Attractiveness of Distractors in Multiple-Choice Listening Tests

Hideki IIMURA

Prefectural University of Kumamoto

Abstract

This study investigates how distractors function in multiple-choice listening tests. Conventionally, each distractor is evaluated by its attractiveness. In other words, distractors that can plausibly be chosen by test takers are considered to perform well on the test. On the other hand, distractors that are chosen by few test takers are recognized as performing poorly. Considering that test takers have to choose only one among several options, it can be assumed that some unselected distractors may have in fact performed adequately. Therefore, it is prudent to independently analyze the attractiveness of distractors. A total of 75 Japanese university students evaluated their confidence in selecting both correct and incorrect answers. The results indicated that (a) the least chosen distractors were not always the least attractive, (b) less proficient listeners were more likely to be allured by distractors, and (c) more proficient listeners were more likely to answer with higher confidence. The researcher explains the process of eliminating distractors and reevaluates the unselected distractors.

1. Introduction

Multiple-choice (MC) format has been widely adopted in listening comprehension tests. A typical MC item consists of three or four options. One of the options is correct, often called the *key*, while the other options, the *distractors*, are incorrect answers. Test takers are expected to choose the key, and not be distracted by the other options.

Downing (2006) explains that MC format is favored in large-scale achievement tests because this format has significant validity advantages. Notwithstanding its efficiency for automated scoring, MC format, if properly constructed, can assess a wide range of content and measure higher cognitive skills. However, MC format has often been criticized for its limitations. It has been purported that MC format induces guessing (Hughes, 2003). Furthermore, valid MC questions are very difficult to construct. Specifically, it is difficult to create effective distractors. Moreover, it is extremely difficult to create a sufficient number of plausible distractors (Thompson, 1995). In fact, it has been suggested that only one or two distractors among four or five options will function as intended (Haladyna, 2004).

Evidently, distractors play a vital role in the success of MC format. One of the keys to creating good MC tests is indeed the effectiveness of its distractors. The effectiveness of a distractor can be evaluated on its functionality. Conventionally, distractors' functionality has been evaluated in two ways: (a) discriminability, and (b) response frequency (Fulcher, 2010; Haladyna, 2004;

Haladyna & Downing, 1993; Henning, 1987; Hughes, 2003). The former measures the extent to which a distractor can distinguish among high-ability and low-ability test takers. The latter measures the number of test takers that choose each distractor. Thus, dysfunctional distractors have no discriminatory power and are unpopular among test takers. Namely, distractors that succeed in attracting several test takers are functional, whereas distractors that attract few test takers are dysfunctional. Thus, those distractors judged as unattractive are targets for replacement.

The following two empirical studies, which investigated the optimal number of options in MC format, demonstrated the manner in which unattractive distractors were detected and discarded. Shizuka, Takeuchi, Yashima, and Yoshizawa (2006) investigated the effects of reducing the number of options in a reading comprehension test of a Japanese university entrance exam. They changed the standard four-option test items to three-option items by eliminating the least chosen option from a previous administration of the test. The three-option tests were given to Japanese university students who had not taken the original tests. Their results indicated that mean item facility and mean item discrimination between the four- and three-option test items were not significantly different. Furthermore, they found that the mean number of functioning distractors was lower than two, regardless of the number of options per item.

Using a listening section of a university entrance exam in Korea, Lee and Winke (2013) compared three different option formats; five-, four-, and three-option tests. Contrary to Shizuka et al. (2006) relying on item analysis statistics, they deleted distractors based on judgements of several item evaluations, such as graduate students, English language instructors, and senior high school students. It was suggested that this deletion method produced “a more balanced perspective (from test developers to test takers) on the least plausible options” (p. 8). Ultimately, the results were not straightforward. While mean test scores of the three-option tests were significantly higher than those of four- and five-options tests, no difference was found in the mean item discriminations across the three different test formats.

Although the previous studies employed different option deletion methods, namely, item analysis (Shizuka et al., 2006) and evaluators’ judgements (Lee & Winke, 2013), both studies relied on the identical assumption that the least chosen distractors cannot attract the attention of the test taker.

Considering the nature of the test-taking process in MC format, it is prudent to further consider the relationship between response frequency and attractiveness in MC formats. When taking an MC format test, test takers are supposed to answer three or four times per item, depending on the number of options, before they make a final decision. It is reasonable to assume that, before their final choice, test takers consider other options. In other words, unselected distractors cannot be uniformly unattractive. Given that test takers must select only one option, it can be postulated that some unselected distractors function adequately. In other words, we should evaluate each distractor’s attractiveness independent of its response frequency.

Thus, the conventional evaluation of an item’s functionality based on response frequency is insufficient. We must develop an alternate method for evaluating distractor attractiveness. The present study adopts a questionnaire from Shizuka (1999a, 1999b), where test takers were asked to

both choose the correct response and choose one of three confidence levels with which they have selected their response. Although he focused on test takers' perception about one of the options (the key), the present study will examine the distractors as well.

2. Study Aims

The purpose of this study is to investigate the function of distractors in MC format. Specifically, this study seeks to do the following:

1. Compare keys with distractors with regard to test takers' confidence levels;
2. Compare the attractiveness of distractors in terms of test takers' confidence levels and response frequency.

3. Method

3.1 Participants

Seventy-five undergraduates from two private universities in Japan participated in the study. Participants were divided into three proficiency groups based on the results of the listening tests in this study (see 4.1).

3.2 Materials

3.2.1 Listening Test

Fifteen listening test items from the Test of English for International Communication (TOEIC) preparation book (Educational Testing Service, 2010) were used in the present study. All items were question-response tasks consisting of three options: one key and two distractors. In this task, test takers heard a question or statement and selected the most likely response. Neither questions nor options were printed in the answer sheet, and both were played once.

In the original TOEIC CD, an approximately one-second pause was inserted between options and a three-second pause between items. Considering the test takers' cognitive load in this experiment, the original pauses were not long enough. Therefore, the researcher edited the CD to have three-second pauses between options for answering the questionnaire and eight-second pauses between items to both answer the questionnaire and select the best answer for the item itself (see Figure 2).

3.2.2 Questionnaire

To elicit test takers' perceptions of distractors and keys, the researcher developed a questionnaire that served as both, an answer sheet and a survey of confidence.¹ As seen in Figure 1, the questionnaire is symmetrically divided at the middle point (0: I'm not sure) into two sides for correctness (correct or incorrect) with three confidence levels for both correct and incorrect options (H: very confident; M: moderately confident; L: not confident).

Test takers were expected to circle at the middle of the line according to their confidence level in their judgement for each option's correctness. In addition, they were required to circle one of the capital letters—A, B, or C—representing each option.

Figure 1 illustrates a case in which a participant judged A as incorrect with moderate confidence, B as correct with high confidence, and C as incorrect with low confidence. Therefore, the test taker decided to choose B as his/her final decision.

Figure 1. Sample of the questionnaire.

3.3 Procedure

Data was collected during July 2013. Each session was administered by classroom teachers at their respective universities. Before each session, participants were informed that all identifying information would remain confidential.

Figure 2 shows the task flow in the present study. First, participants were instructed to choose their confidence level for Option A (see also Figure 1) immediately after hearing this option. This step was identical for Option B and C. Participants were then asked to circle one of the three options as their final answer for the item. Prior to the session, participants were given two sample questions to familiarize themselves with the procedure. The listening test was played with a CD player.

Figure 2. Task Flow in the study.

3.4 Analysis

3.4.1 Listening Test

Descriptive statistics and the reliability with Cronbach's alpha were calculated. Item analysis was also conducted. As a discrimination index, the point-biserial correlation (r_{pbi}) was computed.

3.4.2 Questionnaire

3.4.2.1 Confidence Level (Keys)

The questionnaire data were coded with values ranging from one to seven based on the correctness of the response and test taker's confidence levels. Table 1 shows how test takers' confidence levels were converted into scores. Also, descriptive statistics and reliability with Cronbach's alpha were calculated.

Table 1

Conversion Table for Confidence Ratings on Keys

Confidence level	Correctness	Point
Very confident	Correct	7
Not very confident	Correct	6
Not confident	Correct	5
I'm not sure	-----	4
Not confident	Incorrect	3
Not very confident	Incorrect	2
Very confident	Incorrect	1

3.4.2.2 Distractor's attractiveness

Distractor attractiveness was coded based on correctness and confidence on a scale with values ranging from one to seven. Table 2 demonstrates the conversion procedure for generating an attractiveness score from correctness and confidence. An attractive distractor leads the test taker into deeming it *correct* with high confidence. Therefore, a code of seven was given to those answers where a participant deemed the distractor correct with high confidence. Conversely, the unattractive distractor is the one that only makes test takers believe it *incorrect*. Therefore, the distractor that was judged as incorrect with high confident was coded as one.

Table 2

Conversion Table on Distractor's Attractiveness

Confidence level	Correctness	Attractiveness
Very confident	Correct	7
Not very confident	Correct	6
Not confident	Correct	5
I'm not sure	-----	4
Not confident	Incorrect	3
Not very confident	Incorrect	2
Very confident	Incorrect	1

Also, descriptive statistics and reliability with Cronbach's alpha were calculated. Moreover, to determine the effect of distractor's attractiveness (within-subjects: two distractors) and the effect of listening proficiency level (between-subjects: three proficiency levels), two-way mixed design ANOVAs were performed for each item. The alpha level was set to .003 (.05/15), since 15 significance tests were conducted, one for each item.

4. Results and Discussion

4.1 Test Takers' Listening Proficiency

Descriptive statistics and internal consistency for the listening test are presented in Table 3. The mean score of all the participants was 9.15 out of the total score 15, with the standard deviation of 2.96. This suggests that the test was appropriate in terms of difficulty. The reliability of the total items was .70, which was acceptable considering the small sample size and number of items in the questionnaire.

Table 3

Descriptive Statistics and Cronbach's Alpha of the Listening Test (K = 15)

<i>N</i>	<i>M</i>	<i>SD</i>	<i>Max</i>	<i>Min</i>	<i>α</i>
75	9.15	2.96	15	3	.70

Note. Possible maximum score is 15.

Based on the distribution of the test scores (Figure 3), participants were divided into three groups: low-proficiency group (LP), whose scores were lower than 7; middle-proficiency group (MP), whose scores ranged from 7 to 11; and high-proficiency group (HP), whose scores were higher than 11. Table 4 shows the mean scores and standard deviations for each group.

Figure 3. Distribution of the listening test scores.

Table 4

Three proficiency Groups on the Listening Test

Group	<i>n</i>	<i>M</i>	<i>SD</i>
LP	14	4.71	1.07
MP	44	9.07	1.42
HP	17	13.00	1.06

4.2 Comparison of Distractors With Keys

Table 5 outlines descriptive statistics as well as reliability coefficients for both distractors and keys between the three proficiency groups.

First, we examined internal consistency reliability. Both distractors ($\alpha = .92$) and keys ($\alpha = .80$) demonstrated strong internal consistency. Internal consistency for distractors was higher because the number of distractors was higher (30—two for each item) than the number of keys

(15—one key for each item for 15 items). Since Cronbach's alpha depends on the number of items on the scale (Field, 2009), it is evident that the number of items influenced the ultimate value for Cronbach's alpha.

Next, we investigated the attractiveness of distractors and confidence in keys across the three proficiency groups. These scores are outlined in Figure 4 and 5. Attractiveness of distractors decreased as the proficiency level increased, whereas the average confidence levels in keys increased as proficiency levels increased.

Table 5

Average Attractiveness of Distractors and Test Takers' Confidence in Keys

Level	<i>n</i>	Distractors <i>M</i> (<i>SD</i>)	α	Keys <i>M</i> (<i>SD</i>)	α
LP	14	3.8 (0.4)	.92	3.9 (0.4)	.80
MP	44	3.5 (0.6)		4.7 (0.5)	
HP	17	2.3 (0.7)		5.7 (0.5)	
Total	75	3.3 (0.8)		4.8 (0.8)	

Note. Possible maximum attractiveness/confidence score is 7.

Figure 4. Attractiveness of distractors across groups.

Figure 5. Test takers' confidence in correct answers across groups.

The one-way ANOVA was performed on the mean attractiveness of distractors between groups. Table 6 outlines a significant difference in the average attractiveness of distractors between the three groups; $F(2, 72) = 29.85, p < .001$. The effect size was large (eta squared = .21). The Bonferroni's post-hoc test demonstrated that the mean attractiveness of distractors for LP and MP groups did not differ, but both were significantly higher than the attractiveness of distractors for HP participants, $p < .05$ (Table 7). These results indicate that distractors tend to be more attractive to less proficient test takers. In other words, less proficient test takers are more likely to be distracted by incorrect answers than proficient test takers.

Table 6

One-way ANOVA for Average Attractiveness of Distractors Across Groups

Source	Sum of Squares	df	Mean Square	F	p	η^2
Between	22.05	2	11.02	29.85	<.001	.21
Within	26.59	72	0.37			
Total	48.64	74				

Table 7

Multiple Comparison of Distractors' Attractiveness Across Groups

	Low proficiency	Middle proficiency	High proficiency
LP	—	<i>ns</i>	*
MP		—	*
HP			—

Note. * $p < .05$.

Another one-way ANOVA was conducted to determine whether there was a significant difference in confidence levels for correct answers between the three proficiency groups. As Table 8 shows, a significant difference was found between groups, $F(2, 72) = 53.22, p < .001$. The effect size was large (eta squared = .36). The Bonferroni's post-hoc test revealed that all groups significantly differed from each other, $p < .05$ (Table 9). These results suggest that proficient test takers tend to choose their answers with more confidence than less proficient test takers.

Table 8

One-way ANOVA for Test Takers' Confidence in Keys Across Groups

Source	Sum of Squares	df	Mean Square	F	p	η^2
Between	25.31	2	12.66	53.22	<.001	.36
Within	17.12	72	0.24			
Total	42.43	74				

Table 9

Multiple Comparison of Test Takers' Confidence in Correct Answer

	Low proficiency	Middle proficiency	High proficiency
LP	—	*	*
MP		—	*
HP			—

Note. * $p < .05$

In sum, test taker proficiency differences appear to affect attractiveness of distractors and confidence level ratings in their choice. These findings seem plausible, as proficient learners can control their listening process by utilizing metacognitive strategies, primarily comprehension monitoring (Goh, 2002; O'Malley, Chamot, & Kupper, 1989, Vandergrift, 2003, Vandergrift & Goh, 2012). Their skills should have manifest in success of the MC formats.

4.3 Comparison Between Attractiveness and Frequency in Distractors

Table 10 presents aggregate data to examine the relationship between distractor attractiveness and response frequency. To look into this point closely, another factor, namely, proficiency difference, was taken into consideration. Thus, two-way mixed ANOVAs were carried out on coded data to evaluate the effect of distractor attractiveness and proficiency level (see Appendix 2 for detailed information on these analyses). In addition to the summary of two-way ANOVAs, this table also reports item facility (%), response frequency (%), and discriminatory power (point-biserial correlation coefficient) for each item.

Before examining results from the ANOVAs, it is prudent to first discuss discriminability, as this is another facet of distractor functionality. As seen in Table 10, all distractors produced negative point-biserial correlation. Considering negative or low correlations are desirable for functional distractors (Green, 2013), all distractors in this experiment were functional.

Table 10

Summary of the Survey: Item Facility, Two-way ANOVAs, Frequency of Response, and Discriminatory Power in Distractors

Item # (IF: %)	Pro× Attract	Proficiency	Attractiveness	Response ^a (%)	Discriminatory (r_{pbi})
1 (36.0)	<i>ns</i>	<i>sig</i> HP < MP $\div$ LP (2.1) (4.0) (4.6)	<i>sig</i> A < C (2.4) (4.8)	A < C (2.7) (61.3)	B = .53 A = -.14 C = -.47
2 (81.3)	<i>ns</i>	<i>sig</i> HP < MP $\div$ LP (1.4) (2.7) (3.6)	<i>ns</i> B $\div$ C (2.5) (2.6)	C < B (8.0) (10.7)	A = .37 B = -.26 C = -.24
3 (64.0)	<i>ns</i>	<i>sig</i> HP < MP $\div$ LP (2.6) (3.6) (3.8)	<i>sig</i> A < B (2.9) (3.6)	A < B (17.3) (18.7)	C = .47 A = -.32 B = -.27
4 (66.7)	<i>ns</i>	<i>sig</i> HP $\div$ MP $\div$ LP (2.7) (3.6) (4.1)	<i>ns</i> C $\div$ A (3.3) (3.7)	C < A (13.3) (20.0)	B = .54 A = -.33 C = -.36
5 (70.7)	<i>ns</i>	<i>sig</i> HP < MP < LP (1.9) (3.0) (4.1)	<i>ns</i> C $\div$ B (2.9) (3.2)	C < B (5.3) (24.0)	A = .64 B = -.52 C = -.31
6 ^b (46.7)	<i>sig</i>	A: HP < MP $\div$ LP (2.2) (3.7) (4.5) B: LP $\div$ HP $\div$ MP (3.3) (3.5) (3.6)	LP: B < A (3.3) (4.5) MP: B $\div$ A (3.6) (3.7) HP: A < B (2.2) (3.6)	B < A (25.3) (28.0)	C = .39 A = -.44 B = -.01
7 (50.7)	<i>ns</i>	<i>sig</i> HP < MP $\div$ LP (2.3) (3.6) (3.7)	<i>ns</i> A $\div$ C (3.1) (3.3)	C < A (20.0) (29.3)	B = .42 A = -.27 C = -.22
8 (50.0)	<i>ns</i>	<i>ns</i> HP $\div$ MP $\div$ LP (3.4) (3.7) (4.1)	<i>ns</i> C $\div$ A (3.5) (4.0)	C < A (23.0) (27.0)	B = .37 A = -.08 C = -.36
9 (77.3)	<i>ns</i>	<i>sig</i> HP < MP $\div$ LP (1.7) (3.4) (3.7)	<i>sig</i> B < A (2.7) (3.2)	B < A (8.0) (14.7)	C = .29 A = -.19 B = -.20

(Continued)

Table 10

Summary of the Survey: Item Facility, Two-way ANOVAs, Frequency of Response, and Discriminatory Power in Distractors (continued)

Item # (IF: %)	Pro × Attract	Proficiency	Attractiveness	Response ^a (%)	Discriminatory (r_{pbi})
10 (94.7)	<i>ns</i>	<i>sig</i> HP < MP $\div$ LP (1.5) (2.9) (3.1)	<i>ns</i> A $\div$ C (2.4) (2.5)	A = C (2.7) (2.7)	B = .31 A = -.24 C = -.19
11 (82.7)	<i>ns</i>	<i>sig</i> HP < MP $\div$ LP (2.2) (3.3) (3.4)	<i>sig</i> B < A (2.7) (3.1)	B < A (8.0) (9.3)	C = .29 A = -.17 B = -.22
12 (86.5)	<i>ns</i>	<i>sig</i> HP < MP $\div$ LP (1.3) (2.9) (3.6)	<i>ns</i> B $\div$ C (2.4) (2.8)	B < C (1.4) (12.2)	A = .38 B = -.15 C = -.35
13 (29.7)	<i>ns</i>	<i>sig</i> HP < MP $\div$ LP (2.5) (3.8) (4.1)	<i>sig</i> A < B (3.0) (3.9)	A < B (18.9) (51.4)	C = .59 A = -.18 B = -.40
14 (45.2)	<i>ns</i>	<i>sig</i> HP < MP $\div$ LP (2.9) (3.9) (4.3)	<i>ns</i> B $\div$ C (3.5) (3.8)	B < C (26.0) (28.8)	A = .47 B = -.13 C = -.39
15 (36.5)	<i>ns</i>	<i>ns</i> HP $\div$ MP $\div$ LP (3.0) (3.9) (4.0)	<i>ns</i> A $\div$ C (3.6) (3.7)	C < A (27.0) (36.5)	B = .34 A = -.30 C = -.05

Note. Boldface (A, B, or C) in "Response (%)" indicates the least chosen distractor. Italicized face (A, B, or C) in "Discriminatory" denotes the key. Shaded items show that the least chosen distractor is the least attractive. IF = item facility. Pro = proficiency. Attract = attractiveness. LP = low proficiency group; MP = middle proficiency group; HP = high proficiency group. Numbers in parentheses in both "Proficiency" and "Attractiveness" denote distractor's attractiveness ranging from one to seven.

^aMathematical signs in "Response (%)" are not based on inferential statistics. ^bThere was a significant interaction effect between proficiency and distractor attractiveness, $F(2, 72) = 7.09$. Subsequently, both the main effect of proficiency and of distractor attractiveness was analyzed.

sig: $p < .003$.

Following the two-way ANOVAs, the fifteen items can be classified into four distinct categories: (a) an item with a significant interaction between proficiency and attractiveness, (b) items with a significant main effect for both proficiency and attractiveness, (c) items with a significant main effect for proficiency, and (d) items with neither a significant interaction nor main effects. These are summarized in Table 11.

Table 11

Items Categorized Based on the Results of Two-way ANOVAs

Category	Proficiency × Attractiveness	Proficiency	Attractiveness	Item
(a)	✓			6
(b)		✓	✓	1, 3, 9, 11, 13
(c)		✓		2, 4, 5, 7, 10, 12, 14
(d)				8, 15

There was an interaction evident for item 6, the only item in the first category (a), which suggests that one of the distractors had almost the same degree of attractiveness for every proficiency group, while the other distractor was more attractive only to less proficient groups. For items in the second category (b), the degree of attractiveness between two distractors differed as a function of proficiency levels. For items in the third category (c), the degree of attractiveness between the two distractors did not differ, but the degree of attractiveness between two distractors differed as a function of proficiency levels. Lastly, for items in the fourth category (d), the two distractors did not differ in attractiveness regardless of proficiency levels.

It is pertinent to note that in 9 of 15 items (in Category c & d), the two distractors did not differ from each other in attractiveness. It follows that frequency of response is not always consistent with the attractiveness of distractors. In fact, there were only five items in which the least chosen distractor was also the least attractive (item 1, 3, 9, 11, & 13; see shaded areas in Table 10). Conversely, in 7 of 15 items, the least chosen distractor was not the least attractive. The least frequent choice in item 7, for example, was C, which was selected by 20% of test takers, but the attractiveness of that option (3.3) was almost the same (even somewhat higher) as that of A (3.1) chosen by approximately 29% of test takers. A similar pattern was observed in item 15, where C (27%) was less chosen by test takers than A (36.5%), but the attractiveness of the former (3.7) was almost equal to the latter (3.6).

It is unreasonable, however, to apply this rationale to extremely easy items (e.g., items 10 and 12; $IF = 94.7\%$ and 86.5% , respectively), since it is possible that the items themselves did not function well. Nevertheless, we could probably say that distractors cannot be adequately evaluated with conventional methods, i.e., response frequency and discriminability. Namely, these methods do not take into account the nature of the test-taking process. Therefore, it is prudent to incorporate test takers' perspective in the development of MC format testing.

5. Conclusion

To my knowledge, this is the first study to specifically examine the attractiveness of distractors. Functionality of distractors has long been evaluated by response frequency. Namely, prior investigations have assumed a direct relationship between response frequency and

attractiveness.

The current study tried to examine the relationship between attractiveness and frequency in distractors. Specifically, an MC test served as a questionnaire that queried test taker confidence in each option. The main finding of this study suggests that the least frequently chosen distractors are not always the least attractive. Therefore, it can be stated that conventional item analyses might be insufficient in evaluating the effectiveness of distractors during test development. A new kind of survey in which test takers can evaluate each distractor independently should be included in test development. Since the questionnaire employed in this study had strong internal consistency, it would be advisable for test developers to use and analyze this questionnaire for future use.

Furthermore, test takers' proficiency affects the attractiveness of distractors. Specifically, less proficient test takers are more likely to be actually distracted by distractors. Finally, proficient test takers are more likely to choose their answer with greater confidence.

This study has the following limitations. First, the sample size was relatively small and the questionnaire was brief. To generalize current findings, more data, many more items, and a wider range of proficiency levels must be aggregated. Second, this study employed only a question-response task in which where both questions and options were given verbally to test takers. Future research should include various MC tasks, such as picture-description and passage comprehension, that might contain more information and more complex discourse structure. Third, this study focused on a three-option test format. However, the four-option test format is more widely employed in major domestic and international English proficiency tests such as TOEFL ITP, EIKEN, and TOEIC (except question-response tasks). Including these factors will surely lead to further revelations about distractor attractiveness and test validity.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number (23520692). I would like to thank Dr. Rie Koizumi, Dr. Yo In'nami, Dr. Sachiyo Takanami, and two anonymous reviewers for their valuable comments on an earlier version of this article.

Notes

1. The original version of the questionnaire was written in Japanese (see Appendix 1).

References

- Downing, S. M. (2006). Selected-response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287–301). Mahwah, NJ: Lawrence Erlbaum Associates.
- Educational Testing Service. (2011). *TOEIC test official practice: Listening*. Tokyo: Institute for International Business Communication.
- Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage.
- Fulcher, G. (2010). *Practical language testing*. London: Hodder Education.
- Goh, C. (2002). Exploring listening comprehension tactics and their interaction patterns. *System*, 30,

185–206.

- Green, R. (2013). *Statistical analyses for language testers*. Hampshire: Palgrave Macmillan.
- Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*, 53, 999–1010.
- Henning, G. (1987). *A guide to language testing: Development, evaluation, research*. Cambridge; MA: Newbury House.
- Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.
- Lee, H., & Winke, P. (2013). The differences among three-, four-, and five-option-item formats in the context of a high-stakes English-language listening test. *Language Testing*, 30, 99–123.
- O'Malley, J. M., Chamot, A. U., & Kupper, L. (1989). Listening comprehension strategies in second language acquisition. *Applied Linguistics*, 10, 418–437.
- Shizuka, T. (1999a). Combining response correctness and confidence level rating to produce polychotomous data from MC dichotomous items. *Kosen Kyoiku (Japanese Colleges of Technology Education Journal)*, 22, 243–252.
Retrieved from <http://sucra.saitama-u.ac.jp/modules/xoonips/detail.php?id=A3000286>
- Shizuka, T. (1999b). Combining response correctness and confidence level rating to produce polychotomous data from MC dichotomous items (II). *KATE Journal*, 13, 73–86.
- 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, 35–57.
- Thompson, I. (1995). Assessment of second/foreign language listening comprehension. In D. Mendelsohn & J. Rubin (Eds.), *A guide for the teaching of second language listening* (pp. 31–58). San Diego, CA: Dominic Press.
- Vandergrift, L. (2003). Orchestrating strategy use: Toward a model of the skilled second language learner. *Language Learning*, 53, 463–496.
- Vandergrift, L., & Goh, C. M. (2012). *Teaching and learning second language listening: Metacognition in action*. New York: Routledge.

Appendices

Appendix 1

Questionnaire for confidence in each option

No.1	<div style="display: flex; justify-content: space-between;"> <div style="width: 30%;"> <p>A</p> <p>とても自信がある やや自信がある あまり自信がない 分からない あまり自信がない やや自信がある とても自信がある</p> <p>高 中 低 なし 低 中 高</p> <p>正 解 不正 解</p> </div> <div style="width: 30%;"> <p>B</p> <p>とても自信がある やや自信がある あまり自信がない 分からない あまり自信がない やや自信がある とても自信がある</p> <p>高 中 低 なし 低 中 高</p> <p>正 解 不正 解</p> </div> <div style="width: 30%;"> <p>C</p> <p>とても自信がある やや自信がある あまり自信がない 分からない あまり自信がない やや自信がある とても自信がある</p> <p>高 中 低 なし 低 中 高</p> <p>正 解 不正 解</p> </div> </div>
No.2	<div style="display: flex; justify-content: space-between;"> <div style="width: 30%;"> <p>A</p> <p>高 中 低 なし 低 中 高</p> <p>正 解 不正 解</p> </div> <div style="width: 30%;"> <p>B</p> <p>高 中 低 なし 低 中 高</p> <p>正 解 不正 解</p> </div> <div style="width: 30%;"> <p>C</p> <p>高 中 低 なし 低 中 高</p> <p>正 解 不正 解</p> </div> </div>
No.3	<div style="display: flex; justify-content: space-between;"> <div style="width: 30%;"> <p>A</p> <p>高 中 低 なし 低 中 高</p> </div> <div style="width: 30%;"> <p>B</p> <p>高 中 低 なし 低 中 高</p> </div> <div style="width: 30%;"> <p>C</p> <p>高 中 低 なし 低 中 高</p> </div> </div>

Appendix 2

Results of Mixed Design Two-Way ANOVAs

Item 1

Source	SS	df	MS	F	p	partial η^2
Between Subjects						
Proficiency	106.30	2	53.15	23.31	< .001	.39
Error	164.20	72	2.28			
Within Subjects						
Distractor	150.32	1	150.32	54.79	< .001	.43
Distractor $\times$ Proficiency	3.37	2	1.68	0.61	.54	.02
Error (distractor)	197.53	72	2.74			
Total	621.72	149				

Item 2						
Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	<i>partial</i> η^2
Between Subjects						
Proficiency	74.54	2	37.27	13.48	< .001	.27
Error	199.04	72	2.76			
Within Subjects						
Distractor	0.07	1	0.07	0.07	.788	.01
Distractor $\times$ Proficiency	3.17	2	1.59	1.66	.197	.04
Error (distractor)	68.83	72	0.96			
Total	345.65	149				
Item 3						
Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	<i>partial</i> η^2
Between Subjects						
Proficiency	29.49	2	14.74	6.44	.003	.15
Error	164.85	72	2.29			
Within Subjects						
Distractor	13.47	1	13.47	7.93	.006	.10
Distractor $\times$ Proficiency	4.55	2	2.28	1.34	.268	.04
Error (distractor)	122.29	72	1.70			
Total	334.65	149				
Item 4						
Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	<i>partial</i> η^2
Between Subjects						
Proficiency	30.64	2	15.32	6.22	.003	.15
Error	177.34	72	2.46			
Within Subjects						
Distractor	4.94	1	4.94	2.08	.154	.03
Distractor $\times$ Proficiency	3.84	2	1.92	0.81	.450	.02
Error (distractor)	171.33	72	2.38			
Total	388.09	149				

Item 5						
Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	<i>partial</i> η^2
Between Subjects						
Proficiency	70.30	2	35.15	14.88	<.001	.29
Error	170.14	72	2.36			
Within Subjects						
Distractor	2.07	1	2.07	2.32	.132	.03
Distractor $\times$ Proficiency	1.50	2	0.75	0.84	.435	.02
Error (distractor)	64.14	72	0.89			
Total	308.15	149				

Item 6						
Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	<i>partial</i> η^2
Between Subjects						
Proficiency	17.90	2	8.95	4.92	.010	.12
Error	130.99	72	1.82			
Within Subjects						
Distractor	0.04	1	0.04	0.02	.877	<.01
Distractor $\times$ Proficiency	25.73	2	12.86	7.09	.002	.17
Error (distractor)	130.61	72	1.81			
Total	305.27	149				

Item 7						
Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	<i>partial</i> η^2
Between Subjects						
Proficiency	49.64	2	24.82	11.93	<.001	.25
Error	149.85	72	2.08			
Within Subjects						
Distractor	2.82	1	2.82	2.34	.130	.03
Distractor $\times$ Proficiency	0.48	2	0.24	0.20	.818	<.01
Error (distractor)	86.61	72	1.20			
Total	289.40	149				

Item 8						
Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	<i>partial</i> η^2
Between Subjects						
Proficiency	8.26	2	4.13	3.30	.042	.08
Error	89.98	72	1.25			
Within Subjects						
Distractor	7.90	1	7.90	4.30	.042	.06
Distractor $\times$ Proficiency	1.54	2	0.77	0.47	.629	.01
Error (distractor)	118.76	72	1.65			
Total	226.44	149				

Item 9						
Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	<i>partial</i> η^2
Between Subjects						
Proficiency	83.59	2	41.79	20.55	< .001	.36
Error	146.45	72	2.03			
Within Subjects						
Distractor	6.48	1	6.48	4.76	.032	.06
Distractor $\times$ Proficiency	2.83	2	1.42	1.04	.359	.03
Error (distractor)	98.06	72	1.36			
Total	337.41	149				

Item 10						
Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	<i>partial</i> η^2
Between Subjects						
Proficiency	52.38	2	26.19	13.26	< .001	.27
Error	142.19	72	1.98			
Within Subjects						
Distractor	0.55	1	0.55	0.66	.419	.01
Distractor $\times$ Proficiency	1.13	2	0.57	0.68	.511	.02
Error (distractor)	60.16	72	0.84			
Total	256.41	149				

Item 11						
Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	<i>partial</i> η^2
Between Subjects						
Proficiency	36.22	2	18.11	6.10	.003	.15
Error	213.82	72	2.97			
Within Subjects						
Distractor	5.17	1	5.17	4.49	.038	.06
Distractor $\times$ Proficiency	8.00	2	4.00	3.47	.036	.09
Error (distractor)	82.99	72	1.15			
Total	346.20	149				

Item 12						
Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	<i>partial</i> η^2
Between Subjects						
Proficiency	93.71	2	46.85	18.95	< .001	.35
Error	178.07	72	2.47			
Within Subjects						
Distractor	5.05	1	5.05	6.99	.010	.09
Distractor $\times$ Proficiency	2.82	2	1.41	1.95	.150	.05
Error (distractor)	52.07	72	0.72			
Total	331.72	149				

Item 13						
Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	<i>partial</i> η^2
Between Subjects						
Proficiency	49.28	2	24.64	10.00	< .001	.22
Error	177.42	72	2.46			
Within Subjects						
Distractor	22.07	1	22.07	14.54	< .001	.17
Distractor $\times$ Proficiency	4.24	2	2.12	1.40	.254	.04
Error (distractor)	109.26	72	1.52			
Total	362.27	149				

Item 14

Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	<i>partial</i> η^2
Between Subjects						
Proficiency	34.61	2	17.45	11.11	<.001	.24
Error	113.09	72	1.57			
Within Subjects						
Distractor	1.71	1	1.71	1.30	.257	.02
Distractor $\times$ Proficiency	4.75	2	2.38	1.82	.170	.05
Error (distractor)	94.21	72	1.31			
Total	248.37	149				

Item 15

Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	<i>partial</i> η^2
Between Subjects						
Proficiency	20.24	2	10.12	4.99	.009	.12
Error	146.10	72	2.03			
Within Subjects						
Distractor	0.26	1	0.26	0.13	.724	.01
Distractor $\times$ Proficiency	10.06	2	5.03	2.42	.094	.06
Error (distractor)	148.28	72	2.06			
Total	324.94	149				