A Comparison of Distractor Selection Among Proficiency Levels in Reading Tests: A Focus on Summarization Processes in Japanese EFL Learners

Takahiro Terao1 and Hidetoki Ishii2

Abstract
This study aimed to compare selection patterns of distractors (incorrect options) according to test taker proficiency regarding Japanese students’ summarization skills of an English paragraph. Participants included 414 undergraduate students, and the test comprised three summarization process types—deletion, generalization, and integration. Within the questions, which represented summary candidates for a final version of a test, distractors were created reflecting typical student errors related to each summarization process. Six distractor types were tested. Results showed that distractors that were missing important information for the summary functioned well for determining low-, middle-, and high-proficiency students regarding deletion items. For generalization items, both distractor types, those containing examples and those with inappropriate superordinates, were attractive for low- and middle-proficiency students. Regarding integration items, it was found that distractors missing the author’s viewpoint in the summary were more attractive only for less-proficient students. Several tips to guide future item writing are provided.

Keywords
distractor, summarization, item development, typical errors

Introduction
In educational testing, test items play an important role in gathering evidence of students’ skills and abilities. Multiple-choice items are one of the most widely used item formats, through which evidence of students’ proficiency can be obtained from their chosen responses. To obtain rich information from test items and options, it is necessary to define the skills and abilities to be measured and reflect these components in the items and options.

Test development requires various investigations. Lane et al. (2016) pointed out that test development often involves 12 steps. The test development procedure includes several steps before item writing, including the overall planning of the test, content and construct definition and description, and item specification. These are all initial and essential parts of ensuring the validity of test items. Specifically, in construct definition and description, the exploration of the target construct is a principal and influential step that guides item writing. Construct definition and description are also called domain description and refer to how performance in the target domain is linked to the observation of performance in the test (Chapelle et al., 2008).

Test items must reflect only the targeted skills and abilities (Messick, 1995; Wilson, 2005) and assign scores that correspond with high and low proficiency. In multiple-choice items, each option reflects a part of the targeted skills; proficient test takers choose the correct option, whereas less-proficient ones choose distractors (i.e., incorrect options). For gathering evidence of students’ skills through multiple-choice items, it is necessary for the targeted skills and the option development to be connected with each other. Nonetheless, during the actual item development, test items may often miss important dimensions or contain other components related to the targeted skills in the distractors. When creating distractors, students’ common errors should be reflected (Haladyna & Rodriguez, 2013; Rodriguez et al., 2014). Common errors are mistakes that made frequently by

1 National Center for University Entrance Examinations, Tokyo, Japan
2 Nagoya University, Japan

Corresponding Author:
Takahiro Terao, National Center for University Entrance Examinations, 2-19-23, Komaba, Meguro-ku, Tokyo 153-8501, Japan.
Email: terao@rd.dnc.ac.jp
students in solving the item. Indeed, such errors are difficult to identify, and uncommon errors that include other components of mistakes not related to the targeted skills are often reflected in distractors. Distractors with uncommon errors are less attractive and rarely selected; therefore, less-proficient test takers are more easily able to choose the correct option. As a result, the item becomes relatively easy for low-proficient students, resulting in unpredicted easiness. Conventionally, distractors should be chosen by at least 5% of test takers (Haladyna & Downing, 1993). According to previous studies, test items with plausible distractors are not easier (DiBattista & Kurzawa, 2011; Drum et al., 1981; Tarrant et al., 2009; Ware & Vik, 2009).

Furthermore, test items with attractive distractors also revealed stronger discrimination power. The discrimination power of a test item means the degree of contribution an item makes in assigning higher scores to proficient test takers and lower scores to less-proficient ones (Haladyna & Rodriguez, 2013). To measure a skill or skills in educational testing, it is important that the test items provide accurate and representative test scores to students. The discrimination power of a distractor has an expanded definition of item discrimination power. The distractor discrimination power means that the selection of that distractor is related to lowering the test score.

In another case, when there are unrelated components to confuse test takers, the distractors must be tricky. A tricky distractor is sometimes chosen even by highly proficient test takers. In other words, a test item with a tricky distractor prevents even proficient test takers from selecting the correct option due to unrelated factors. Item development guidelines suggest that a tricky distractor should not be prepared (Haladyna & Rodriguez, 2013).

To prevent such situations, it is very important to perform statistical checks for distractors through a process called distractor analysis. To create construct-relevant distractors, test developers and item writers must understand the influence of common errors related to distractors and their attractiveness. However, the typicality of the errors cannot be examined before gathering student response data. In the initial phase of test development, exploratory and iterative empirical investigations of common errors and repeated revisions of construct descriptions are necessary.

Unfortunately, there have been few studies that investigate the relationship between common misconceptions related to target skills and distractor selections. While there is significant research on the effects of distractor characteristics—such as surficial and linguistic ones—on the selection of correct options (e.g., Freedle & Kostin, 1993), most studies on distractors addressed only the statistical characteristics of distractors (e.g., DiBattista & Kurzawa, 2011; Tarrant et al., 2009), without focusing on the connection between students’ typical errors. Therefore, it is necessary to study the types of errors that contribute to high discrimination power of distractors.

Several studies focused on students’ common errors, especially those in reading comprehension items (e.g., King et al., 2004). King et al. (2004) proposed the distractor rationale taxonomy (DRT) for characterizing the features of students’ errors included in distractors into four understanding levels (King et al., 2004). This theory was validated by evidence of a strong correlation between test scores and evaluation of the subject matter experts to the level of distractors (Lin et al., 2010). The DRT contributed to clarifying the meaning of the scores because of ordering the levels of students’ common errors. This theory, however, mainly focused on the local reading process. Although the ability to locally understand the meaning of sentences in the text is essential, the ability to understand a paragraph and the text level is also an important reading comprehension skill, especially for learners of English as a second language (L2 learners). L2 learners naturally comprehend texts through individual sentences and do not focus on the paragraph structure (e.g., Kozminska & Graetz, 1986). Moreover, many previous studies showed that L2 readers focus on the paragraph and text structure and improved their reading performance by learning the strategies of summarization (e.g., Hare & Borchardt, 1984; Karbalaei & Rajyashree, 2010; Khoshnevis & Parvinnejad, 2015). Based on this research, summarization skills need to be addressed during the development of reading items for L2 readers.

The aim of this study is to develop distractors with common errors made by L2 learners in summarizing a text and to investigate the attractiveness of these distractors for the test takers. This study comprised three components: (a) identifying common errors made in L2 summarization through a literature review, (b) editing distractors with such common errors, and (c) testing the attractiveness of such distractors. First, this study extracted the common errors in summarizations made by L2 learners from previous studies on summarization. We reviewed the theories on summarization in cognitive psychology. In addition, we reviewed several studies on the L2 characteristics of summarization in second-language learning. Second, we developed distractors along with their common errors for summarization items. We edited three types of test items focusing on the sub-processes of summarization: deletion, generalization, and integration. For each item type, we investigated two types of distractors that reflect students’ common errors in the sub-processes, based on previous findings. The characteristics of L2 summarization are presented in the following section. Third, we performed distractor analysis to examine how attractive distractors with common errors in each sub-process of the summarization were. In this analysis, the attractiveness of distractors was assessed for a comparison of the proportion of each distractor selection by proficiency level. We investigated the selection patterns of distractors by proficiency level through item and distractor analysis approaches as a descriptive research. The independent variable was proficiency groups (low, middle, and high proficiency), and the
dependent variable was the proportion of each distractor selected for each item type. Through this analysis, we revealed the attractiveness of each distractor in test items to measure L2 summarization skills.

This study has two significant merits for educational testing and measurement. The first advantage is that it clarifies the distinction between related and unrelated components of targeted skills. When including typical errors based on the cognitive theory in a distractor, test takers answer an item correctly when they complete the target process, and incorrectly when they make mistakes in that process. This study’s findings contribute to defining the related and unrelated components of the summarization skills of L2 learners. The second merit is that our findings may help item writers to create more functional distractors. While writing distractors, item writers may unconsciously include the irrelevant features of errors, making the distractors less attractive even for low-proficiency test takers, or tricky for those with high proficiency. Statistical analysis of incorrect options with focal common errors clarifies the types of errors that should be included in distractors, and how frequently selected are such distractors. These findings would promote effective development of items and distractors, which is necessary during the initial stage of the validation process.

Literature Review

Construct Representation and Distractors in Item Development

To connect the construct definition or description with distractor development, the concept of construct representation in the validity theory provides a useful framework. Construct representation is referred to as the identification of the cognitive mechanism that underlies test items (Embretson, 1983). When investigating construct representation, important dimensions related to the targeted skills are necessary to be identified and reflected in test items. However, in practice, two issues may negatively affect the test item validity. First, items may not fully reflect the target construct. This is called construct underrepresentation and indicates that a test lacks important dimensions or aspects that are closely related to the target skill (Messick, 1995). Alternatively, an item may contain factors that are not related to the target construct, called construct-irrelevant variance; these other factors can influence the test score and cause the item to be either easier or more difficult than intended (Messick, 1995). In the construct definition and development of valid test items, it is important to prevent construct underrepresentation, as well as to minimize construct-irrelevant variance.

Previous studies indicated that in a construct’s definition, an information processing model in cognitive psychology relevant to the target skills is utilized for defining the relevant and irrelevant components (Embretson & Gorin, 2001). Cognitive items prevent construct underrepresentation when the theory and findings capture students’ cognitive behaviors in demonstrating their abilities. In addition, test items based on cognitive theory contribute to theory-based interpretation, thus indicating a clear interpretation of test scores (Embretson & Gorin, 2001; Kane, 2013). While solving a cognitive-based item, test takers answer the item correctly when they complete the target process and incorrectly when they make mistakes in that process.

In the development of multiple-choice items, identifying students’ common errors and misconceptions based on cognitive theories and including them in the distractors play a significant role in creating cognitive-based items and making it easier to interpret test scores.

Distractors in Reading Items

Previous studies have investigated the characteristics of distractors in reading comprehension items (e.g., Drum et al., 1981; Freedle & Kostin, 1991, 1993; King et al., 2004; Lin et al., 2010; Ushiro et al., 2007). For example, the larger number of new content words in distractors could be easily falsified, leading to high p-values (easier) because test takers considered those words as not included in the text (Drum et al., 1981). Distractors with negation and/or lengthy distractors make an item difficult (Freedle & Kostin, 1993). Furthermore, Ushiro et al. (2007) investigated the distractor characteristics of reading items for L2 learners and found that distractors with text keywords and those containing more information were selected frequently by test takers.

Some studies on distractors that include students’ common errors in reading comprehension proposed the DRT (King et al., 2004; Lin et al., 2010). DRT defines the features of students’ errors included in distractors in terms of four understanding levels (King et al., 2004). Level 1 distractors contain errors that reflect a misunderstanding of the passage’s topic; Level 2 distractors include errors that reflect confusion regarding the relationship between sentences; Level 3 distractors have errors that reflect poor analysis or interpretation; and Level 4 is the correct option. Lin et al. (2010) developed experimental test items in an ordered multiple-choice format based on the DRT and revealed a strong correlation between the test scores and evaluation of the subject matter experts based on the levels of distractors. The DRT contributed to clarifying the meaning of the score, as it provides information on common errors made by students, orders them, and proposes an interpretation of the test scores. The DRT demonstrated the importance of extracting the common errors made by test takers.

Cognitive Theories on Summarization Skills

Summarization has been well researched in cognitive psychology. When students read and summarize a passage, whether they are first-language (L1) or L2 learners, a macrostructure is developed (Kintsch & van Dijk, 1978).
macrostructure is an integrative representation of the several propositions of a text. For integrating each proposition into a macrostructure, the following three rules (macrorules) are adopted: deletion, generalization, and integration. Deletion refers to selecting the important topic in a text and discarding redundant information. Generalization refers to replacing the subordinates (episodes and examples) with superordinates (more generic words and phrases that extract the essence of episodes and examples). Integration refers to developing a consistent proposition that represents a global fact of the text from a sequence of propositions. These processes are activated while reading a passage and constructing a coherent mental representation of the text. A. L. Brown and Day (1983) expanded this model to include the following six rules: (a) deleting trivial and unimportant contents or (b) redundant information even if it is important, (c) substituting actions or (d) events for a list of components, and thus, (e) selecting or (f) inventing a topic sentence, which is the author’s summary of the paragraph. The first two rules correspond to deletion, the middle two to generalization, and the final two to integration.

Since Kintsch and van Dijk’s research, additional studies have expanded the understanding of the integrated theory of reading and the utility for second-language learning. Like the three macrorules proposed by Kintsch and van Dijk (1978), Coffman (1994) categorized the three strategies used in summarization: reproduction, transformation, and intrusion. This theory is essentially similar to the three macrorules. A reproduction is a restatement of the original content, a transformation is a combination of multiple pieces of the content, and an intrusion is a statement regarding the writer’s opinion and the reader’s prior knowledge. While Coffman’s (1994) theory originally targeted primary school students, Kintsch and van Dijk’s (1978) three macrorules are applicable to L2 learning and comprehension (Johns & Mayes, 1990), and these models have been adopted in the L2 context. More recent studies examined the cognitive characteristics of summarization in L2 learners and tested the effect of learning summarization strategies on reading comprehension (Johns & Mayes, 1990; Karbalaei & Rajyashree, 2010; Keck, 2006; Kim, 2001; Mori, 2015). The three macrorules framework remains relevant to understanding the characteristics of an L2 summary; thus, this study also employed these macrorules and captured the erroneous features of L2 learners’ summarizations.

L2 Summarization Characteristics: Identifying Typical Errors

Ordinarily, summarization skill is often measured using constructed-response items, which do not contain options and require students to write their own answers (Haladyna & Rodriguez, 2013). Although students’ written responses to constructed-response items contain a vast amount of information about their thinking process, scoring their responses requires considerably more time and effort and often contains numerous human-related rating errors. In large-scale, high-stakes testing programs, gathering students’ written responses and scoring them precisely is not the best strategy for identifying their skills. In addition, it has been reported that a deficit of targeted skills is compensated by writing fluency when adopting the construct-response format (Breland et al., 2017). This study aims to resolve this problem by using distractors in a multiple-choice format to reflect students’ typical errors, thus separating these skills from writing fluency.

Some research on L2 acquisition has investigated the summarizing strategies used by L2 learners (e.g., Johns & Mayes, 1990; Keck, 2006, 2014; Kim, 2001) based on the cognitive models described above. From these studies, the two main characteristics of summarizing a text can be extracted. The first characteristic is applying textual borrowing, or copying (Johns & Mayes, 1990; Keck, 2006, 2014; Shi, 2004), a strategy that was originally called the copy-delete strategy (A. L. Brown et al., 1983). Johns and Mayes (1990) revealed that less-proficient L2 students applied direct copying to summary writing, whereas proficient L2 students did not. Shi (2004) and Keck (2006, 2014) showed that L2 learners used textual borrowing significantly more than L1 students. This result was observed regardless of the differences between the native languages of the L2 learners (Keck, 2006). Moreover, Japanese students formed a summary that was directly excerpted from a sentence in the original text (Terao & Ishii, 2018). These findings indicated that Japanese students tended to construct an inappropriate and generalized representation obtained from detailed information. Therefore, it is assumed that low-proficient test takers considered a summary that contained only detailed information to be correct.

The second characteristic of L2 summarization is the lack of integrative idea units involving several sentences (Johns & Mayes, 1990; Kim, 2001). Kim (2001) revealed that L2 learners tended not to include descriptions that unify propositions from several sentences in the paragraph. Terao and Ishii (2018) further revealed that Japanese students partially integrated the entire paragraph in their summarization. Combining numerous propositions from several sentences into an integrative phrase is related to both the generalization and integration processes. When generalizing a list of things, the episodes or examples stated in several sentences must be presented as more abstract propositions. When integrating the content of a paragraph, ideas from several sentences in the paragraph should be represented in the form of a single proposition. This feature was observed regardless of the test takers’ proficiency levels (Johns & Mayes, 1990); nonetheless, the frequency of this feature’s occurrence may be different by proficiency level. During the generalization process, the most proficient students often succeed in rejecting a summary using concrete examples, although several less-proficient students may consider it to be correct. In the integration
process, a large number of less-proficient students cannot develop a representative summary of the paragraph, whereas proficient students are able to.

Based on these findings, some common errors related to each process in summarization were found by Terao and Ishii (2018). They examined the summarization errors made by Japanese students in each process along with the construct-response summarization tasks. The researchers reported that the inclusion of unimportant information, lack of important information, and excessive deletion were observed in tasks that focused on the deletion process. Furthermore, the students' summaries in the generalization task contained excerpted detailed information, such as episodes or examples, along with generalization; in addition, partial descriptions of the target paragraph (not representative of the paragraph) and the inclusion of different viewpoints from the original text were demonstrated to be errors in the integration process. Of these features, the excerpting of detailed information should come from copy-and-delete strategy. Inappropriate generalization, partial descriptions, and the inclusion of different viewpoints may have been a result of the failure to unify a global proposition.

When including these characteristics in distractors, we predicted that less-proficient students would select these, whereas proficient students would not. Moreover, regarding deletion, there are insufficient findings for predicting the relationship between proficiency levels and distractor characteristics; therefore, this study exploratorily investigated this relationship.

**Method**

**Participants**

A total of 414 Japanese undergraduates (346 female, 68 male) participated in this study. Participants were recruited in public and private universities in Aichi prefecture in Japan. Of the total participants, 167 were from Kinjo Gakuin University, 109 from Aichi Shukutoku University, 75 from Nagoya University, and 63 from Chubu University. While this study could not ask the age of participants, most of university students in Japan are in the same age group and their age ranged from 18 to 22 years old. All of them were Japanese students and L2 learners of English. They participated in this study for course credit for an introductory course in psychology.

We divided participants into three proficiency groups based on the scores to understand distractor functions by discrete proficiency level. We used calibrated scores, or estimated thetas, of the six booklets described from the next section. We did not employ any other external tests because the construct to be measured in this study was different from other tests. Cut-off values were determined based on Kelley (1939), and the 27th and 73rd percentiles of distribution of scores were used. This method can produce three valid groups with sufficient sample size in every group and contrasting differences in proficiency levels (Kelley, 1939). After the experiment, 119 participants were classified into the low-proficiency group, 187 in the middle-proficiency group, and 109 in the high-proficiency group.

**Procedure**

One of the six test booklets were randomly assigned to each participant. The resulting number of participants assigned to each booklet was as follows: 72 participants answered Booklet 1, 68 answered Booklet 2, 68 answered Booklet 3, 69 answered Booklet 4, 68 answered Booklet 5, and 69 answered Booklet 6. We provided explanations of the study and obtained informed consent from each participant, including data use. The test was administered by paper-and-pencil. Participants had 30 min to complete the booklet.

**Instruments**

**Test design.** To investigate functions of distractors in as many items as possible, this study developed seven testlets, each with one English passage and the corresponding test items. Common-item design was employed to create six booklets. Each booklet had two sections; one was a common-item section, and the other was a unique-item section. Each testlet packaged one English text and three test items. Each of the three items was experimentally developed to measure the subskills of summarization: deletion, generalization, and integration. Below the test section, we gave detailed information on the test in this study from two points of view: (a) characteristics of English passages used in this study and the selection criteria and (b) the methodology of question and option development of three types of test items.

**English passages.** We employed seven English passages that had previously been administered in entrance examinations at Japanese universities. We evaluated a variety of passages using the Flesch-Kincaid readability index to determine the text's difficulty level and selected the seven texts within the appropriate difficulty level for participants in this study. The Flesch-Kincaid readability score was calculated as a function of the total number of sentences, total word count, and total number of syllables. The index corresponds to U.S. grade levels for which the texts are appropriate. According to previous studies that investigated the text's difficulty level in reading tests of the university entrance examinations in Japan, the Flesch-Kincaid index of English texts ranged from 10 to 12 (J. D. Brown & Yamashita, 1995; Kikuchi, 2006).

Detailed information about the passages used in this study is described below in terms of word count, topic, and Flesch-Kincaid readability. Passage 1 was a 742-word text about the history of homesickness and the relationship between modern technology and homesickness; Flesch-Kincaid readability was 9.48. Passage 2 was a 508-word text about language, that is, the difficulty of recording and understanding...
non-English culture in English; Flesch-Kincaid readability was 13.83. Passage 3 was a 755-word text about mysterious features of human behavior; Flesch-Kincaid readability was 8.42. Passage 4 was a 920-word text about reasons Dutch people love cheese; Flesch-Kincaid readability was 8.10. Passage 5 was an 876-word text on psychological issues in the Digital Age; Flesch-Kincaid readability was 11.84. Passage 7 was a 643-word about veganism; Flesch-Kincaid readability was 8.64.

Passage 7 (veganism) and test items attached to this passage were included in the common-item section of all six booklet versions. Passages 1 to 6 and their related test items were presented in the unique-item section in Booklets 1 to 6, respectively.

Test items. Three types of test items that were focused on a specific macrorule were edited in this study. There are two main points shared among three kinds of items. First, all items asked participants to understand the designated paragraph’s main idea and choose the appropriate summary. In the question part of the test item, we asked participants to read a specific paragraph in the text which assumed to activate the targeted subskill. The stem in every type of question was described as “Which of the following statements best summarize the paragraph (X)?” Second, each item had three answer options: one correct answer and two distractors.

The first type of test item was the deletion item and required test takers to judge the importance of a topic in the paragraph. To elicit the deletion process, we chose a paragraph that included many supporting sentences related to the topic and required participants to evaluate the important pieces of information in the paragraph. The second type was the generalization item and tasked students with understanding the gist of the paragraph by replacing subordinates with superordinates along with the topic sentence. To activate the generalization process, we selected a paragraph that included detailed information such as examples and episodes to extract an essence. The last type was the integration item and asked students to evaluate the representativeness of the statement about which summary described the intent of the text’s author in the paragraph. We picked a paragraph in which the topic sentence was clearly stated. At the end of test item development, we constructed seven packages that included one passage and three experimental items, including deletion, generalization, and integration items for each.

Under each three types of test questions, one correct and two distractors were edited. Correct options were developed to represent the appropriate result of summarization to activate the targeted process in those items. In a deletion item, a correct summary was written to contain only necessary pieces of information and to exclude unnecessary or redundant ones. In a generalization item, a correct option was edited to express a more generalized description based on detailed pieces. In an integration item, a correct summary was authored to represent the topic sentence of that paragraph.

Each two distractors in deletion, generalization, and integration items were created to reflect typical errors in L2 learners’ summarization process, as shown in Terao and Ishii (2018). For deletion items, two different types of distractors were prepared: one summary missing important information and one with unnecessary information. For generalization items, one summary describing episodes or examples from the passage word-for-word and one using inappropriate expressions were prepared as distractors. For integration items, one partial summary (i.e., one that does not include content that is integrated throughout the paragraph) and one with a description different from the author’s viewpoint were used as distractors. Examples of each type of test item are presented in the Appendix.

Research Design and Data Analysis

This study investigated functions of each distractor and tested the difference of the proportion of distractor selection between proficiency groups. The independent variable was proficiency groups (low, middle, and high proficiency), and the dependent variable was the selection of each distractor. Distractor analyses were performed to check functions of distractors with students’ common errors focusing on a specific process of summarization.

The data analysis comprised five parts. First, the descriptive statistics of the total raw scores were calculated. Second, we applied the one-parameter logistic model (OPLM) to calibrate the proficiency estimate of students who answered different booklets. Third, item and option statistics were calculated and examined to check psychometric characteristics such as difficulty and discrimination power. Fourth, a graphical check of distractor functioning in each item type was performed. Finally, the difference in choice ratios of distractors between proficiency groups was statistically investigated. The analytic steps were as follows.

First, participants’ responses were dichotomously scored for each item (1 for correct and 0 for incorrect) and the characteristics of total raw scores were assessed. A histogram of the total score was depicted and the mean, the standard deviation, and the minimum and maximum raw scores were evaluated.

Second, we applied the OPLM in item response theory (IRT) to estimate the latent ability $\theta$ for each participant. Every booklet in this study included different sets of experimental items, so comparing total raw scores of each booklet directly was inappropriate. In such circumstances, test takers’ ability should be evaluated independently of the booklet assigned. We utilized an expected a posteriori (EAP) estimation with standard normal distribution as a prior distribution of the population of $\theta$. We used common item equating to calibrate latent ability scales from all six booklets, and
concurrent calibration was applied to the overall data matrix, in which the score of items not included in the booklet assigned to the participants was missing values.

Third, we calculated the descriptive statistics of option selection and discrimination power, called item analysis. Option selection was evaluated by the proportion of participants who selected that option for all the participants of this study. Option discrimination power was evaluated with the point biserial correlation between estimated thetas and option scores (1 for selecting that option, and 0 for not selecting). Positive discrimination power for correct options would indicate that test takers with higher proficiency chose correct options and those with lower proficiency did not. In contrast, negative discrimination power for distractors would demonstrate that students with higher proficiency did not select these answers, but low-proficient ones did. Point biserial correlations are usually calculated using raw total scores, but due to the reason stated above, we used estimated thetas instead.

The last two steps of analysis were designed to compare distractor selection among test takers’ proficiency levels. Distractor selection by proficiency level was evaluated by graphical and statistical means. Before these analyses, we divided participants into three groups by proficiency level, based on the estimated scores in OPLM: low proficiency \((n = 119)\), middle proficiency \((n = 187)\), and high proficiency \((n = 109)\).

The fourth step was the graphical examination of distractor selection, and we depicted trace lines to visualize patterns of option selection for each item type. Trace lines are a set of line graphs to plot choice ratios of each option by proficiency level. In the graph of trace lines, discrete proficiency groups were in the horizontal axis, and choice ratios of each option were in the vertical axis. The choice ratio of the option was calculated for each proficiency group and evaluated by the proportion of participants who chose that option in a specific proficiency group. Haladyna and Rodriguez (2013) and Lane et al. (2016) introduced trace lines as a useful tool to examine distractor selection. This analysis is regularly employed to check patterns of option selection item by item. While this framework was also useful for this study, some minor points needed to be modified. This study examined several types of distractors related to cognitive processes in summarization, and the same type of distractors were edited in several items. To analyze overall functions of every type of distractor, we merged their responses to the items so that skills could be measured together, and then depicted trace lines by item type (deletion, generalization, and integration items).

The last step was to evaluate local odds ratios (ORs) to examine distractor selection among adjacent proficiency groups. Local OR is an extended version of OR for a \(r \times c\) contingency table (\(r\) means the number of rows, \(c\) means the number of columns, and both are greater than 2). When constructing local ORs, four arbitrarily selected cells in a contingency table are used. In this study, odds were constructed for each proficiency group such that denominators were choice ratios of correct options, and numerators were those of certain types of distractors. Local ORs were constructed with odds for adjacent proficiency groups. In comparing distractor selection between low- and middle-proficiency groups, we determined local ORs with odds for the middle-proficiency group as a denominator and those for the low-proficiency group as a numerator. In this case, local ORs were larger than 1 when that distractor was more attractive in the low-proficiency group. Size of local ORs represents the discrimination power of distractors between adjacent proficiency levels. Based on Fagerland et al. (2017), 95% confidence intervals (CIs) of local ORs were calculated.

Overall statistical analyses were implemented by R 3.5.0 (R Core Team, 2018). Specifically, an estimation of participants’ proficiency \(\theta\) using OPLM and a calculation of local ORs and CIs were performed with additional packages. Itm (Rizopoulos, 2018) and irtoys packages (Parchev, 2017) were utilized in applying OPLM, and vcd packages (Meyer et al., 2017) were adopted in estimating local ORs.

Results

Descriptive Statistics and Histogram of the Total Raw Score

The score’s histogram is presented in Figure 1. The minimum was 0 and the maximum was 5. The mean of the total raw score was 2.13, and the standard deviation was 1.24. Considering that participants only answered an average of two items correctly, test items were somewhat difficult for them. Figure 1 also showed that the distribution was slightly right-skewed and difficult items were aligned in the test.

Fitting OPLM to Estimate Participants’ Ability

We applied OPLM to dichotomously scored responses and estimated ability for each test taker. The mean of estimated thetas was 0.00, and the standard deviation was 0.34. Histogram of the estimates of thetas is shown in Figure 2.

Item Analysis

Using estimated thetas, difficulty and discrimination were calculated for every option in deletion items (Table 1), generalization items (Table 2), and integration items (Table 3). Discriminations for correct options were sufficient and well-operating, and all distractors were selected by more than 5% of the participants. Some distractors had a point biserial correlation near zero and had weak discrimination power. Negative distractor discrimination was desirable because that distractor was selected more often by test takers with low proficiency and only occasionally chosen by proficient test takers (Attali & Fraenkel, 2000; DiBattista & Kurzawa, 2011). In this study, we continued to include options with negative discrimination in later analyses.
Trace Lines and Inferences on Local ORs to Examine Distractor Selection

To examine distractor selections relative to the correct options among proficiency groups, trace lines were drawn, and local ORs among proficiency groups were constructed by item type (i.e., deletion, generalization, and integration items). Figure 3 shows trace lines for deletion, generalization, and integration items.

Deletion items. Trace lines for deletion items (the left panel of Figure 3) show that both types of distractors went down to the right, meaning both distractors were much more attractive for the low-proficiency group than the high-proficiency group. The slope of distractors that lacked necessary information was steeper than that of distractors with unnecessary information.

The local OR of distractors without necessary information between low- and middle-proficiency groups was large (OR = 4.709, 95% CI = [2.396, 9.255]), and the local OR between middle- and high-proficiency groups was slightly large (OR = 4.318, 95% CI = [2.241, 8.321]). It was indicated that distractors without necessary information for the paragraph summary were judged to be more plausible by test takers at the lower proficiency level. In turn, the local OR of distractors with unnecessary information between low- and middle-proficiency groups was large (OR = 3.526, 95% CI = [1.750, 7.104]), and the local OR between middle- and high-proficiency groups was moderate (OR = 2.707, 95% CI = [1.493, 4.908]). This result showed that this type of distractor was much more attractive for the low-proficiency group than the middle-proficiency group. Overall, distractors without necessary information had more discrimination power than those with unnecessary information.

Generalization items. Trace lines for generalization items (the middle panel of Figure 3) show that selection patterns of both distractors were quite similar: choice ratios in low- and middle-proficiency groups were at the same level, while those in the high-proficiency group were low. It was shown that both distractors were attractive for low- and middle-proficiency test takers.

The local OR of distractors with episodes and examples between middle- and high-proficiency groups was very large (OR = 6.077, 95% CI = [3.223, 11.458]), while it was smaller between low- and middle-proficiency groups (OR = 1.910, 95% CI = [1.102, 3.608]). Similarly, the local OR of distractors with inappropriate superordinates was larger between middle- and high-proficiency groups (OR = 5.755, 95% CI = [2.958, 11.195]) than between low- and middle-proficiency groups (OR = 2.149, 95% CI = [1.125, 4.107]). These results indicate that both distractors were more attractive to test takers at low to middle proficiency, while they were not attractive to high-proficiency students. Either type of distractor had discrimination power between middle- and high-proficiency groups.

Integration items. Trace lines for integration items (the right panel of Figure 3) show that choice ratios of both distractors went down to the right, and slopes were somewhat mild. Distractors with differing viewpoints from the author’s were shown to be more plausible for test takers at the lower proficiency level. In turn, the local OR of distractors with unnecessary information between low- and middle-proficiency groups was large (OR = 3.526, 95% CI = [1.750, 7.104]), and the local OR between middle- and high-proficiency groups was moderate (OR = 2.707, 95% CI = [1.493, 4.908]). This result showed that this type of distractor was much more attractive for the low-proficiency group than the middle-proficiency group. Overall, distractors without necessary information had more discrimination power than those with unnecessary information.
tendency was observed for distractors with a different viewpoint from the author’s: the local ORs were larger between low- and middle-proficiency groups than between middle- and high-proficiency groups (OR = 4.038, 95% CI = [2.067, 7.892] for a comparison with low and middle; OR = 2.346, 95% CI = [1.251, 4.398] for a comparison with middle and high). Both types of distractors including errors in the integration process were attractive only for test takers with low proficiency.

**Discussion**

This study examined selection of distractors on a reading summarization test by student proficiency level. Results showed that all types of the distractors addressed in this study were well-functioning. These results also provide guidance on future item writing to achieve the development of valid distractors and test items.

**Distractor Selection by Item Type**

**Deletion items.** We examined two types of distractors in deletion items to measure students’ judgment about essential information in each paragraph. One distractor was a summary that lacked necessary information and the other was a summary that included unimportant information. Results showed that both kinds of distractors were functioning, and distractors without necessary information had particularly strong discriminating power. Hare and Borchardt (1984) discussed sensitivity to importance, which is required to delete unimportant ideas. Low-proficient test takers might be insensitive to the importance of the contents in the paragraph and choose a distractor without necessary information. The results of this study indicate that identifying the paragraph’s necessary information was an important piece of summarization skills in both constructed-response and multiple-choice formats due to the distractor’s strong discrimination power.
This study also revealed that distractors with unnecessary information in a summary had weaker discrimination power and were relatively less attractive for test takers in the low-proficiency group and rather attractive for the proficient group. Such a result might be obtained partly because proficient students struggled to evaluate whether a summary was correct. Although most proficient participants may be able to identify necessary information appropriately, some proficient students thought redundant information should be excluded from a summary while others thought that a summary with redundant information was acceptable because it also had enough information to represent the paragraph. When developing test items to measure deletion skills, item writers should identify important and unimportant propositions in advance and develop distractors that are missing important pieces of information. When distractors are used with unnecessary contents in a candidate summary, the definition of an ideal summary should be clearly announced to test takers.

Generalization items. The following two distractors were tested in generalization items: distractors that repeated episodes or examples from the text word-for-word and those with inappropriate expressions to generalize examples or episodes. Results indicated that selection patterns of these two distractors were very similar, and the choice ratios were much higher in test takers with middle proficiency than in those with high proficiency. As previous studies suggested, low-proficiency students adopted a textual borrowing strategy in summary writing (Johns & Mayes, 1990; Keck, 2006, 2014). When answering multiple-choice items, lower- and middle-level test takers seemed to use this strategy to understand the gist of the paragraph with examples and episodes. Results of these distractors indicated that students in low- and middle-proficiency groups might represent the gist of the paragraph with borrowed sentences and choose a distractor summary with detailed information such as episodes or examples. Students with low- and middle-proficiency employed copy-and-delete strategies in summarization or main idea comprehension, even without the production of a summary.

At the same time, other students with low and middle proficiency selected another distractor summary with inappropriate generalization. Some students in these groups could evaluate a distractor summary full of detailed information as incorrect and avoid choosing this option but could not evaluate the appropriateness of generalization correctly. As described above, inappropriate generalization of subordinates seems to be an unsuccessful result of unifying multiple propositions in the generalization process. Kim (2001) also
mentioned that Korean L2 students had few opportunities to use transformation (i.e., a combination of multiple contents from the original text) so may not have learned such strategies. When answering generalization items in this study, participants with low and middle proficiency may also struggle to unify detailed descriptions.

From the examination of distractor analysis for generalization items, it was suggested that although two types of errors seemed to have a distinguished feature, identifying detailed information such as episodes or examples and unifying them to represent a more general proposition were closely related. Students can answer generalization items correctly after a success both in detailed information identification and unification. When developing distractors of generalization items, the identification and unification of detailed information should be reflected separately in distractors.

**Integration items.** Two types of distractors were investigated for integration items. One included only a partial description of the passage while the other contained a viewpoint different from the author’s in candidate summaries. Integration items with these distractors asked participants to evaluate the representativeness of summaries of the designated paragraph. Results showed that choice ratios of these distractors were much higher in the low-proficiency group than in the middle-proficiency group, while relatively smaller differences were observed between middle- and high-proficiency groups. Previous studies have revealed that L2 students have difficulty integrating propositions from several sentences into a single sentence when writing a summary, and this characteristic was observed regardless of proficiency level (Johns & Mayes, 1990). The results of this study, however, indicate that the selection pattern of distractors with errors in the integration process differed by proficiency level. Indeed, these types of distractors were less attractive for middle- and high-proficiency students but were attractive enough for low-proficiency students. Moreover, results indicated that distractor selection related to the overall proficiency estimates from item analysis. It was also shown that most of distractors in the integration item were chosen by lower-proficient students and were not by proficient ones.

While similar tendencies were observed between two types of distractors, the size of distractor discrimination powers differed between them. Distractors with different viewpoints had larger discrimination power than those with partial descriptions. As trace lines showed, distractors with different viewpoints were less attractive for proficient test takers. When a summary contains different viewpoints, most proficient students could find it incorrect in terms of the representativeness. On the other hand, when a summary includes only part of the paragraph’s contents, a small proportion of proficient test takers misjudged the representativeness.

When multiple-choice items are developed to measure an integration skill, it can be useful to include a different viewpoint from the author’s in a candidate summary to determine which students have either lower- or middle-level proficiency. If test takers in the population are relatively proficient, however, these types of distractors may not function well. Assessing the students’ current level will be beneficial in targeting the test and determining which types of distractors are necessary to prepare in the item.

**Implications**

This study obtained useful findings on distractor characteristics for the measurement of L2 learners’ summarization skills and makes an important contribution to future test-item writing. This study provides some empirical evidence of typical errors in summarization concerning each macrorule (deletion, generalization, and integration) and those statistical characteristics by their proficiency levels. Results of this study have three implications in item development and understanding common errors of students.

First, findings of this study can help item writers edit distractors effectively when creating test items to measure L2 summarization skills. In the development of reading items, there are some situations where it is appropriate to measure comprehension of the main idea in a paragraph. For example, the skill to judge which information is important in the paragraph would often be measured in a reading section. It might be effective to prepare the list of candidate features of distractors with those psychometric features when editing distractors to ensure that they are reflective of common errors related to the subskill. This study presented evidence for developing guidelines for clarifying the connection between distractor characteristics and statistics in L2 summarization skills toward item writers.

Second, this study provided a foundational evidence not only for manual item writing, but also for automatic item generation. Recent studies have tackled with distractor generation in an algorithm-based way (Chen et al., 2006; Gao et al., 2019; Mitkov et al., 2006, 2009; Susanti et al., 2017). While most of these studies mainly focused on vocabulary and grammar items, a few of them dealt with reading items (Gao et al., 2019). Gao et al.’s (2019) approach was a semantic similarity, not using students’ common errors. An approach based on common errors can be useful in an algorithm-based generation of distractors and may be integrated with natural language processing (NLP) approaches. Indeed, Chen et al. (2006) succeeded in addressing a common error structure in the algorithm-based generation of distractors in vocabulary items. This study offers additional information to algorithmically generate distractors with common errors.

Third, we also obtained some validity evidence on L2 students’ cognition in summarization. Findings on the moderate relationship between students’ cognitive characteristics embedded in distractors and discrimination powers should contribute to represent the construct or summarization skills more appropriately and minimize construct-irrelevant variance. For construct representation, identifying
summarization errors which are made frequently by students will contribute significantly to ensure that important features of these skills are not ignored. For construct-irrelevant variance, this study could suggest potential characteristics of distractors that were turned into tricky ones for L2 learners.

Limitations and Future Directions

One of the limitations in this study is that only Japanese undergraduates were examined as L2 learners; however, Japan is a country in which almost everyone uses Japanese as their official language. To generalize the findings of this study, responses of test takers from other countries, where the official language is English but the native language is not, should be examined. Another limitation is that separating three macrorules is somewhat unnatural in the process of developing test items. This study addressed the unique characteristics of macrorules; however, deletion, for example, may be a prerequisite of generalization and integration. Future studies should examine appropriate item specification to assess summarization skills.

When developing test items, test developers should examine distractors through many approaches; for example, to grasp test takers’ cognitive processes via test items or distractors, various statistical and measurement models have been developed, such as cognitive diagnostic models (de la Torre, 2009) and evidence-centered design, or Bayesian networks (Mislevy et al., 2003). These frameworks may also be useful in analyzing students’ proficiency and misconceptions. Application of various approaches and establishment of distractor development guidelines should be useful for stakeholders in testing, such as item writers, test developers, and educators.

Conclusion

Writing distractors for multiple-choice items requires time and effort mostly because there is a limitation to imagining test takers’ common errors. This study contributed to a framework to understanding common errors in summarization and to promote writing distractors more effectively. We hope to improve the effectiveness of distractor development in measuring English reading proficiency of L2 learners.

Appendix

Examples of Experimental Test Items

Deletion item

Paragraph. The global desire to leave home arises from poverty and necessity, but it also grows out of a belief that such mobility is possible. People assume that individuals can and should be at home anywhere in the world. This view is now accepted as central to a globalized economy. It leads to opportunity and profits, but it also has high psychological costs. In nearly a decade’s research into the emotions and experiences of immigrants, I’ve discovered that many people who leave home in search of better prospects end up feeling displaced and disappointed.

Options
a. As the global mobility becomes available, more and more people who need to earn money travel abroad easily, but some of them feel lonely living apart from their family members.
b. According to the survey on emotional experiences of immigrants for about 10 years, most of them who met their necessity to succeed in another country thanks to the global mobility suffer from homesickness.
c. Due to the easiness of global transportation, people who are poor and distant from members of their family to find more profitable work in another country tend to get opportunity and earn much money.

Generalization item

Paragraph. It is possible that these new technologies actually make feelings of displacement stronger. Maria Elena Rivera, a psychologist in Mexico, believes technology may increase homesickness. Her sister, Carmen, had been living in the United States for 25 years. With the rise of inexpensive long-distance calling, Carmen was able to phone home with greater frequency. Every Sunday she called Mexico and talked with her family, who routinely gathered for a large meal. Carmen always asked what the family was eating and who was there. Technology increased her contact with her family, but also brought a regular reminder that she was not there with them.

Options
a. Modern technology that enables us to contact regularly with distant people remind immigrants of their family, and stronger feelings of displacement occur to them.
b. Cheaper international calling made Carmen frequently call her family to know what they are doing, think of them all the time, and therefore feel alone and lonely.
c. Recent information technology made immigrants to the United States get homesick because a regular phone call with lower price feel that they are separated from members of their family.

Integration item

Paragraph. Homesickness is a pain that has been around since ancient times. In the 19th century, millions of immigrants to America suffered from it. But by the beginning of the 20th century, the technology of the day was seen as the solution to the problem. At the time, some Americans claimed that serious cases of homesickness had grown less common because of the speed with which mail could be
sent and news could be spread around the world. In reality, though, homesickness continued to affect many immigrants.

**Options**

a. Homesickness has existed in immigrants even if information technology advanced in the beginning of the 20th century and a regular contact with their family member became possible.

b. Advanced technology made it possible to talk with member of our own family all the time and to reduce the number of immigrants who suffered from homesickness seriously.

c. Owing to the advance in technology early in the 20th century, people can communicate with their family constantly and homesickness seemed to be less common, but nowadays immigrants tend to suffer from it again.

**Declaration of Conflicting Interests**

The author(s) declared receipt of the following financial support to the research, authorship, and/or publication of this article: This work was supported by Japan Society for the Promotion of Science (grant number 2216J05973).

**ORCID iD**

Takahiro Terao https://orcid.org/0000-0003-3839-1708

**References**


Ware, J., & Vik, T. (2009). Quality assurance of item writing: During the introduction of multiple choice questions in medicine for high stakes examinations. Medical Teacher, 31, 238–243. https://doi.org/10.1080/01421590802155597


Author Biographies

Takahiro Terao is an assistant professor of Research Division in the National Center for University Entrance Examinations. His research interests include item and distractor development considering students’ cognitive process and test validation, as well as computer-based testing and assessment, and automatic item generation with the strong theory.

Hidetoki Ishii is the director of Admission Division and a professor of Department of Education and Human Development in Nagoya University. His research interests include item analysis and test development, especially for school teachers to create appropriate tests. He also has concern on teaching data science and statistics to students in humanities.