　⑤ 모델 - 사진 작가

9. 대화를 듣고, 두 사람이 대화하고 있는 장소를 고르시오.

① airport ② art gallery ③ flower shop

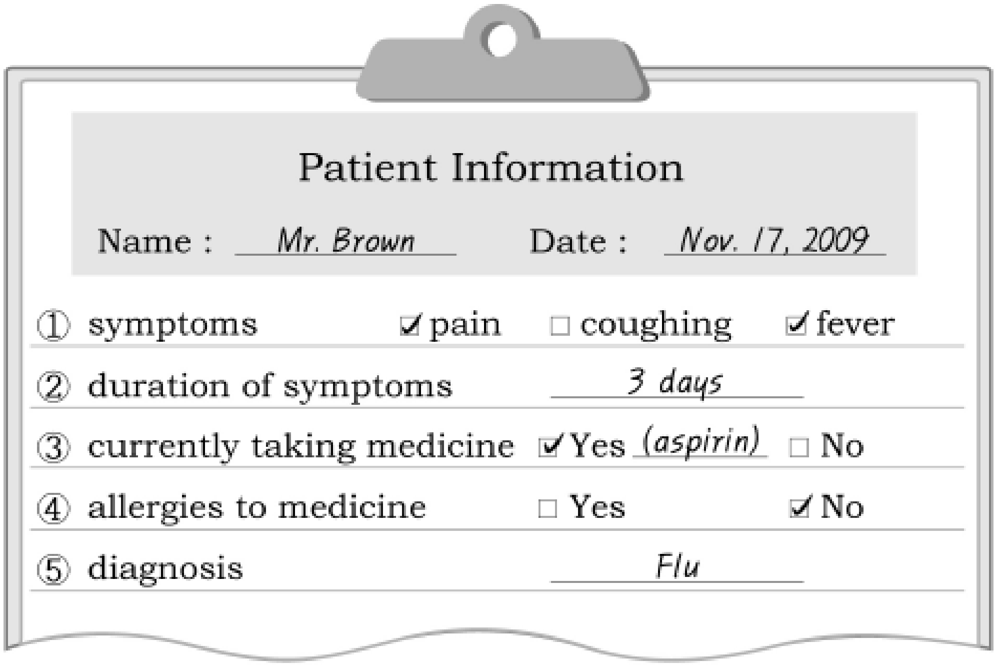
④ restaurant ⑤ amusement park

10. 대화를 듣고, 남자가 할 일로 가장 적절한 것을 고르시오.

① 방 청소하기 ② 숙소 예약하기 ③ 방학 숙제하기

④ 휴가 계획 세우기 ⑤ MP3 플레이어 구입하기

11. 차트를 보면서 대화를 듣고, 대화의 내용과 일치하지 않는 것을 고르시오. [3점]



Patient Information

Name : Mr. Brown Date : Nov. 17, 2009

① symptoms	☑ pain	□ coughing	☑ fever
② duration of symptoms	3 days		
③ currently taking medicine	☑ Yes (aspirin)		□ No
④ allergies to medicine	□ Yes		☑ No
⑤ diagnosis	Flu		

12. Pets & People에 관한 다음 내용을 듣고, **일치하지 않는 것을** 고르시오.

① 위험에 처한 애완동물을 구호하는 단체이다.

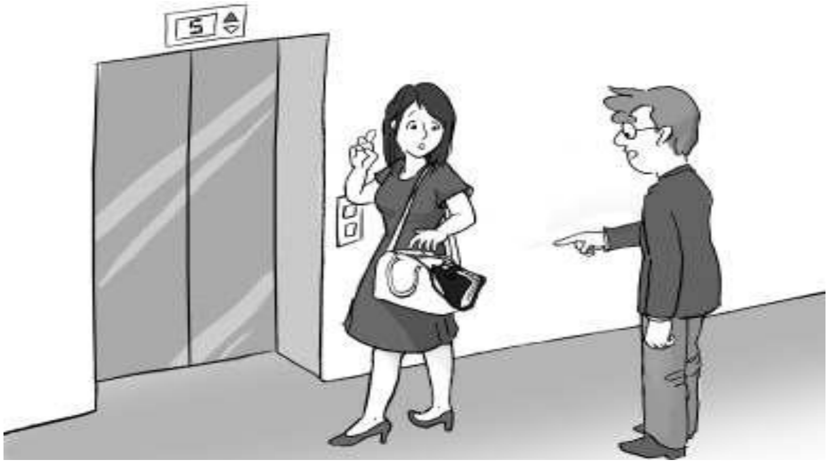
② 집 없는 개와 고양이에게 새 주인을 찾아준다.

③ 회원이 되려는 사람으로부터 가입비를 받는다.

④ 개와 고양이의 입양비를 다르게 받는다.

⑤ 위험에 처한 애완동물을 돕는데 입양비를 사용한다.

13. 그림의 상황에 가장 적절한 대화를 고르시오.



① ② ③ ④ ⑤

14. 대화를 듣고, 남자의 마지막 말에 대한 여자의 응답으로 가장 적절한 것을 고르시오.

Woman:____

① Would you follow the yellow line?

② You can write better by practicing.

③ Take the three o’clock bus at the station.

④ Why don’t you send your painting to a contest?

⑤ Close your eyes and imagine anything about the topic.

15. 대화를 듣고, 여자의 마지막 말에 대한 남자의 응답으로 가장 적절한 것을 고르시오.

Man:____

① Calm down. You can get a refund.

② Keep an eye on my bag for a minute.

③ You can use a locker over there while shopping.

④ Would you start to work in customer’s service?

⑤ Where can I buy a bag in this department store?

16. 대화를 듣고, 남자의 마지막 말에 대한 여자의 응답으로 가장 적절한 것을 고르시오.

Woman:____

① Try not to watch TV for 24 hours.

② There’s something strange in my food.

③ That’s why we have some pizza and chicken.

④ Think about those poor children you can help.

⑤ I’ve finished a report about the starving children.

17. 다음 상황 설명을 듣고, Matthew가 Tiffany에게 할 말로 가장 적절한 것을 고르시오.

Matthew:____

① Tell your parents the truth and they'll help you.

② Don't use your parents' phones any more.

③ Sorry. You have the wrong number.

④ You'd better go to the repair shop.

⑤ That number is not in service now.

Appendix B-1

Script: Test (I5i)

W: female voice M: male voice

1. W:How may I help you, sir?

M:I’m looking for a mobile for my 3-month-old baby.

W:Do you have anything in mind?

M:Not really.

W: Then what do you think of ‘Starry night’? Your baby will enjoy stars and moons.

M:It’s cute. But, the mobile with a baseball over there looks better.

W: Which one? You mean the one with four cute bears?

M:Not that one. That looks too complicated for a young baby.

W:Then you like this simple baseball mobile with bats, gloves, and a ball.

M:You got it.

W:Great! It's machine washable. It’s only $30.

M:Sounds great. I’ll buy it.

(10 seconds)

2. *[telephone rings]*

W:Asia-Pacific Airways. Can I help you?

M:Yes. I need a flight from Seoul to Sydney on Tuesday.

W:Let me see. Yes. We have an 8:30 flight in the evening.

M:Eight thirty! What’s the check-in time?

W:One hour earlier than your flight time. Will you take that?

M:No. I won't get to the airport in time. When will the next flight leave?

W: There won't be another direct flight on Tuesday. There will be one on Wednesday at the same time.

M: Then can I make it to the business meeting in Sydney? It is at 9 o'clock on Thursday.

W: *[pause]* I'm sorry, but I don't think you can make it on time.

M: Really? Then, what am I going to do?

W:I'm afraid you have no choice right now.

(10 seconds)

3. W: Bill, what would you like to have for dinner?

M: I don't have much time, so I just want to eat a hamburger at the cafeteria. What about you?

W: I've heard there is a nice Italian restaurant near here. Can't we go there?

M: I'd love to, but I should finish this report here before they close. I need this book.

W: Why don't you just borrow it and finish your report after a nice dinner?

M: I can't. I've already borrowed all that I can.

W: Then let me check it out for you.

M: That's a great idea. Thank you.

(10 seconds)

4. W: David! How have you been lately?

M: Not bad. And you?

W: Much the same, except I do have some big news.

M: Big news? Come on! I’m dying to hear it.

W: Hold on now... *[pause]* I had a blind date a couple of weeks ago.

M: Jenny! You said that was not your style. I can’t believe it.

W: That was then and this is now.

M: This is all news to me. What is he like?

W: He seems nice, but I’m not sure.

M: You aren’t?

W: Not yet. Meet him for me and tell me about him, will you?

(10 seconds)

5. M: May I help you, ma’am?

W: Yes, I am looking for exercise equipment that I can use easily.

M: We have dumbbells. Women easily use them anywhere. Besides, the price is reasonable.

W: But I don’t know how to use them. Is it simple?

M: Sure. Let me show you. Hold your upper body straight. And swing the dumbbells like this with your elbows close to your sides.

W: Not that difficult. I think I can do that.

M: There are a wide range of weights but as a beginner you should use one kilogram dumbbells.

W: OK. I will take them.

(10 seconds)

6. W: So, how are things going, Steve?

M: Well, to be honest, Carla, I’ve got a cold.

W: Are you OK now?

M: Not really. I’m worried because I’m scheduled to hand in a report by Friday.

W: Don’t worry. You are going to get better in no time.

M: Well, I took some medicine, but it didn’t seem to help.

W: Listen, forget about that medicine! My mom’s herbal medicine will get rid of your cold.

M: Oh, Carla. You are so kind but...*[pause]* no thanks.

W: Come on! You’ll be up and dancing around soon.

M: OK. I’ll give it a try.

W: Great. My mom is at home right now. Let’s go get some herbal tea.

M: OK. I will.

(10 seconds)

7. M: Any toothbrush that you choose should have a soft brush and should be comfortable in your hand. You can choose between a manual and an electric toothbrush depending on your lifestyle and situation. If you prefer a manual toothbrush, make sure that the tip is small enough to reach all areas of your mouth easily. People with arm and shoulder problems might prefer an electric toothbrush for convenience as well as comfort. If purchasing an electric toothbrush, be sure that the head is soft and the brushes move in a back and forth motion.

(10 seconds)

8. W: Mr. Brown, here are your pictures.

M: Wow, these pictures are great. As I ordered, you enlarged the picture of me alone. I like this.

W: I’m happy you are satisfied.

M: How much is a small-sized one?

W: It’s 200 won.

M: I received 30 small-sized pictures and one large picture. What about the large one?

W: Originally it is 1,000 won, but I will give you a 50% discount.

M: Thank you. Here is 10,000 won.

W: Here’s your change, 3,500 won.

(10 seconds)

9. W: Attention, please. Today is the first day at the lab. So I’m going to explain how to use the tools on the table. The first thing on the left can be used to pick up and release small amounts of liquid like water or alcohol. I’ll show you how to use this. Squeeze the round bulb at one end of the stick and put the other end into the water like this. Release the bulb and the water will move up into it. If you want the water to go out, you just have to squeeze the bulb again. Understand?

(10 seconds)

10. M: How’s your new apartment working out, Ann?

W: Well, I like the apartment, but it’s too far from campus. I want to look for a new place.

M: Then did you go to a real estate agency?

W: No, you know, I have a final exam this Friday. Would you help me?

M: Sure. What kind of place are you looking for?

W: Above all, I want an apartment within walking distance to school.

M: Okay, anything else?

W: Uh, some place under $200 a month, including utilities, if possible.

M: I’ll drop by a real estate agency for you on my way to class today.

W: Do you think there are vacancies around the campus?

M: I saw the sign they have many available apartments.

(10 seconds)

11. W: November 20th is my daughter’s first birthday.

M: Congratulations! Are you going to hold a party?

W: Yes, I will. I want to give her a memorable party.

M: Are you going to prepare it by yourself?

W: No, I’m considering these party planning companies.

M: How much money are you planning to spend?

W: Um... Less than $1,000.

M: How long do you want your party to last?

W: I need 4 hours. And I want a family picture to remember her birthday.

M: Well, this company looks good for your party.

(10 seconds)

12. *[News Signal]*

M: This is Robert Brown with the Opening World. Here are today’s headlines:

Ozone has a stronger climate effect than we have thought until now.

Gorillas are in danger of extinction because of hunting in Congo.

Serious floods hit southern England suddenly.

Mobile phone users are increasing sharply in India.

That’s all for now. I’ll be back in a few minutes.

[News Music]

(10 seconds)

13. ① W: Where is everybody?

M: They are all in the living room watching TV.

② W: Is this seat taken?

M: No, go ahead.

③ W: What can I do for you?

M: I’d like to open a savings account.

④ W: How many people in your party?

M: There are three of us.

⑤ W: Excuse me, how can I get to City Hall?

M: Get on this bus. It is three stops away.

(10 seconds)

14. W: Time to wake up, James. Wake up!

M: What time is it?

W: It’s 7:30.

M: 7:30! Why didn’t you wake me up at 7? I’m late.

W: I’m sorry I didn’t know you had to wake up early today.

M: Today I have to give a presentation in class.

W: Why didn’t you set the alarm, if you had to wake up early?

M: I forgot to. I’m in a hurry. Mom, where is my USB?

W: On your desk. James, get dressed first and have your breakfast.

M: I don’t have time to eat. Oh, I don’t know what to wear. I’m really late, mom.

W: ____

(15 seconds)

15. *[telephone rings]*

W: Hello, John?

M: Oh, Jane. How are you?

W: Why are you still at home?

M: What do you mean?

W: We promised to meet at six today. Don’t you remember?

M: Oh, no! I thought we would meet each other on Friday.

W: No, we planned to meet on Thursday. I am waiting for you at Tim’s restaurant.

M: I’m really sorry. Please wait until I get there. I will go as soon as possible.

W: I see. I’ll wait for you. Drive safely.

M: ____

(15 seconds)

16. M:The flowers on your desk are very nice. Who sent them to you?

W:Patricia, my best friend. She is in Japan now.

M:She must be the friend that you’ve told me about before.

W:Right. We actually grew up together.

M:So, how long have you known each other?

W:Let me see. I guess we’ve known each other since elementary school. We’ve been friends for almost 15 years!

M:That’s a long time.

W:Yeah, in some ways I feel like she’s my sister.

M:How often do you keep in touch with her?

W:\____

(15 seconds)

17. W: Jim and his friend want to see the latest blockbuster. They promise to meet in front of the box office after school. Jim gets to the theater and looks around. But his friend doesn't come yet. Jim tries to buy tickets before his friend comes. It is such a nice movie that the theater is crowded. He finds a lady who seems to be waiting to buy a ticket. Jim wants to stand behind her after checking if she is standing there to buy one. In this situation, what would Jim most likely to say to the lady?

Jim: Excuse me, ____

(15 seconds)

Appendix B-2

Script: Test (II5i)

W: female voice M: male voice

1. W: Tom, why were you late this morning?

M: My alarm clock broke again.

W: Again? Why don't you buy a new one? Here's a shop that sells clocks.

M: Okay. Let's look around to see if we can find the right one.

W: Oh, the pyramid-shaped one over there looks exotic.

M: Well, it is not my style. I like round clocks.

W: Then, how about the one with a dog picture on it? You can hear the barking sound of a dog.

M: But I think it is for children. The apple-shaped one next to it is childish as well.

W: Hmm.... What about the one that looks like a computer mouse?

M: It's not bad. But I like sports, so I'd like to buy the sporty one, instead.

W: Great! Get it. Just don't kick it around in the playground.

2. M: What are you doing, honey?

W: I'm looking up a telephone number. Can you help me out?

M: Sure. What number are you looking for?

W: The Consumer Protection Union.

M: The Consumer Protection Union? Why?

W: You know, I bought a new computer game for Mike on the Internet.

M: Yeah. You told me he was very happy with it. What's wrong?

W: Every time he tries to play it, there is a new problem.

M: Then, why don't you exchange it for another one?

W: I tried to several times, but the company won't take it back. How could such a big company do that?

3. M: This is used to measure time in sports like swimming and track-and-field events. This can also be used in laboratory experiments. This has buttons that you press at the beginning and end of an event, so you can measure exactly how long it takes. Nowadays in schools, we see some students using this while studying to check if they can solve the questions within the given time.

4. W: Kevin, you don't look good. What's the matter?

M: Ms. Park caught me reading a comic book in her class.

W: Oh, boy. Did she scold you?

M: No, she just told me not to do it again.

W: Then, why are you so blue?

M: Because I'm worried she might be disappointed with me. I was so embarrassed that I couldn't say anything.

W: Why don't you apologize to her now?

M: Now? Don't you think it's too late?

W: Better late than never.

M: OK. I'll take your advice.

5. M: Wow, there are a lot of neckties.

W: Yeah. Take your time and call me if you need any help.

M: Um.... This striped one looks good.

W: Yes. This is very popular for young men.

M: I like it. How much is it?

W: It's 10 dollars, but you can get 20% off.

M: Really? That's great.

W: Today's our special bargain day. If you buy two, you get one free.

M: Sounds good, but I don't need that many. I only need one. I'll take it.

W: All right. One moment, please.

6. W: Hello, students. I'd like to inform you of the way to get counseling. First of all, you have to make an appointment with our staff member, Ms. Rey. She will give you the next available appointment. We encourage you to make an appointment before or after school, or during lunch

time. During lunch time, you don't have to wait long, so it is best to utilize that time. Please don't hesitate to ask for counseling service. We'll try hard to be available and accessible to you!

7.M: Is there any juice in the refrigerator?

W: No. You drank it all last night.

M: Then I'll go to the supermarket and buy some.

W: Wait. Could you do me a favor?

M: Sure, what is it?

W: I have an exam tomorrow, and I need some coffee to stay up all night. Will you buy me coffee?

M: Is that a good idea? Caffeine in coffee isn't healthy for you.

W: I know. But without it, I can't concentrate on studying.

M: You'd better not drink too much coffee, though.

W: Yeah, but I really need a cup of coffee tonight.

M: OK. But just this time. I'll be right back.

8. M: Hi, what can I do for you?

W: Hi, I'm a student of this dormitory. I have a problem with my mailbox.

M: Tell me the details.

W: I lost my mailbox key, so I haven't been able to get my mail. Do you have a spare key?

M: Of course I do, but I'm not supposed to give it out. These days we've had a lot of problems with mailbox theft in the dormitory.

W: I know what you mean. But I need to get an important piece of mail.

M: Um.... I see. Do you have your ID card?

W: Yes. Here is my student ID.

M: Okay. This time I'll help you.

W: Thank you!

9. W: Mark, we were supposed to turn left.

M: Oh, no! I'll turn left at the next corner.

W: But, I'm afraid we'll get lost.

M: Don't worry. I have a good sense of direction.

W: Don't you remember when we spent one hour finding the cinema last weekend?

M: Please forget that ever happened.

W: Mark! Why didn't you stop? We could get a ticket.

M: Oops! I didn't even see the stop sign.

W: You'd better slow down a little bit.

10. M: Say, Monica, could you do me a favor?

W: Sure, what would you like?

M: I'm typing my paper on the computer right now.

W: And....?

M: I'm afraid I will lose my data.

W: Oh, do you want to know how to back up the data on a CD?

M: Yes, you're right. I just want to do it in case I should mess up my data.

W: Better to be safe than sorry. I can let you know how to do it right now. Do you have a new CD?

M: Of course, I have one on my desk.

W: Bring your CD and I'll show you how to do it. It's really a piece of cake.

M: Thanks. I'll be back in a minute.

11. M: Wow, finally the exam is over. And it's Sunday. I'll relax and watch TV.

W: Frank, even though the exam is over, you have to keep studying.

M: I know, mom. But I want to watch TV today.

W: Then why don't you watch English Grammar on channel 5?

M: Please.... Don't say that. I've been looking forward to watching sports.

W: I see. What sports program do you want to watch?

M: Football on channel 7 at 7:00 and baseball on channel 6 at 9:30.

W: But I want to watch SBC Weekend News on channel 6.

M: Do you? Hmm.... Then, I'll study while you're watching the news and I'll watch what I want at 9:30.

W: Now that's a good plan!

12. M: Are you looking for an ideal place for your vacation? Gloria Hotel is the answer. We invite you to experience our wonderful hotel. Our hotel is located in the city center, within a short walk of major attractions. We have a variety of rooms with nice views. We especially recommend deluxe rooms which have been recently renovated. They have new furniture and wallpaper. For all our guests we offer free drinks at the snack bar. You can also enjoy 24-hour access to the indoor swimming pool. We look forward to seeing you soon. Thank you.

13. ① W: Can you tell me where the ladies' room is?

M: Yes, it's right down the hall. Just look for the sign.

② W: There's a park across the street.

M: Let's cross the street. The light is green.

③ W: You look busy. What are you doing here?

M: I'm gardening. It is my favorite hobby.

④ W: Look at the signs. We have to take the road on the right.

M: Right. I guess it'll take only half an hour to get to the falls.

⑤ W: It's getting colder and colder.

M: Yes, and all the leaves have fallen off the trees.

14. W: David, what's the matter with your face? You are so sunburned!

M: I know! I went hiking yesterday, and I walked for five hours in the sun without a hat.

W: Oh, dear! Does it hurt?

M: Yeah, it really does. I feel like I'm on fire.

W: Have you tried cucumber on your skin?

M: Cucumber? Does it really work?

W: Of course, it does. It helps to cool down your skin.

M: Does it? What should I do with the cucumber?

W: You just cut up a cucumber, and you put the slices on your skin. That's it.

M: ____

15. M: It's too dry these days.

W: You can say that again.

M: I think a drought has set in. It hasn't rained for months.

W: But didn't it rain last month?

M: It did rain last month, but a news reporter said the rainfall was only half the monthly average.

W: Oh, that bad?

M: Yeah, it's really bad. I'm very worried.

W: Come to think of it, the grass has almost dried out.

M: It's getting worse every day.

W: ____

16. M: Hi, Jane!

W: Hi, Michael. I heard you were sick last week. Are you okay?

M: I'm getting better.

W: I'm glad you're okay.

M: Well, sometimes I still don't feel very well.

W: Really? Why don't you see a doctor?

M: I went to the hospital this morning. My doctor advised me to start exercising.

W: That's a good idea. You need to do some exercise.

M: I agree, but I don't know what kind of exercise I should do.

W: How about working out at a fitness center?

M: A fitness center? Can you recommend one?

W: I heard the fitness center next to my office is giving a 30 percent discount this month.

M: ____

17. W: Judy was very happy to know that she was accepted to a university she wanted to enter. But this morning she got another letter of admission from a university guaranteeing her a scholarship. Now she has to select one university. If she chooses the first university, she can major in politics, which she really wants to study. If she chooses the second one, she won't be able to major in politics, but she can receive the scholarship. She has to make a decision by this weekend. So she has decided to get advice from her father. In this situation, what would Judy most likely say to her father?

Appendix B-3

Script: Test (III5i)

W: female voice M: male voice

1. W: Hey, they're having a sale today.

M: That's nice. I'm thinking of buying a humidifier. It's very dry these days.

W: That's right. The pig shaped one and the house shaped one are both cute.

M: Yes, but I want a simpler one.

W: Then, what do you think about this round one?

M: Well, that's too ordinary. What about this square one?

W: Which one? The one with flowers on it?

M: No, I mean the one with a mouse on it.

W: Yeah, that is cute and simple.

M: Yes. I'll get that one.

2. M: Welcome to Grand Cinema. How can I help you?

W: I reserved four seats for the two o'clock movie by phone.

M: Your membership number, please.

W: It's 3826.

M: Please, wait a moment. [*keyboard sound*] I'm sorry, but we have no reservation under that number.

W: No way! I reserved it just yesterday. Can you check one more time?

M: Okay. [*keyboard sound*] Sorry. Are you sure you used that membership number?

W: Definitely. [*pause*] Oh, wait. Will you check my husband's membership number? It's 3825.

M: Sure, ma'am. [*keyboard sound*] Yes, we have four seats reserved under that number.

W: Oh, that's great. I'm glad to hear that.

3. M: These are signaling devices used to control the flow of traffic. You can find them at the places where roads meet. You can find them in most cities around the world. They usually consist of a set of three lights. Red indicates "stop," yellow indicates "caution" and green indicates "go." They help drivers to avoid accidents and pedestrians to cross the street safely. You should always pay attention to these when using the roads.

4. W: Honey, are you going out now?

M: Yes. Are you staying home this evening? My laundry will be delivered around 8:00 P.M.

W: Oh, no. I'm going out to see my friend, Jina.

M: Oh, Jina? You haven't seen her since she got back from England, right?

W: Right. She came back to Korea last weekend.

M: Sounds great! [*pause*] But what am I going to do with my shirts?

W: When do you need them?

M: Tomorrow morning. You know, I'm leaving for my business trip.

W: Okay. Then I'll drop by the dry cleaner's and pick them up in the afternoon.

M: Thank you, honey.

5. M: Hi, Amy. I see you're having a garage sale.

W: Hi, Robert. Look around for something you need.

M: I see. Hmm... This bike looks almost new. How much is it?

W: I can give it to you for $80.

M: $80? That’s a bit expensive.

W: But I bought it just two months ago and haven’t ridden it much at all. It was $120, when it was new.

M: Well, I’m not sure...

W: If you buy this bike, I’ll give you this basket for free. I bought it for $20.

M: Well, that’s not bad. I’ll take it.

W: Thank you.

6. W: May I have your attention, please? Tomorrow morning, we’re going to explore a cave. Please remember the following: first, don’t forget to have a good breakfast. It will give you enough energy to explore the cave. Second, wear a long-sleeved T-shirt, long pants and strong boots to protect your body. Third, you will need a helmet and a good flashlight with spare batteries. Exploring caves can be a wonderful experience. But, if you’re not well prepared, it can be dangerous.

7. M: Hi. How are things going?

W: I’ve been so busy recently.

M: Yeah, I know. I heard that you’re preparing for a school festival.

W: Yes. I have to make pamphlets, invite guests and schedule the events.

M: Wow, lots of things to do! Oh, I can help with the pamphlets.

W: Thank you, but I’ve already asked someone else to help with that.

M: That’s great.

W: But I haven’t designed the invitation card. Could you do that for me?

M: Sure. When do you need it by?

W: As soon as possible. Thank you.

8. W: Welcome to Energy Factory. I’d be happy to show you our exhibitions today.

M: How many exhibits do you have in Energy Factory?

W: We have five exhibits. And this exhibit is the most popular in this exhibition.

M: Wow, it’s full of wonders.

W: Yes. This shows how electricity is used in our daily life.

M: Interesting! Oh, what’s this?

W: This model shows how water produces electricity. Would you like to try this?

M: Sure, I’d love to. How can I do it?

W: Pump the water into the tank until it is full.

M: Amazing! I powered the TV by pumping the water.

W: Good job. Now let’s move on to the next hall.

9. M: Good afternoon, ma’am. What can I do for you?

W: Good afternoon. I brought an empty flowerpot.

M: So, you want to plant something in it?

W: Yes. But I don’t know what to plant.

M: We have various kinds of flowers and trees. Would you like to look around?

W: Wow, they’re beautiful. Well... Could you recommend one?

M: How about sansevieria or English ivy? They're good for keeping the air clean.

W: Sansevieria sounds good to me. How much is it?

M: Ten dollars. It'll take about half an hour to plant it in your pot.

W: Okay. I'll be back after lunch.

10. M: Mom, you promised to buy me an MP3 player.

W: What? Why should I buy it for you?

M: How could you forget? You promised to buy it this vacation.

W: Yes. But you promised to clean your room three times a week. You didn't keep your promise.

M: Yeah. You're right. But could I have one more chance?

W: Well... [*pause*] Okay. But it's the last chance for you to earn an MP3 player.

M: I'll keep my room tidy and clean this time.

W: I'll watch you for a month and then decide.

M: Thanks for giving me another chance. I love you, Mom.

W: Why don't you go upstairs and start right now?

M: Okay. I'll sweep and wipe my room until it's shiny.

11. W: Good morning, Mr. Brown. Take a seat. What seems to be the problem?

M: Good morning. I have a pain in my stomach.

W: Okay. Do you have a fever?

M: A little. I haven't eaten anything since yesterday.

W: How long have you had the symptoms?

M: For two days.

W: Are you taking any medicine now?

M: Yes. I'm taking aspirin.

W: Okay. Are you allergic to any medicine?

M: No, I'm not.

W: I see. I think you have the flu. Take the medicine I prescribe and take a rest.

M: Okay. Thank you.

12. M: You should become a member of Pets & People today. Pets & People is an organization to assist pets in danger. We help homeless cats and dogs and keep them until we find someone to take care of them. Anyone who loves cats or dogs can become a member for free. As a member of Pets & People, you can help take care of the dogs and cats or even adopt them from us. The adoption fee is $40 for a cat and $50 for a dog. This money will be used to save other pets still in danger.

13. ① M: Would you give me a hand?

W: Yes. I can help you carry your luggage.

② M: Oh, hurry up. The elevator is going up.

W: Thank you for holding the door.

③ M: Excuse me. Something's falling out of your bag.

W: Oh, I forgot to zip it up. Thanks.

④ M: Oh, I'm sorry. I stepped on your foot.

W: That's okay. The elevator is so crowded.

⑤ M: I'd like to borrow these books.

W: Okay. Give me your membership card, please.

14. M: Sarah, how’s it going with your painting?

W: I’ve just finished the sketch. How about you?

M: Don’t ask me. I didn’t even start.

W: What’s the matter? Ms. Smith told us to finish by three o’clock.

M: Yeah, I know. But I don’t have any good ideas about today’s topic.

W: You know, painting is just putting your own ideas on canvas.

M: Putting my ideas on canvas?

W: That’s right. You can create anything from your imagination.

M: Then, how can I put my ideas on canvas?

W:\____

15. M: Excuse me, ma’am. You can’t have that bag while shopping.

W: What’s wrong?

M: I’m very sorry. But customers are not allowed to shop with a big bag.

W: Oh, I didn’t know that. But if that’s the case, you should at least notify customers about it.

M: I think you missed the sign. It says that no one may enter this place with such a big bag.

W: Oh, my... I’m sorry. I didn’t see it.

M: That’s all right, ma’am.

W: Then, what shall I do with my bag?

M:\____

16. W: Michael, I'm excited to be participating in the 24 Hour Fast with you.

M: Me, too. But eating nothing for 24 hours is not easy.

W: You're right. But it will help starving children.

M: It's much harder than I expected.

W: Right. After this experience, you'll be proud of yourself.

M: Yes, I will. But I'm so hungry now.

W: You've been doing a good job until now.

M: Oh, dear! I can't stop thinking of food.

W:____

17. W: Tiffany is happy to get a brand new cell phone. She likes to talk on it. She enjoys having long conversations over the phone. She doesn't care about the cell phone charge until she checks her phone bill. When she finally does, she is worried about it. It's too much. So she asks her friend, Matthew, what to do. He wants to suggest that she should explain the situation to her parents, and get help from them. In this situation, what would Matthew most likely say to her?

Matthew:____

Appendix C

Survey Questionnaire

1. Which multiple-choice item format do you most prefer among the three-, four-, and five-option items?

2. Why do you prefer it?

3. What do you think about a three-option item format?

Most dislike Most like

1 2 3 4 5 6 7

4. Can you think of any other reasons why the three-option-item format would be good?

5. Can you think of any other reasons why the three-option-item format would be bad?

6. If the CSAT were to be administered with a three-option item format, do you think you would like it?

7. Why do agree or why do you disagree with the three-option-item format on the CSAT?

Note. The original questionnaire is written in Korean, the native language of the participants.

References

References

Brown, J. D. (2001). Using surveys in language programs. Cambridge language teaching library. Cambridge, U.K: Cambridge University Press.

Brown, J. D. (2005). *Testing in language programs*. New York, NY: McGraw-Hill.

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

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

Card, J. (2005, November, 30). Life and death exams in South Korea. *Asia times.* Retrieved on Mar. 23, 2010 from http://www.atimes.com/atimes/Korea/GK30Dg01.html

Cizek, G. J., & Rachor, R. E. (1995, April). *Nonfunctioning options: A closer look.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Choi, I-C. (2008). The impact of EFL testing on EFL education in Korea *Language Testing*, *25*(1), 39-62.

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

Costin, F. (1972). Three-choice versus four-choice items: implications for reliability and validity of objective achievement tests. *Educational and Psychological Measurement*, *32,* 1035-1038.

Costin, F. (1976). Difficulty and homogeneity of three-choice versus four-choice objective test items when matched for content of stem. *Teaching of Psychology*, *3*, 144–145.

Cox, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, *17*, 407-422.

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

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

Dörnyei, Z. (2003). *Questionnaires in second language research: Construction, administration, and processing.* Mahwah, N.J: Lawrence Erlbaum Associates.

Ebel, R. L. (1969). Expected reliability as a function of choices per item. *Educational and Psychological Measurement, 29,* 565-570.

Green, K., Sax, G., & Michael, W. B. (1982). Validity and reliability of tests having different numbers of options for students of differing levels of ability. *Educational and Psychological Measurement 42*, 239-245.

Grier, J. B. (1976). The optimal number of alternatives at a choice point with travel time considered. *Journal of Mathematical Psychology, 14*, 91-97.

Gyeonggi Provincial Office of Education. (2007, November 22) *The CSAT English prep-test* [Data file]. Available from EBSi Web site, http://www.ebsi.co.kr/ebs/ent/enta/retrieveNaarPrdRdrInfo.ebs

Gyeonggi Provincial Office of Education. (2008, November 18) *The CSAT English prep-test* [Data file]. Available from EBSi Web site, http://www.ebsi.co.kr/ebs/ent/enta/retrieveNaarPrdRdrInfo.ebs

Gyeonggi Provincial Office of Education. (2009, November 17) *The CSAT English prep-test* [Data file]. Available from EBSi Web site, http://www.ebsi.co.kr/ebs/ent/enta/retrieveNaarPrdRdrInfo.ebs

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

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

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

Hedrick, W. B., & Cunningham, J. W. (1995). The relationship between wide reading and listening comprehension of written language. *Journal of Reading Behavior, 27*(3), 425-438.

Hedrick, W. B., & Cunningham, J. W. (2002). Investigating the effect of wide reading on listening comprehension of written language. *Reading Psychology, 23*(2), 107-126.

Hogben, D. (1973). The reliability, discrimination and difficulty of word knowledge tests employing multiple choice items containing three, four or five alternatives. *Australian Journal of Education, 17,* 63-68.

Kolstad, R. K., Briggs, L. D., & Kolstad, R. A. (1985). Multiple-choice classroom achievement tests: Performance on items with five vs. three choices. *College Student Journal, 19,* 427–431.

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

Lee, S. (2005, October, 23). Samsip samilganeui gamgeun suneung chujewuiwoneui bimil. *The hankyoreh*. Retrieved on Mar. 23, 2010 from http://www.hani.co.kr/arti/society/schooling/73560.html

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

Lord, F. (1944). Reliability of multiple-choice test as a function of choices per item. *Journal of Educational Psychology, 35*, 175-180.

Millman, J., Bishop, H. I., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement, 25*, 707-726.

Owen, S. V., & Froman, R. D. (1987). What’s wrong with three-option multiple choice items? *Educational and Psychological Measurement 47,* 513-22.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of Research. *Educational Measurement: Issues & Practice, 24*(2), 3-13.

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

Ruch, G.M., & Stoddard, G. D. (1927). *Tests and measurements in high school instruction.* Chicago: World Book.

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

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

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

Torabi-Parizi , R., & Campbell, N. J. (1982). Classroom test writing: Effect of item format on test quality. *Elementary School Journal, 83*(3), 155-160.

Trevisan, M. S., Sax, G., & Michael, W. B. (1991). The effect of the number of options per item and student ability on test validity and reliability. *Educational and Psychological Measurement*, *51*, 829-837.

Trevisan, M. S., Sax, G., & Michael, W. B. (1994). Estimating the optimum number of options per item using an incremental option paradigm. *Educational and Psychological Measurement, 54,* 86–91.

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

University of Washington (1983). *Washington Pre-College Admissions Test Battery*. Seattle, WA: University of Washington, Washington Pre-College Department.