frontiers in Education

ORIGINAL RESEARCH
published: 11 February 2020
doi: 10.3389/feduc.2020.00009

Difficulty of Summarization Items for Japanese Learners: Effects of Passages, Distractors, and Response Formats

*Takahiro Terao**

The National Center for University Entrance Examinations, Tokyo, Japan

This study aims to examine factors affecting the difficulty of summarization items for Japanese learners. In the process of item development, creating a connection between cognitive features related to target construct and the difficulty of test items is necessary to define the abilities to be measured. Previous studies have mainly focused on local reading comprehension, while this study addressed summarization skills at the paragraph level. The study originally developed items for an experiment that elicited three macrorules of the paragraph and text: deletion, generalization, and integration. This study evaluated the influence of passages, distractor characteristics central to summarization processes, and response formats on item difficulty, using item difficulty modeling. When editing distractors, characteristics of L2 learners' summarization were carefully reviewed and reflected. The participants included 150 freshmen from Japan, who were asked to answer experimental summarization items. The results of the linear logistic test model (LLTM) indicated that the main source of difficulty in summarization items was distractor characteristics. In particular, summaries with unnecessary information or lacking necessary information increased the level of difficulty. In addition, summaries with detailed information, such as episodes and examples, and with a viewpoint different from the author's, also increased difficulty. The effect of passage differences was found to be minimal. A difference in response formats moderately affected item difficulty, and the extended-matching format was slightly more difficult than the conventional multiple-choice format. This study suggested that test developers and item writers should pay attention to distractor development, to limit students' errors when measuring summarization skills of L2 learners.

Keywords: item difficulty modeling, summarization skills, passages, distractors, response formats, item development

OPEN ACCESS

Edited by:
Okan Bulut,
University of Alberta, Canada

Reviewed by:
Jinnie Shin,
University of Alberta, Canada
Joseph A. Rios,
University of Minnesota Twin Cities, United States

***Correspondence:**
Takahiro Terao
terao@rd.dnc.ac.jp

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *17 October 2019*
Accepted: *22 January 2020*
Published: *11 February 2020*

Citation:
Terao T (2020) Difficulty of Summarization Items for Japanese Learners: Effects of Passages, Distractors, and Response Formats. Front. Educ. 5:9. doi: 10.3389/feduc.2020.00009

INTRODUCTION

Automatic item generation (AIG) is a promising methodology to reduce the cost and effort of human item writers and to create test items systematically (Gierl and Lai, 2012). In the AIG framework, test items are generated from an item model, which includes manipulable elements in the item stem and options. Such elements are divided into two types: radicals and incidentals (Gorin, 2005). Radicals are the essential components of cognitive modeling or variables related

to the target construct, and produce a change in cognitive processing (Bejar, 1993; Gorin, 2005). Incidentals are the surface characteristics of test items; even if the incidentals were to be manipulated, the difficulty of test items would not change. Item generation with the manipulation of incidentals, called the weak-theory approach, is easier and more realistic, but it has been pointed out that the difficulty of generated items cannot be predicted accurately, and a family of generated items looks very similar (Gierl and Lai, 2012). In contrast, item generation with the manipulation of radicals, called the strong theory approach, is much more difficult because there are few theories to identify radicals; thus, there are a limited number of applications of item generation in which radicals can be manipulated. Applying the strong theory in item generation can contribute not only to the prediction of item difficulty, but also to systematic development of valid test items, to maintain an item bank effectively, and in the future to reduce the cost of pilot testing. Of course, application of the strong-theory approach is beneficial to systematic item development by human item writers.

The strong theory approach requires the development of cognitive models related to targeted skills and the identification of candidate radicals in that model. Previous studies introduced examples of original cognitive models established by subject matter experts in medical education (e.g., Pugh et al., 2016). An alternative methodology can be applied for other skills such as reading comprehension; extracting radicals from theories and findings in cognitive psychology. For reading comprehension, theories and findings have been accumulated in cognitive psychology (e.g., Kintsch and van Dijk, 1978; Brown and Day, 1983). While those theories and findings may not be directly applicable toward item development and generation, such rationale are useful to examine radicals in test items. Concerning the reading comprehension skills of English as a second language (L2) learners, studies on English learning may also support the understanding of radicals. Unfortunately, a limited number of studies present cognitive models on item development and generation using cognitive psychology (e.g., Embretson and Gorin, 2001; Gorin, 2005).

This study focused on one important aspect of reading comprehension skills in L2 learners: summarization skills. Pearson et al. (1992) summarized features of expert and novice readers' strategies and considered the selection of important ideas to be a reading skill. In general, L2 learners read texts by focusing on individual sentences, and do not focus on paragraph structure (e.g., Kozminsky and Graetz, 1986). Many previous studies have shown that summary writing training sessions and instructions containing summarization strategies make readers focus on paragraph and textual structure, and improve L2 learners' reading performance (e.g., Hare and Borchardt, 1984; Karbalaei and Rajyashree, 2010; Khoshnevis and Parvinnejad, 2015). While two of these studies obtained such results only from Iranian students, Hare and Borchardt (1984) collected participants from a variety of backgrounds. These findings imply that summarization skills may be a universally important dimension of L2 learners' reading skills. Despite the importance of these skills, there are no studies that examined the source of test item difficulty when measuring summarization skills. Cognitive theories on summarization processes in L2 learners assist in the understanding of radicals and incidentals in summarization items. When developing summarization items, three main factors should be considered in assessing difficulty: English passages, options, and response formats. This study aimed to examine whether these factors have an impact on the difficulty of summarization items for L2 learners, and to identify radicals and incidentals of summarization items.

This study makes two major contributions toward the field of educational measurement; (a) to provide a framework of summarization skills for L2 learners, and (b) to help item writers systematically create multiple-choice summarization items. Theories and findings on summarization processes may benefit the development of a cognitive model in item writing and generation, and the identification of radicals related to cognitive models promotes item development in the strong theory approach. In addition, item writers may be concerned about the effect of task characteristics on item difficulty when editing questions. The primary task-related characteristics of summarization items are passages, distractors, and response formats. An investigation of the source of difficulty of summarization items must help item writers understand which components of the cognitive model increase or decrease difficulty, and by how much. This finding may also contribute to reducing discarded items that do not meet statistical criteria in pilot testing.

BACKGROUND

Assessing Construct Representation in Item Development

In educational testing, test item development plays an important role in gathering evidence of student's skills and abilities. Item and test development require extensive investigation of the items and tests themselves, and include processes such as defining the content, developing a scoring rubric, constructing test assembly, evaluating item statistics, and so on. Lane et al. (2016) notes that 12 steps were often experienced in test development. The test development procedure occurs several steps before item writing, and consists of the overall plan of the test, content and construct definition and description, and item specification; these are the first and most essential parts of ensuring test item validity. Specifically, in the process of construct definition and description, identifying the construct that is measured is crucial to guiding item writing.

A test item should reflect only the targeted skills and abilities, and assign a score that corresponds with high or low proficiency (Wilson, 2005). Such items are considered to accomplish construct representation (Embretson, 1983). However, in practice, the following situations often arise, resulting in a loss of validity: (1) the test items do not fully reflect the target construct, or (2) they contain other factors unrelated to the target construct. The first is known as construct underrepresentation, and indicates that a test lacks an important dimension or aspect that is closely related to the target skill (Messick, 1995). Construct underrepresentation is often caused

by an unclear distinction between the component that is central to the targeted skill and components that are irrelevant. To ensure construct representation, a clear definition of the target skill is necessary (Haladyna and Rodriguez, 2013). The second situation is known as construct-irrelevant variance, and indicates that other factors may also influence a test score. Construct-irrelevant variance results from construct-irrelevant difficulty and easiness (Messick, 1995). When developing valid test items, it is important to prevent construct underrepresentation and minimize construct-irrelevant variance.

In particular, an investigation of construct representation provides a useful framework for systematic item writing. Under the strong theory of AIG, radicals are an important primary source of item difficulty. In other words, defining radicals contributes to increasing construct-relevant variance. While various construct-irrelevant factors exist, such as the demographic attributes of test takers and differences in educational system and native language, construct-relevant factors are often a smaller number of variables that elicit targeted skills. It should be straightforward to manipulate radicals to increase construct-relevant variance rather than to minimize construct-irrelevant variance from a variety of sources in item writing. For this reason, the current study focused on construct representation.

According to Embretson (1983), construct representation was often assessed with the following idea: student performance is explained through test items that elicit the cognitive processes, strategies, and knowledge involved in those performances. When a test item feature succeeds in activating more complex processing, the difficulty of the test item with the specific feature will be greater than that of items without the feature. Such empirical investigation provides evidence that a feature related to the target construct can definitely activate students' performances (e.g., Daniel and Embretson, 2010). This is part of the test validation between construct definition and the evaluation process. During the initial stages of test development, the connection between the two is necessary for iterative revision in accordance with the preliminary result.

Item Difficulty Modeling for Checking Construct Representation

One of the most common approaches for checking construct representation is item difficulty modeling (IDM). This process examines how certain test item characteristics affect the difficulty of the question. Both *p*-values (i.e., the proportion of correct answers) in the classical item analysis and *b* (item difficulty) parameters in item response theory (IRT) have been adopted as indices of item difficulty. In item difficulty modeling, radicals, and incidentals should be distinguished to evaluate whether hypothesized radicals definitely affect item difficulty and whether assumed incidentals do not. A clear distinction between radicals and incidentals may contribute to maximizing construct-relevant variance and minimizing construct-irrelevant variance.

To identify radicals and incidentals, cognitive psychology theories, and findings are very useful (Embretson and Gorin, 2001). In such a framework, test items are developed that activate a cognitive process relevant to the target skills, after the theory of the construct to be measured is reviewed. This approach has been termed cognitive psychology principles. When cognitive theories and findings contain radicals of students' thinking and behaviors that reveal their abilities, such cognitive items enable the prevention of construct underrepresentation. Additionally, test items based on cognitive theory make the interpretation of test scores much clearer (Embretson and Gorin, 2001; Kane, 2013). In 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. To provide evidence that candidate radicals in cognitive theories definitely influence item difficulty, test scores correspond with the cognitive process, and indicate the success or failure in processing. Item development with the manipulation of radicals derived from cognitive theories contributes to theory-based validation (Kane, 2013).

Factors Affecting the Difficulty of Summarization Items

Summarization skills are well-researched in the area of cognitive psychology; therefore, theories and findings on summarization can be utilized. In the first part of this section, I review previous IDM studies on reading items to primarily identify the source of difficulty of traditional reading items for L2 learners. The second part focuses on cognitive studies on summarization to identify the source of difficulty of the summarization items addressed in this study.

Sources of Difficulty of Reading Items for L2 Learners

Many previous studies have been conducted to identify features affecting the difficulty of test items in the area of educational measurement. However, there are fewer studies (e.g., Freedle and Kostin, 1991, 1993) aimed at addressing test items that measure something akin to summarization skills (i.e., the main idea item).

Previous studies have found many factors that influence item difficulty. Drum et al. (1981) showed that the number of content words in a passage, unfamiliar words in questions and options, and the inclusion of new content words in correct and incorrect options increased difficulty. Freedle and Kostin (1991, 1993) examined the difficulty of three types of items in the SAT, TOEFL, and GRE exams: the main idea, inference, and supporting items. The main idea items measured learners' skills to understand the central purpose of a text, which is relevant to summarization skills. Freedle and Kostin (1991) found that the concreteness of a text contributed to making the main idea item easier. Freedle and Kostin (1993) also reported that the positioning of information concerning the main idea impacted item difficulty; in their study, the main idea items were easier when information regarding the main idea of the text was located in the first and second paragraphs, and vice versa.

Recent research aimed at identifying features affecting IRT difficulty parameters revealed that text and option features had a significant impact on item difficulty. Embretson and Wetzel (1987) proposed the information processing model in answering reading items and employed the linear logistic test model (LLTM) to estimate text- and response-related effects on item difficulties.

They demonstrated that substantial influences of response-related variables such as word frequencies in correct and incorrect options were observed (Embretson and Wetzel, 1987). Similar results were obtained in Gorin and Embretson (2006); response-related features affected item difficulties. Gorin (2005) extracted four item features and revealed that negative wording and passive voice in texts made items difficult. Sonnleitner (2008) found that the text-related characteristics such as propositional complexity and inference process of causality from written resources increased item difficulty. Baghaei and Ravand (2015) investigated factors affecting item difficulty in Iranian reading items and suggested that the items required to understand the main idea of a text are difficult.

As shown in these previous studies, text characteristics, question specifications, and word usage in options were found to be essential in explaining item difficulty. In addition, tasks that grasp the main idea of a paragraph are difficult, and may depend on several features such as the textual and paragraph structure, word familiarity, and question and option specifications. As this study examines the features of summarization items, it is necessary to examine these factors. Moreover, additional factors, such as text characteristics and the sub-processes of summarization, need to be considered to explain item difficulty. For a more detailed investigation on item features, it will be necessary to conduct a review of the cognitive models of summarization.

Identifying the Sources of Difficulty of Summarization Items

In this section, cognitive theories and models are reviewed to identify the sources of difficulty of the summarization items treated in this study. This section addressed three main factors: passages, distractor characteristics, and response formats.

English passages

It is necessary to address text characteristics when measuring summarization skills as well as general reading skills. Hidi and Anderson (1986) insisted that three main textual characteristics affect students' summarization skills: the text length, genre, and complexity. First, the text length has a great impact on students' performance. Previous studies showed that longer texts increased cognitive load, and summarization became more difficult (Hidi and Anderson, 1986; Anderson and Hidi, 1988). Empirical studies revealed that the paragraph length (the number of sentences in the longest paragraph) increased the difficulty of the main idea items (e.g., Freedle and Kostin, 1991). Concerning text genre, Freedle and Kostin (1993) reported that texts covering the humanities were easier, while those covering the social sciences were harder. Kobayashi (2002) investigated the effects of text genre and students' proficiency levels on scores for summarization, and revealed that texts with cause-and-effect and problem-solving structures were easier for highly proficient students to summarize than texts with only monotone descriptions; for less proficient students, genre did not affect summarization skills. Concerning text complexity, Hidi and Anderson (1986) pointed out that this aspect was difficult to define. Examples of elements used to measure text difficulty include the number of low-frequency words and elaborate sentence structures (Hidi and Anderson, 1986). One of the measures to assess text complexity is readability. The index of readability has a variety of factors; linguistic features including word- and structure-related components were evaluated in most indices. It was found that readability of the text increased item difficulty (Mosenthal, 1998).

Distractor characteristics

The second factor that affects summarization difficulty is distractor characteristics. In principle, distractors should be developed so as to reflect students' common errors (e.g., Haladyna and Rodriguez, 2013). Common errors made by students in each process have been found to be prominent sources of difficulty (Gorin and Embretson, 2012). Most previous works reported that the increased number of falsifiable distractors made test items more difficult (e.g., Embretson and Daniel, 2008; Ozuru et al., 2008). However, these studies did not focus on the cognitive characteristics of distractors. It is important for item writers to be aware of what cognitive feature should be included in distractors, and whether such distractors make test items difficult.

Previous studies suggested the common characteristics of L2 summarization (e.g., Brown and Day, 1983; Johns and Mayes, 1990; Kim, 2001; Keck, 2006). These studies were largely based on the most fundamental model of summarization, proposed by Kintsch and van Dijk (1978). The Kintsch model assumes that readers apply three macrorules–deletion, generalization, and integration–to establish a global meaning from propositions at the sentence level (Kintsch and van Dijk, 1978). Deletion is the first process of summarization, in which a proposition that does not have clear connections with other propositions (Kintsch and van Dijk, 1978), or contains trivial and redundant information (Brown and Day, 1983), is deleted. Generalization is the second process, in which a set of propositions are substituted by a general and global proposition (Kintsch and van Dijk, 1978) or replaced with a superordinate phrase (Brown and Day, 1983). Integration is the process of expressing a global fact of propositions that include normal conditions, components, and consequences (Kintsch and van Dijk, 1978). Readers represent a macrostructure of the paragraph or the text by applying these three macrorules.

In the area of second language learning, several characteristics of summarization have been discovered based on the studies by Kintsch and van Dijk (1978) and Brown and Day (1983). Understanding the characteristics of summarization relevant to L2 learners is fruitful for identifying errors in L2 summarization and guiding the development of functioning distractors in summarization items.

The first feature of summarization is the copy-delete strategy, which refers to the tactic of borrowing existing sentences in the text and including them in the summary with little or no modification. Brown and Day (1983) examined readers' use of macrorules by grade. In their study, fifth and seventh graders employed the copy-delete strategy to develop summaries. Similar patterns have been observed in other studies that examined the summarization skills of L2 learners (e.g., Johns and Mayes, 1990; Keck, 2006). Johns and Mayes (1990) compared the summaries

of proficient and less-proficient readers, and showed that less-proficient L2 readers were more likely to directly copy text in their summaries than proficient readers. Keck (2006) examined the differences in paraphrasing skills between L1 and L2 learners; the results showed that L2 learners employed exact and near copies during summarization, while L1 learners moderately or substantially revised the phrases in the text, which was consistent with the results of Johns and Mayes' (1990) study. Overall, the copy-delete strategy is one of the common characteristics observed in summaries written by L2 learners, particularly less-proficient readers.

The second feature is that L2 summaries rarely contain an integrated idea unit from several micro-propositions (e.g., Johns and Mayes, 1990; Kim, 2001). Kim (2001) analyzed summaries written by Korean college students, and showed that their summaries rarely contained an integrated idea unit derived from two or more micro-propositions. This feature was consistently observed regardless of participants' proficiency levels (Johns and Mayes, 1990). L2 learners may process each micro-proposition separately and select the important ones for summarization, but they do not integrate multiple propositions into a general one.

Based on these findings, Terao and Ishii (2019) developed multiple-choice questions to measure summarization skills, and compared the functions of two types of distractors each that reflected summarization errors in the deletion, generalization, and integration processes. For the deletion process, a summary with unnecessary information and one without the necessary information were examined. For the generalization process, a summary with detailed information, such as episodes and examples, and one with inappropriate superordinates were tested. For the integration process, a summary that partially described the original author's intent and one that presented viewpoints different from the author's opinions were examined. Terao and Ishii (2019) found that, in the deletion process, summaries without necessary information were selected by less-proficient test takers, and were not selected by proficient ones. In the generalization process, summaries with detailed information were attractive for low- and middle-proficiency test takers. In the integration process, summaries with a viewpoint that differed from the author's opinion worked as a functioning distractor.

Response formats

The third factor that affects summarization difficulty is response formats. The item format is closely related to cognitive demand and item performance, and the effect of the format depends on the construct to be measured (Haladyna and Rodriguez, 2013). Many previous studies investigated the influence of multiple-choice and construct-response formats, but few examined the impact on item difficulty of different types within multiple-choice formats (e.g., Case and Swanson, 1993). This study considered two multiple-choice item formats in the context of measuring summarization skills: conventional multiple-choice (CMC) and extended matching (EM) formats.

A CMC item consists of a test question with one correct answer and several distractors (Haladyna and Rodriguez, 2013). By using the CMC format, test items ask participants to select the most appropriate option from a list of candidates. Distractors must be presented together, and responses to distractors cannot be gathered when they answer the item correctly. An EM item has (a) a theme, (b) a lead-in statement, (c) an option list, and (d) two or more item stems (Case and Swanson, 1993). In the EM format, options are commonly used as candidates' answers for multiple stems. Case and Swanson (1993) also revealed that multiple-choice items with the EM format were more difficult than five-option multiple-choice items. By employing the EM format, it is possible to collect the students' responses to every option, including appropriate and inappropriate answers.

Study Design

The current study investigated the influence of English passages, distractor characteristics, and response formats on the difficulty of summarization items. The research question addressed what increased or decreased the difficulty of summarization items, and what were the radical components in measuring such skills. The study assumed that passage difference was an incidental; distractors and response formats were radicals.

First, this study expected that substantial effects of the texts were not detected. While text-related factors, such as text length, genre, and complexity, affect students' summary writing, according to previous studies (e.g., Hidi and Anderson, 1986), most testing programs would like to target reading skills independent of passages. This study carefully selected texts with similar characteristics to be summarized in terms of these three aspects, and treated text differences as an incidental.

Second, this study hypothesized that each distractor characteristic contributes to increasing item difficulty. Distractors were created to reflect students' errors, based on Terao and Ishii's (2019) study. This study developed experimental items to focus on each summarization process; three types of summarization items (deletion, generalization, and integration items) were examined. And distractors with two different common errors were contained in each type of summarization item. Six kinds of distractor summaries were identified as follows: containing unimportant information and missing necessary information for deletion items; including examples or episodes and inappropriate superordination for generalization items; and partially describing the original author's intent and presenting viewpoints different from the author's opinions for integration items. In this study, all six types were treated as common errors in summarization for L2 learners. Since distractors contained common errors in summarization, and thus central components of the targeted construct, this study considered distractor characteristics as radicals. Attractive distractors make it more difficult for test takers to answer that item correctly.

Third, two types of multiple-choice formats, the CMC and EM formats, were examined. In both formats, test takers select one of the appropriate options presented. This study predicted that two response formats elicit different cognitive demands, and thus, impact the difficulty of summarization items as a radical. The item specification in this study was as follows: one item with three options was used to measure one of the three summarization processes. In the item, one of the options was an appropriate summary that succeeded in activating the

targeted sub process and the remaining two were distractors that reflected students' errors in that process. When answering the CMC items, participants select the most appropriate candidate summary; when answering the EM items, participants evaluated each summary and selected the appropriate status from an option list. The assessment of each summary required much higher proficiency in the EM format than the CMC format, which may result in the difference in item difficulty.

MATERIALS AND METHODS

Participants

The participants were 150 freshmen (28 female and 122 male) from 11 national universities in the Kanto region. All participants were Japanese L2 learners.

All participants were recruited through the 4 day survey program held by the Research Division of the National Center for University Entrance Examinations. This program consisted primarily of an operational and an experimental part. In the operational part, participants completed the National Center Test on the same day the actual test was administered. In the experimental part, researchers conducted experiments on new testing technologies and methods. There were four experimental slots of 80 min each. This study used one slot to collect response data. Participants used the first 30 min for this study, and the remaining 50 min for another study. Those who participated in the 4 day survey program received 40,000 JPY in compensation. This study was approved by the ERB Board in the Research Division of the NCUEE.

Materials

English Passages

This study utilized English passages excerpted from reading tests that were previously administered in Japanese university entrance examinations, because it was necessary that the texts have an appropriate difficulty level for the participants in this study. The answer duration was only 30 min, so two passages were selected to develop the experimental items.

The passages used in this experiment were carefully chosen to have similar text length, genre, and level of complexity. While Hidi and Anderson (1986) noted that text complexity was difficult to define, their study also listed elaborate sentence structure as a partial index of complexity. This study adopted the Flesch-Kincaid readability index as a readability measure, as well as to evaluate text difficulty and determine text complexity. This index represents the U.S. grade level appropriate for reading the selected passages; it was also used in previous studies to investigate the text difficulty of reading tests in Japanese entrance examinations (e.g., Brown and Yamashita, 1995; Kikuchi, 2006).

The characteristics of the two English passages were as follows. Passage 1 was 643 words in length, and its Flesch-Kincaid readability was 8.64. This passage described several characteristics of veganism, and suggested a realistic diet for situations wherein a vegan cannot manage what they eat. Passage 2 was 742 words in length, and its Flesch-Kincaid readability was 9.48. This passage presented the history and current situation of homesickness, and suggested that modern technology made people who were away from home feel lonelier. The two texts were both expository, so the genre was the same. According to previous studies (e.g., Brown and Yamashita, 1995; Kikuchi, 2006), the readability was 10.98 on average and ranged from 7.00 to 12.00 for private universities; for national and public universities, the readability was 9.62 on average and ranged from 8.20 to 15.00. With reference to these values, the two texts chosen in this study were within the range of text difficulty prescribed by Japanese university entrance examinations, and were assumed to match the reading level of participants.

Experimental Test Items

This study developed original test items for the experiment through the following three steps: (1) item specification, (2) development of candidate summaries, and (3) manipulation of response formats.

First, the overarching plan for item development was determined. This study developed three types of test items to focus on each summarization process: deletion, generalization, and integration items. Each test item instructed participants to read the paragraph designated in the stem (the question part). Deletion items asked participants to identify important information in the designated paragraph. Generalization items required them to replace detailed information, such as episodes or examples, with a more concise, abstract description. Integration items required participants to represent the topic sentence of the paragraph. The paragraphs were selected to activate various focal processes in each test item. For deletion items, a paragraph containing both important and unimportant information was chosen; for generalization items, a paragraph including episodes and examples was selected. For integration items, a paragraph without any clear topic sentence was chosen.

Second, candidate summaries were developed for each item. This study edited three candidate summaries per item; one correct (key) and two distractors. The summary representing the correct answer was written to be a successful result of each summarization process, while the two distractor summaries were written to include student errors for each summarization process, as found in Terao and Ishii's study (Terao and Ishii, 2019). In deletion items, a correct summary contained only the important information from the paragraph, while the two distractor summaries, respectively (1) contained unnecessary information (denoted as D-1 below), and (2) missed important information (D-2). In generalization items, a key summary included a generalized statement from the episodes or examples in the paragraph, and two distractors that, respectively, contained (1) episodes or examples themselves (G-1), and (2) misinterpretations of episodes or examples that were inappropriately generalized (G-2). In integration items, a key was edited to reflect the author's intention in a paragraph, and two distractors that were respectively, (1) viewpoints different from the author's (I-1), and (2) a partial description of the author's intent (I-2).

Third, response formats were manipulated for each item. Instructions to the participants in the item stem were changed between the two formats. In the CMC version, participants were asked "Which of the following statements is the best summary

TABLE 1 | Test form design.

Test form	CMC-1	CMC-2	EM-1	EM-2
1	○	○		
2	○			○
3		○	○	
4			○	○

of Paragraph (X)?" and to select one of three listed summaries. In contrast, the instruction for the EM version was to "Classify every sentence into the following categories below," and they chose the appropriate evaluation for each summary. Categories were edited by the targeted process to represent the status of the summary (correct or incorrect) and the type of misconception if it is incorrect. The three categories for deletion items were displayed as follows: (a) a correct summary, (b) an incorrect summary because important information was missing, and (c) an inappropriate summary because unimportant information was included. For generalization items, the categories were: (a) a correct summary, (b) an incorrect summary because detailed information was included, and (c) an incorrect summary because an expression was mistakenly generalized. For integration items, they were: (a) a correct summary, (b) an incorrect summary because viewpoints different from the author's were included, and (c) an incorrect summary because the author's intent was not fully described.

Due to the use of different response formats, the unit "item" varied between the CMC and EM versions. In the CMC format, a task to choose one of the three candidate summaries (options) represented an item. In the EM format, however, a task to select the most appropriate category of each summary represented an item. Thus, the CMC version comprised three items in one text, while the EM version comprised nine. Overall, four kinds of testlets were edited: the CMC version for Passage 1 (CMC-1, three items) and Passage 2 (CMC-2, three items); and the EM version for Passage 1 (EM-1, nine items) and Passage 2 (EM-2, nine items). Item stems and options were checked by a subject matter expert. After the check, stems, and options were revised in terms of grammar and content issues. Examples of both formats are shown in the **Appendix**.

Construction of Test Forms

Four testlets, CMC-1, CMC-2, EM-1, and EM-2 are included in the four test forms shown in **Table 1**. A test form contained two of four testlets, provided that testlets for the same passage were not included in the same form. A common-item design was employed in constructing test forms to calibrate item parameters of different forms on the same scale. The number of participants assigned to each booklet were as follows: 35 participants answered booklet 1; 35 answered booklet 2; 40 answered booklet 3; and 40 answered booklet 4.

Procedure

In this study, printed versions of the test forms were prepared and delivered. Four test forms were randomly assigned to each participant who had 30 min to answer two testlets from the booklet delivered to them.

Data Analysis

Data analysis comprised two major parts: checking descriptive statistics for each item and applying the LLTM (Fischer, 1973) to estimate the effect of several item features on item difficulty.

Before the main analysis, a check for descriptive statistics was performed by classical item analysis. The proportions of correct responses (p-values) and the point biserial correlation between item scores (1 for correct, and 0 for incorrect) and participants' proficiency (r_{pbs}) were evaluated for each item. The raw total score could not be used as participants' proficiency scores because one of the four different test forms was assigned to participants; hence, the meaning of the total score might vary based on form. In this study, the proficiency estimates of theta in the Rasch model described later were used as proficiency scores and as correlations between dichotomous item scores and estimated theta. Results of this analysis were reported in **Supplementary Table 1**.

The first analysis checked the assumption of applying the Rasch model; the unidimensionality and local independence. For a unidimensionality check, this study employed item fit indices to test misfit, and could not use standard methodology such as principal component analysis or Rasch residual factor analysis. Parts of tetrachoric correlation were missing since two versions of response formats that had the same text were not presented to the same participants. In this case, any correlational methods could not be applied to the data. This study employed item fit statistics such as outfit and infit measures. Outfit statistic is the unweighted mean square of the residual of the Rasch model, and infit is the weighted mean square of the residual (Bond and Fox, 2015). These statistics are larger than 1 when other constructs influence the dimension we would like to use; smaller than 1 when response patterns are deterministic. Outfit and infit statistics can be transformed into values following t-distribution, so the transformed statistics were also reported. Bond and Fox (2015) indicated that, except in high-stakes situations, outfit and infit statistics fall between 0.70 and 1.30. This study also adopted such criteria.

The primary analysis in this study was to estimate the influence of item features on item difficulty using the LLTM. This study compared the following three models and examined the goodness of fit indices. The first model was the null model, which constrained item difficulty as equal. The second was the Rasch model, which estimated item difficulty parameters for every item.

Generally, the Rasch model is expressed as

$$P\left(\theta_i \middle| b_j\right) = \frac{\exp(\theta_i - b_j)}{1 + \exp(\theta_i - b_j)}$$

where θ_i denotes the ability parameter of student i, and b_j denotes the difficulty parameter of test item j. The difficulty parameter was constrained as equal in the null model while it was freely estimated in the Rasch model.

The third model was LLTM, in which several features of test items predicted item difficulty. LLTM expands the Rasch

model to explain the variance of difficulty parameters by K item features:

$$b_j = \sum_{k=1}^{K} q_{jk}\eta_k$$

Here, q_{jk} is the component of the design matrix, which expresses the relationship between test items and features. Regularly, q_{jk} is coded as 1 when the item j has a feature k, and as 0 when that item does not have a feature k. η_k signifies the magnitude of the influence on item difficulty parameters of that feature. When η_k is large, the feature k has a large impact on difficulty of test items.

In the LLTM, the design matrix was determined by the following nine dummy variables. A portion of the design matrix is shown in **Table 2**. Differences in passages are coded 0 for Passage 1 and 1 for Passage 2 in column V1. Each distractor characteristic was coded in columns V2 to V7, respectively. When a test item had a summary with a certain feature, it was coded as 1; otherwise, it was coded as 0. The CMC version had two types of distractors; the corresponding two columns of that item were both coded as 1. The EM version had only one summary; therefore, when a correct summary was presented in the EM item, only one column concerning the related distractor characteristic was coded as 1. Differences in response formats were represented in column V8; the EM format was coded as 1, and the CMC format as 0. Finally, column V9 represented the intercept, and all items were coded as 1 in this column.

TABLE 2 | Design matrix of LLTM.

	V1	V2	V3	V4	V5	V6	V7	V8	V9
Q11–1	1	0	0	1	1	0	0	0	1
Q11–2	1	1	1	0	0	0	0	0	1
Q11–3	1	0	0	0	0	1	1	0	1
Q12–1	1	0	0	0	0	0	0	1	1
Q12–2	1	0	0	0	1	0	0	1	1
Q12–3	1	0	0	1	0	0	0	1	1
…	…	…	…	…	…	…	…	…	…
Q22–9	0	0	0	1	0	0	0	1	1

The row is a list of test items, and the column is a set of item features. V1, passage differences; V2, distractor D-1; V3, distractor D-2; V4, distractor G-1; V5, distractor G-2; V6, distractor I-1; V7, distractor I-2; V8, response formats; V9, the intercept.

The goodness of fit was evaluated for these three models, such as chi-square differences between models, $-2\ln L$ (-2 times the log likelihood), AIC (Akaike information criteria), and BIC (Bayesian information criteria). In addition, the incremental fit index, $\Delta^{1/2}$, was also calculated to examine the proximity of likelihood compared to the saturated model. In this study, the saturated model was the Rasch model because it parameterized difficulty for every item, while the null model and the LLTM constrained item difficulty parameters. Previous studies using LLTM also assessed these indices to evaluate the goodness of fit of LLTM (e.g., Embretson, 2002; Embretson and Daniel, 2008). This index was found to correspond with the multiple correlations between the difficulty parameters in the Rasch model and predicted values in the LLTM (Embretson, 1999). Estimates of item difficulty parameters in the Rasch model and LLTM were compared graphically in the scatterplot to check how approximate two sets of estimates were.

Parameters in these three models were estimated by the flirt package (Jeon et al., 2014) in R language (R Core Team, 2018). Parameter estimation was conducted by applying the expectation-maximization (EM) algorithm. Using this package, the sign of item difficulty parameter b was reversed to quickly make the calculation. Hence, one must use caution when interpreting the parameters; a positive sign means that a certain feature makes test items easier, and a negative sign means that the feature increases test item difficulty. This study employed the concurrent calibration method, not chain equating. In using this method, item difficulty parameters were estimated simultaneously with the overall data matrix in which responses to test items in the other forms were treated as missing.

RESULTS

The results of classical item analysis, such as p-values and point biserial correlation, are presented in **Supplementary Table 1**.

TABLE 3 | Item fit statistics in the Rasch model.

Item	Outfit MNSQ	Outfit *t*	Infit MNSQ	Infit *t*	Item	Outfit MNSQ	Outfit *t*	Infit MNSQ	Infit *t*
Q11–1	1.069	0.776	1.045	0.518	Q21–1	1.005	0.069	1.008	0.124
Q11–2	0.991	−0.159	0.990	−0.173	Q21–2	0.921	−1.107	0.934	−0.921
Q11–3	1.013	0.179	1.004	0.054	Q21–3	1.057	0.728	1.045	0.594
Q12–1	0.958	−0.769	0.964	−0.646	Q22–1	0.982	−0.308	0.988	−0.209
Q12–2	1.179	0.698	1.037	0.221	Q22–2	1.093	1.585	1.082	1.404
Q12–3	0.797	−1.291	0.890	−0.648	Q22–3	1.031	0.518	1.028	0.467
Q12–4	1.008	0.141	1.010	0.178	Q22–4	0.980	−0.309	0.979	−0.319
Q12–5	0.970	−0.499	0.971	−0.477	Q22–5	1.011	0.129	1.000	0.019
Q12–6	0.949	−0.916	0.951	−0.874	Q22–6	1.057	1.036	1.052	0.950
Q12–7	0.942	−0.892	0.945	−0.834	Q22–7	1.055	0.935	1.048	0.834
Q12–8	1.031	0.526	1.032	0.541	Q22–8	1.044	0.713	1.034	0.555
Q12–9	0.970	−0.314	0.983	−0.169	Q22–9	0.976	−0.136	0.968	−0.188

TABLE 4 | The goodness of fit in the null model, Rasch model, and LLTM.

	$-2\ln L$	AIC	BIC	$\Delta^{\frac{1}{2}}$
Null	2512.874	2516.874	2522.895	–
Rasch	2354.055	2404.055	2479.321	–
LLTM	2420.312	2444.312	2480.439	0.762

Item Fit Statistics

Table 3 shows item fit statistics when applying the Rasch model to the data. Mean squares of outfit and infit measures (Outfit MNSQ and Infit MNSQ in **Table 3**) were within the acceptable ranges for the experiment (0.70–1.30). Outfit and infit t statistics ranged from −2.00 to 2.00 and were within the appropriate values. Based on these statistics, the misfit due to multidimensionality and the overfit because of local dependence were not detected. Therefore, the Rasch model was considered to fit the data. This study assumed the unidimensionality and local independence and continued to apply a family of the unidimensional Rasch models.

Fitting LLTM to Estimate the Impact on Item Difficulty

Comparing the Fit of Statistical Models

We constructed three statistical models and compared the fit of these models. **Table 4** shows $-2\ln L$, AIC, BIC, and the delta index ($\Delta^{1/2}$) for each model. Chi-square differences were large between the null model and the Rasch model [$\Delta\chi^2$ (23) = 158.81; $p < 0.001$] and between the null model and LLTM [$\Delta\chi^2$ (8) = 92.21; $p < 0.01$]. Differences were also observed between the Rasch model and LLTM [$\Delta\chi^2$ (15) = 66.61; $p < 0.001$]. These fit indices suggested that the Rasch model was the most appropriate out of the three models constructed in this study.

Two information criteria, AIC and BIC, both suggested the Rasch model, but a different pattern of the proximity between two indices was observed; while the difference in AIC values between two models were larger, the BIC values were small. It depends on the characteristics of these two criteria: AIC suggests the model that predicted future data, and BIC favors the model that explained the data-generating structure and weighs the simplicity of the model (Sober, 2002). Considering the focus of this study, these three models needed to be compared in terms of explanation. Thus, the proximity of LLTM relative to the Rasch model should be further examined.

The delta index, which is comparable to the magnitude of multiple correlation, indicated a moderate level. This means that about 58% of the variance of item difficulty parameters were explained by the features addressed in this study. Gorin and Embretson (2012) indicated that most IDM studies reported 30% to 60% of variance in item difficulty parameters (e.g., Embretson and Wetzel, 1987; Enright and Sheehan, 2002; Gorin, 2005). LLTM was indicated to be relatively approximate to the Rasch model. Estimated item difficulty parameters in the two models are presented in **Table 5** and **Figure 1**. The correlation between parameters in these two models was 0.768, showing the relatively strong relationship. Therefore, predictors included in the LLTM were sufficient to explain the item difficulty.

TABLE 5 | Item difficulty parameters in the Rasch model and LLTM.

Item	Rasch	LLTM	Item	Rasch	LLTM
Q11–1	−0.633	0.038	Q21–1	−0.207	−0.349
Q11–2	−0.193	−0.036	Q21–2	0.190	0.459
Q11–3	0.357	0.360	Q21–3	−0.133	0.519
Q12–1	−0.043	0.232	Q22–1	0.062	0.231
Q12–2	2.184	0.897	Q22–2	−0.341	−0.192
Q12–3	−1.461	−1.233	Q22–3	−0.410	−0.280
Q12–4	−0.043	−0.170	Q22–4	0.043	−0.395
Q12–5	−0.259	0.232	Q22–5	−0.397	−0.695
Q12–6	−0.097	−0.239	Q22–6	0.062	−0.052
Q12–7	0.388	0.213	Q22–7	0.727	−0.193
Q12–8	−0.204	−0.226	Q22–8	0.062	0.267
Q12–9	0.789	0.232	Q22–9	−1.403	−1.220

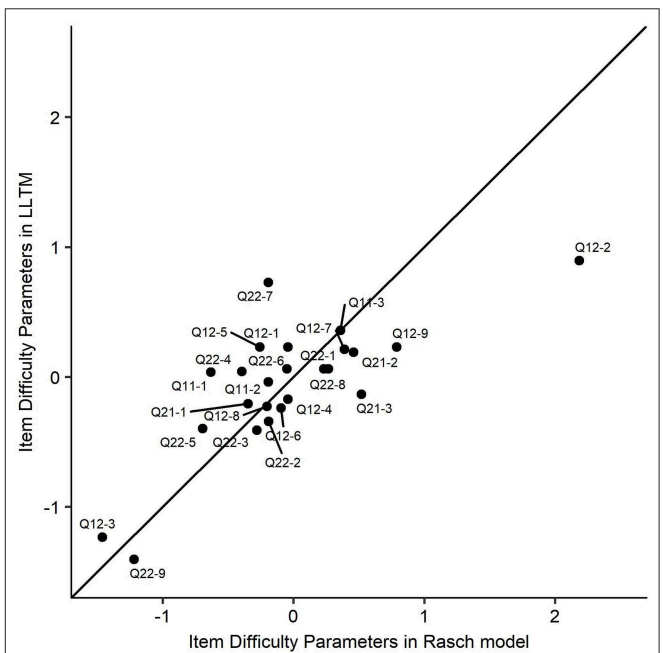

FIGURE 1 | Scatterplot of item difficulty parameters in Rasch model and the LLTM.

Examining the Impact on Item Difficulty

Table 6 shows the result of estimation for ηs, the magnitude of the impact of each item predictor on item difficulty. The information regarding the statistical significance was contained in the 95% confidence interval in **Table 6**. When the confidence interval of a predictor did not include zero, the predictor was statistically significant at the 0.05 level, and the null hypothesis (H_0: $\eta = 0$) was rejected. When the confidence interval included zero, this predictor was not statistically significant, and the null hypothesis held.

TABLE 6 | Parameter estimates and 95% confidence intervals in LLTM.

Variables in the design matrix	η	SE	95%CI
Passage	0.170	0.101	[−0.027, 0.368]
Distractor characteristics			
Unnecessary information	−0.403	0.176	[−0.749, −0.057]
Missing necessary information	−0.472	0.195	[−0.818, −0.125]
Concrete	−1.465	0.195	[−1.847, −1.083]
Inappropriate	0.665	0.184	[0.304, 1.026]
Partial description	−0.019	0.176	[−0.363, 0.325]
Different viewpoint from the author's	−0.458	0.176	[−0.804, −0.112]
Response format	−0.605	0.204	[−1.008, −0.208]
Intercept	0.667	0.267	[0.144, 1.190]

Passage difference was not a significant predictor of item difficulty. It was indicated that test item difficulty did not differ between the two passages. The results of distractor characteristics revealed that most of them increased item difficulty. These estimates indicated that avoiding most types of distractor summaries and selecting correct answers required higher proficiency. Response formats, either the CMC or EM version, have a significant impact on item difficulty. The result shows that EM items were more difficult than CMC items.

Detailed results of the effects of distractor characteristics are described below. For the deletion process, including a distractor summary with unimportant information (D-1) and a distractor summary without necessary information (D-2) made the item difficult. It was indicated that in the CMC format, including these types of student errors in the candidate summaries successfully distracted participants from choosing a key; in the EM format, a summary with such types of errors was misclassified.

Concerning the generalization process, including a summary with concrete expressions (G-1) in the option list increased item difficulty. In contrast, including inappropriate superordinates in a candidate summary decreased item difficulty. It was suggested that the summary that included inappropriate replacement of episodes or examples may be easier to falsify in the CMC format. It may also be easier to find a fault in that summary and to classify it into the appropriate category.

For the integration process, a summary that partially described the author's intent (I-1) had no explanatory power for item difficulty. In turn, including the statements different from the author's (I-2) made the item difficult. Detecting a different point of view from the author's and classifying that summary into the corresponding category required a higher level of summarization skills.

DISCUSSION

This study performed difficulty modeling of summarization items for Japanese learners, and investigated the effect of item characteristics, such as passages, distractor features, and response formats. The results of this study suggested that the extended matching (EM) format was more difficult than the conventional multiple-choice (CMC) format. It was also found that most types of distractors contributed to making items difficult, while a certain type of distractor decreased the difficulty of summarization items. The goodness of fit of LLTM was at a moderate level for predicting item difficulty, and about 58% of the variance of item difficulty parameters in the Rasch model could be explained by item features in this study. In light of these findings, the model fit of this study is sufficient to explain item difficulty.

The next sections discuss the effects of each item feature on item difficulty in measuring summarization skills, and consider future item development and generation.

Sources of Difficulty of Summarization Items

Passage

Differences in item difficulty could not be observed between the two passages. In general, passage differences influenced student responses of summarization tasks. Since the passages were selected after careful consideration and the text had similar characteristics in terms of the length, genre, and complexity, there are just small differences in item difficulty between the two passages and a substantial effect was not detected. The results of this study suggest that the careful selection of passages with similar characteristics can minimize the influence of summarization material.

During the item development process, various passages are often used to develop test items. It is much more difficult to control all the characteristics of the passages used in tests. However, L2 learners are required to understand the primary idea of various kinds of passages. In such situations, the influence of text characteristics on item difficulty should be minimized. While further research on passage variability is needed, there is a possibility of minimizing the influence of texts on item difficulty when factors concerning English passages, such as length, genre, and complexity, are carefully examined.

Distractor Characteristics

This study suggested that distractor characteristics had an enormous impact on the difficulty of summarization items. As suggested by Embretson and Wetzel (1987), cognitive processes concerning the response largely influenced item difficulty rather than the text itself, in reading tests. This study also revealed that distractor types, as a variable of question features, explained a large proportion of the variance of difficulty parameters. The inclusion of a well-functioning distractor leads test takers to answer incorrectly; therefore, the estimated effect of including a certain distractor on item difficulty was found to be a negative value, implying that the distractor makes the test item difficult. Most estimates concerning distractors were negative, but one distractor, a summary with inappropriate superordinates in generalization items, showed a positive value.

In deletion items, students' common errors, such as the lack of necessary information or inclusion of unnecessary information, were embedded in candidate summaries to edit distractors. It was shown that both distractors made items difficult. Rejecting summaries with these errors and answering the item correctly

required a higher proficiency from students. Hidi and Anderson (1986) stated that readers were required to focus on the importance of each proposition in the text when selecting the necessary information. Consistent with Hidi and Anderson (1986), this study showed that judgement of the inclusion of necessary information and exclusion of unnecessary information in the list of summaries was an essential component of summarization skills, from the view of construct representation.

In generalization items, summaries with episodes or examples that contained inappropriate superordinates were examined. Results showed that two types of distractors were functioning differently. Including concrete expressions in candidate summaries made test items much more difficult. Falsifying the summary with episodes or examples was relatively more difficult than other items. Terao and Ishii (2019) revealed that summaries with episodes and examples were frequently chosen by low- and mid-proficient test takers, while they were not chosen by proficient test takers. The results of their study were consistent with those of the current study; that is, falsifying the distractor that includes detailed information in a summary requires higher summarization skills. Hidi and Anderson (1986) and Kim (2001) suggested that the novice's summary included few expressions across sentences even within the paragraph. L2 readers tend to process information at a sentence level (Kozminsky and Graetz, 1986), even when more detailed pieces have to be generalized. This study empirically indicated that such cognitive operations can be central to the generalization process, as well as overall summarization skills.

In contrast, including summaries with inappropriate superordinates promoted correct answering of items. This distractor may be easily falsified and increased the possibility of reaching the correct answer, suggesting that this type of distractor did not require high proficiency. In developing generalization items, it should be remembered that such errors in candidate summaries may lead to easier items than others. This result was partly inconsistent with Terao and Ishii's study (Terao and Ishii, 2019), in which only college students with higher proficiency could falsify such distractors. The current study used a few test items, and a specific distractor (e.g., Q12-2) may have affected the parameter estimates. Further research should employ a larger number of test items and examine this effect again.

In integration items, a summary that partially described the author's intent and another that contained a viewpoint different from the author's were included as distractors. Results showed that only the latter type of distractor functioned well, while the former did not work as a predictor of item difficulty. This result indicated that judging whether a summary is representative of the paragraph can be an important component of summarization skills. In contrast, including a distractor that contained partial descriptions did not change the item difficulty as much. Although additional analysis using a larger number of test items is necessary, it was suggested that developing this distractor does not contribute to making test items easier or more difficult. Results of the current study suggest that it is not necessary to develop this type of distractor.

In developing summarization items, the following characteristics can be included in candidate summaries to work as distractors: the inclusion of unnecessary information, lack of necessary information, inclusion of detailed information, and perspectives that are different from the author's. Including distractors is one of the most difficult aspects of the item developmental process. Evidence of distractor quality and its impact on item difficulty is very useful for future item development. The findings of this study can help item writers edit effective and functional distractors for summarization items.

Response Formats

The result of response formats showed that the EM was more difficult than the CMC. Case and Swanson (1993) also indicated that EM items tended to be more difficult than CMC items. This tendency may result from the level of cognitive demand. When participants answered the CMC format, they needed to evaluate which summary was correct; in contrast, when answering the EM format, they were required to additionally assess why a summary was incorrect.

When developing summarization items, response formats are often critical to the difficulty of a summarization item. Cognitive demands may differ between two formats, and a careful examination is required. While there are some psychometric advantages in the EM format, such as high discrimination power and reliability (Case and Swanson, 1993), test developers and item writers need to choose a format that is suitable for the targeted population.

Limitations and Future Directions

There are some limitations to this study. First, the study employed a limited number of English passages and test items. Additional studies should be tailored to address the generalized results of this study. Texts with various characteristics could not be fully assessed because this study used only two passages. As stated above, various passages with different lengths, genres, and levels of complexity are used in the actual tests, and item difficulty might be affected by these factors. Future research should investigate the effect of texts with similar or different characteristics on the difficulty of summarization items. Similarly, as the number of test items under one text was limited, it is necessary to confirm whether the study results are consistent with a large number of test items.

Second, the number of participants in this study was relatively small at 150, and about half of them answered each item. The small sample size might result in three major issues: the stability of parameters, the interaction effects between item properties or a certain item property and person covariates, and the fit of other models including item discrimination power. Standard errors of the estimates in LLTM were not small, and this study only addressed the main effect of variables in the design matrix. Resolving the variability of parameters and including the interactions may increase the proportion of explained variance of difficulty in the Rasch model; the findings in this study should be tested again using a larger sample size and models with item-property covariability. Due to the relatively small sample size, as well as the larger numbers of predictors, entering the effects of test forms into the design matrix led to a divergence, and it was not possible to estimate the effects after controlling

for a test-form effect. Future studies need to examine the effects of passages or distractors that control the effect of test forms. In addition, it was difficult to compare the models with equal item discrimination parameters (i.e., Rasch model and LLTM) and the two-parameter models. Of course, the primary focus of the current study was the difficulty of test items, not item discrimination power. However, the two-parameter logistic model (2PLM) and a two-parameter constrained model (2PLCM; Embretson, 1999) did not align the data; the estimates did not converge and standard errors were larger mostly because of the small sample size. Future studies are necessary to check the fit of models with equal discrimination constraint and those without constraints. Additionally, to understand the effect of item features on item discrimination power, a two-parameter constrained model (2PLCM) may also be useful. When generating test items, identifying factors that affect item discrimination power, as well as difficulty, is essential to monitor their psychometric properties. Future studies need to investigate the impact of item features on discrimination power with a larger sample size.

Third, the specification of summarization items requires improvement. This study originally developed experimental items to focus on a specific cognitive process. Such specifications may inform participants about what is expected of them in terms of solving the item. Indeed, for the EM version of deletion items, a list of the status of summaries, such as "An incorrect summary because important information was missing," and "An incorrect summary because unimportant information was included," indicated the type of summaries that were aligned. This possibility may result from the focus on a specific item process. In reality, a combination of processes may be activated in summarizing a certain paragraph. For instance, both deletion and generalization may be required in substituting a general statement for detailed information. Additional investigations are required to evaluate item difficulties for summarization tasks developed with other specifications. These could include investigations on activating a mixture of processes or edited items by simultaneously including distractors concerning multiple processes, such as deletion and generalization.

Fourth, distractor development requires sophistication and automerization. This study manually edited distractors through the standards it had determined. Significant effort and time are required to manually prepare a list of candidate summaries. Alternatively, automating the system may be a solution for developing summaries. Research on automatic text summarization (e.g., Das and Martins, 2007) in the area of natural language processing (NLP) can be applied to future distractor development. Within this framework, important sentences in the text are selected in terms of word frequency, similarity, and word sequence. Such technology will be helpful to reduce the effort required in preparing candidate summaries.

Fifth, this study targeted only Japanese L2 learners. While cognitive models and characteristics of summarization patterns are applicable to all L2 learners (e.g., Johns and Mayes, 1990; Kim, 2001; Keck, 2006; Terao and Ishii, 2019), results of this study may be tested again between Japanese and other L2 learners. To generalize the study findings, further research is required on L2 learners from other countries.

Findings of the current study can be applied to measure L2 learners' summarization skills in learners other than Japanese. Indeed, previous studies found that Korean and Iranian students had similar tendencies as Japanese students (Kim, 2001; Karbalaei and Rajyashree, 2010; Khoshnevis and Parvinnejad, 2015). Although the generalization of findings requires greater attention to differences in participants' native language and educational system, this study may provide an initial framework for measuring L2 summarization and tips for developing and generating summarization items.

Recently, a large number of studies on AIG has been conducted to generate test items algorithmically. Within this framework, a strong theory approach to AIG was proposed to generate test items that control cognitive variables to precisely predict item difficulty. As Lai et al. (2009) described, the strong theory approach promises to reduce the cost of pretesting. For a perfect prediction of item difficulty parameters through cognitive features related to target construct, experimental studies to understand the possible source of the difficulty are the first step toward developing and generating test items effectively and economically. Understanding the source of difficulty is also necessary to ensure quality of testing and assessment. This study provides the basic outline of an empirical tryout on test items to measure summarization skills. In the process of construct definition and description, the results of item difficulty modeling in this study can improve discernment of the dimension both related and unrelated to target construct. Findings in this study may contribute to improving the process of item development when measuring summarization skills of L2 learners. A possible next step is to guide item writers using suggestions from the current study to manually develop good test items. Another step is to generate large numbers of test items with the algorithm, and to evaluate psychometric properties. Such accumulated results will help the systematic and effective development of summarization items. They will also help in establishing quality items for an assessment tool for summarization skills.

CONCLUSION

This study revealed that distractor characteristics had a huge influence on the difficulty of summarization items for L2 learners. Editing distractors has been considered difficult, but cognitive theories and findings provide us with a guideline to identify the source of difficulty of test items. Summarization skills are very important for L2 learners, so quality measurement and assessment is required to keep track of their skills. Summarization items after controlling the difficulties with cognitive components can qualify the measurement of these skills.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The ERB Board in Research Division of the NCUEE. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

TT designed the study, developed experimental summarization items, analyzed and interpreted data, wrote the initial draft of the manuscript, and also agrees to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

FUNDING

This work was supported by JSPS KAKENHI Grant-in-Aid for Research Activity Start-up (Grant Number 18H05826).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2020.00009/full#supplementary-material

REFERENCES

Anderson, V., and Hidi, S. (1988). Teaching students to summarize. *Educ. Leadersh*. 46, 26–28.

Baghaei, P., and Ravand, H. (2015). A cognitive processing model of reading comprehension in English as a foreign language using the linear logistic test model. *Learn. Individ. Differ*. 43, 100–105. doi: 10.1016/j.lindif.2015.09.001

Bejar, I. I. (1993). "A generative approach to psychological and educational measurement," in *Test Theory for a New Generation of Tests*, eds N. Frederiksen, R. J. Mislevy, and I. I. Bejar (Hillsdale, MI: Lawrence Erlbaum Associates, Inc.), 323–357.

Bond, T. G., and Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences. 3rd Edn*. Mahwah, NJ: L. Erlbaum. doi: 10.4324/9781315814698

Brown, A. L., and Day, J. D. (1983). Macrorules for summarizing texts: the development of expertise. *J. Verbal Learn. Verbal Behav*. 22, 1–14. doi: 10.1016/S0022-5371(83)80002-4

Brown, J., and Yamashita, S. (1995). English language entrance exams at Japanese universities: what do we know about them? *JALT J*. 17, 7–30.

Case, S. M., and Swanson, D. B. (1993). Extended-matching items: a practical alternative to free- response questions. *Teach. Learn. Med*. 5, 107–115. doi: 10.1080/10401339309539601

Daniel, R. C., and Embretson, S. E. (2010). Designing cognitive complexity in mathematical problem-solving items. *Appl. Psychol. Meas*. 34, 348–364. doi: 10.1177/0146621609349801

Das, D., and Martins, A. F. (2007). A survey on automatic text summarization. *Lit. Surv. Lang. Stat. Course CMU* 4, 192–195.

Drum, P. A., Calfee, R. C., and Cook, L. K. (1981). The effects of surface structure variables on performance in reading comprehension tests. *Read. Res. Q*. 16, 486–514. doi: 10.2307/747313

Embretson, S., and Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *J. Educ. Meas*. 38, 343–368. doi: 10.1111/j.1745-3984.2001.tb01131.x

Embretson, S. E. (1983). Construct validity: construct representation versus nomothetic span. *Psychol. Bull*. 93, 179–197. doi: 10.1037/0033-2909.93.1.179

Embretson, S. E. (1999). Generating items during testing: psychometric issues and models. *Psychometrika* 64, 407–433. doi: 10.1007/BF02294564

Embretson, S. E. (2002). "Measuring human intelligence with artificial intelligence," in *Cognition and Intelligence*, eds R. J. Sternberg and J. E. Pretz (New York, NY: Cambridge University Press), 251–267. doi: 10.1017/CBO9780511607073.014

Embretson, S. E., and Daniel, R. C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychol. Sci. Q*. 50, 328–344.

Embretson, S. E., and Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. *Appl. Psychol. Meas*. 11, 175–193. doi: 10.1177/014662168701100207

Enright, M. K., and Sheehan, K. M. (2002). "Modeling the difficulty of quantitative reasoning items: Implications for item generation," in *Item Generation for Test Development*, eds S. H. Irvine and P. C. Kyllonen (New York, NY: Routledge), 129–157.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychol*. 37, 359–374. doi: 10.1016/0001-6918(73)90003-6

Freedle, R., and Kostin, I. (1991). The prediction of SAT reading comprehension item difficulty for expository prose passages. *(ETS Research Report, RR-91-29)*. Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.1991.tb01396.x

Freedle, R., and Kostin, I. (1993). The prediction of GRE reading comprehension item difficulty for expository prose passages for each of three item types: main ideas, inferences and explicit statements. *(ETS Research Report, RR-93-13)*. Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.1993.tb01524.x

Gierl, M. J., and Lai, H. (2012). "Using weak and strong theory to create item models for automatic item generation," in *Automatic Item Generation: Theory and Practice*, eds M. J. Gierl and T. M. Haladyna (New York, NY: Routledge), 26–39. doi: 10.4324/9780203803912

Gorin, J. S. (2005). Manipulating processing difficulty of reading comprehension questions: the feasibility of verbal item generation. *J. Educ. Meas*. 42, 351–373. doi: 10.1111/j.1745-3984.2005.00020.x

Gorin, J. S., and Embretson, S. E. (2006). Item diffficulty modeling of paragraph comprehension items. *Appl. Psychol. Meas*. 30, 394–411. doi: 10.1177/0146621606288554

Gorin, J. S., and Embretson, S. E. (2012). "Using cognitive psychology to generate items and predict item characteristics," in *Automatic Item Generation: Theory and Practice*, eds M. J. Gierl and T. M. Haladyna (New York, NY: Routledge), 136–156.

Haladyna, T. M., and Rodriguez, M. C. (2013). *Developing and Validating Test Items*. New York, NY: Routledge. doi: 10.4324/9780203850381

Hare, V. C., and Borchardt, K. M. (1984). Direct instruction of summarization skills. *Read. Res. Q*. 20, 62–78. doi: 10.2307/747652

Hidi, S., and Anderson, V. (1986). Producing written summaries: task demands, cognitive operations, and implications for instruction. *Rev. Educ. Res*. 56, 473–493. doi: 10.3102/00346543056004473

Jeon, M., Rijmen, F., and Rabe-Hesketh, S. (2014). Flexible item response theory modeling with flirt. *Appl. Psychol. Meas*. 38, 404–405. doi: 10.1177/0146621614524982

Johns, A. M., and Mayes, P. (1990). An analysis of summary protocols of university ESL students. *Appl. Linguist*. 11, 253–271. doi: 10.1093/applin/11.3.253

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *J. Educ. Meas*. 50, 1–73. doi: 10.1111/jedm.12000

Karbalaei, H. R., and Rajyashree, K. (2010). The impact of summarization strategy training on university ESL learners' reading comprehension. *Int. J. Lang. Soci. Cult*. 30, 41–53.

Keck, C. (2006). The use of paraphrase in summary writing: a comparison of L1 and L2 writers. *J. Second Lang. Writ*. 15, 261–278. doi: 10.1016/j.jslw.2006.09.006

Khoshnevis, I., and Parvinnejad, S. (2015). The effect of text summarization as a cognitive strategy on the achievement of male and female language learners'

reading comprehension. *Int. J. Learn. Dev*. 5, 57–75. doi: 10.5296/ijld.v5i3.8271

Kikuchi, K. (2006). Revisiting English entrance examinations at Japanese universities after a decade. *JALT J*. 28, 77–96.

Kim, S. A. (2001). Characteristics of EFL readers' summary writing: a study with Korean university students. *Foreign Lang. Ann*. 34, 569–581. doi: 10.1111/j.1944-9720.2001.tb02104.x

Kintsch, W., and van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychol. Rev*. 85, 363–394. doi: 10.1037/0033-295X.85.5.363

Kobayashi, M. (2002). Method effects on reading comprehension test performance: text organization and response format. *Lang. Test*. 19, 193–220. doi: 10.1191/0265532202lt227oa

Kozminsky, E., and Graetz, N. (1986). First vs second language comprehension: some evidence from text summarizing. *J. Res. Read*. 9, 3–21. doi: 10.1111/j.1467-9817.1986.tb00107.x

Lai, H., Alves, C., and Gierl, M. J. (2009). "Using automatic item generation to address item demands for CAT," in *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing,* eds D. J. Weiss. Retrieved from http://docplayer.net/140964378-Using-automatic-item-generation-to-address-item-demands-for-cat-hollis-lai-cecilia-alves-and-mark-j-gierl-university-of-alberta.html (accessed April 16, 2019).

Lane, S., Raymond, M. R., Haladyna, T. M., and Downing, S. M. (2016). "Test development process," in *Handbook of Test Development, 2nd Edn,* eds S. Lane, M. R. Raymond, and T. M. Haladyna (New York, NY: Routledge), 3–18.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am. Psychol.* 50, 741–749. doi: 10.1037/0003-066X.50.9.741

Mosenthal, P. B. (1998). Defining prose task characteristics for use in computer-adaptive testing and instruction. *Am. Educ. Res J*. 35, 269–307. doi: 10.3102/00028312035002269

Ozuru, Y., Rowe, M., O'Reilly, T., and McNamara, D. S. (2008). Where's the difficulty in standardized reading tests: the passage or the question? *Behav. Res. Methods* 40, 1001–1015. doi: 10.3758/BRM.40.4.1001

Pearson, P. D., Roehler, L. R., Dole, J. A., and Duffy, G. G. (1992). "Developing expertise in reading comprehension," in *What Research Says to the Teacher*. eds S. J. Samuels and A. E. Farstup (Newark, NJ: International Reading Association), 145–199.

Pugh, D., De Champlain, A., Gierl, M., Lai, H., and Touchie, C. (2016). Using cognitive models to develop quality multiple-choice questions. *Med. Teach*. 38, 838–843. doi: 10.3109/0142159X.2016.1150989

R Core Team (2018). *R: A Language and Environment for Statistical Computing [Computer Software Manual]*. Vienna, Austria. Retrieved from https://www.R-project.org/ (accessed October 1, 2019).

Sober, E. (2002). Instrumentalism, parsimony, and the Akaike framework. *Philos. Sci*. 69, S112–S123. doi: 10.1086/341839

Sonnleitner, P. (2008). Using the LLTM to evaluate an item-generating system for reading comprehension. *Psychol. Sci. Q*. 50, 345–362.

Terao, T., and Ishii, H. (2019). Analyses of distractors in english summarizing test items: focusing on cognitive processes,"in *Paper Presented in the Annual Meeting of the National Council on Measurement in Education* (Toronto, ON).

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach*. Mahwah, NJ: Laurence Erlbaum Associates. doi: 10.4324/9781410611697

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

Examples of test items used in this study (Q21-3, Q22-7, Q22-8, and Q22-9)

Text:

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.

CMC version:

(lead-in) Which of the following statements is the best summary of Paragraph (7)?

(1) Recent information technology made immigrants to the United States feel homesick because a regular phone call at lower prices make them feel like they are separated from the other members of their family.

(2) Modern technology that enables us to regularly contact people at a distance tends to remind immigrants of their families, and result in stronger feelings of displacement in them.

(3) Cheaper international calling made Carmen frequently call her family to know what they were doing, to think of them all the time, and therefore feel alone and lonely.

EM version:

(lead-in) These are summaries of the Paragraph (7). Classify every sentence into the following categories listed below.

(1) Recent information technology made immigrants to the United States feel homesick because a regular phone call at lower prices make them feel like they are separated from the other members of their family.

(2) Modern technology that enables us to regularly contact people at a distance tends to remind immigrants of their families, and result in stronger feelings of displacement in them.

(3) Cheaper international calling made Carmen frequently call her family to know what they were doing, to think of them all the time, and therefore feel alone and lonely.

(an option list)

(a) A correct summary.

(b) An incorrect summary because detailed information was included.

(c) An incorrect summary because an expression was mistakenly generalized.