Language Testing in Asia

CASE STUDY Open Access

Assessing Japanese junior high school students' English achievement through computer-based testing in the classroom: a case of integrated reading-into-writing continuous task

Noriyasu Niimi* and Nobukazu Matsuura

*Correspondence:
d222901@hiroshima-u.ac.jp

Graduate School of Humanities and Social Sciences, Hiroshima University, 1-1-1 Kagamiyama, Higashihiroshima, Hiroshima 739-8524, Japan

Abstract

Introduction: This paper describes the exploratory case and initial evaluation of the computer-based testing (CBT) prototype. The advantage of CBT over paper-based testing (PBT) is that it allows us to control the order of questions and provides test takers with continuous tasks capturing their thought processes. Additionally, their response process data such as response time (RT) can be obtained. Taking advantage of these, we created a CBT prototype in the classroom for Japanese junior high school students.

Case description: A CBT model was created to assess integrated reading and writing ability and was administered to 32 junior high school students. Their process achievement and the relation between the process response and writing quality were analyzed. Students' RT for each screen was analyzed using hierarchical cluster analysis.

Discussion and evaluation: We identified not only students facing difficulties at each stage of a series of thought processes but also five clusters that include students spending too much time reading source texts or organizing their ideas. We suggest how CBT can be developed to identify students with difficulties and applied to teaching.

Conclusions: CBT has the possibility of detecting students who are able to complete the language performance task by controlling the order of answers, asking questions sequentially, and obtaining RT effectively.

Keywords: PBT, CBT, Integrated reading-into-writing, Criterion-referenced test, Classroom assessment, Japanese junior high school students

Introduction

Language proficiency has been regarded as one of the most important skills to be cultivated in school education worldwide. In Japan, the National Curriculum Standards have positioned language proficiency as a fundamental competency for learning and emphasized its importance. Since English language lessons in Japan are a major subject that

aims to improve language proficiency, the central goal is to develop the "students' competencies that form the communication such as understanding, expressing, and communicating simple information and thoughts (MEXT, 2017, p. 10)" though language activities in English. Thus, the importance of developing language proficiency has been pointed out both domestically and internationally; however, in Japan, no large-scale English academic achievement tests had been conducted to determine the students' English language proficiency, therefore, a comprehensive understanding of their English language proficiency was not clarified.

In 2019, for the first time in Japan, the nationwide English academic achievement test was administered through paper-based testing (PBT) to about one million third-year junior high school students. This is a criterion-referenced test created based on the goals outlined in the National Curriculum Standards. The test aims to grasp the actual academic performance of junior high school students nationwide and to use the results to improve instruction. The results of the test revealed a variety of issues, particularly the correct response rates for the reading-into-writing task and independent writing task being 11.6% and 1.9%, respectively (MEXT, 2019). This indicates that the English writing ability of Japanese junior high school students is significantly low, and there is a need for improved instruction that contributes to improving writing ability. However, such criterion-referenced PBT provides little useful information for improving instruction and learning.

Currently, the development goals for the thinking abilities for writing in English language in Japan are: (a) to be able to write coherently about everyday topics while organizing facts, their own thoughts, feelings, etc. by using simple words, phrases, and sentences; (b) to be able to write about their thoughts, feelings, and the reasons for them about social topics they have heard about or read by using simple words, phrases, and sentences. Teachers must develop and implement more effective means of instruction to help students learn the abilities indicated in these objectives. When assessing whether students have learned the aforementioned thinking abilities through PBT, the data obtained is limited to the English writings that the students actually produced; their quality is used to determine whether the goal has been reached. Hence, if the number of students reaching the goal is small, we know that there are problems in the goal development. In other words, for teachers who are aware of the instructional process and develop instructions to help students reach their goals, product-only information is not sufficient for improvement. To link assessment to instruction, it is necessary to have information that identifies at what point in the process each student has reached the product. Once this information is obtained, it can lead to the development of group-specific and individualized instruction. Therefore, it is necessary to develop new assessment methods that will provide an elaborate picture of student achievement.

International large-scale educational assessments have transitioned from PBT to computer-based testing (CBT). The Program for International Student Assessment (PISA) introduced CBT in 2015 to set more interactive questions in various contexts. The National Assessment of Educational Progress (NAEP) has conducted digitally based assessments since 2017, which involve the test-takers engaging in problem-solving by utilizing knowledge and skills. In Japan, the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) is proposing to administer CBT for the nationwide

academic achievement test for third-year junior high school students in a few years. In response to the shift to CBT for the national achievement test, the CBT system provided by the MEXT is currently developing questions that are CBT versions of previous PBT questions and questions that take advantage of CBT features. At present, schools that wish to use this system can only access it from their students' ICT devices to conduct unit tests and learning. In the future, however, teachers are expected to use CBT for periodic classroom tests, for example, by creating tests, allowing students to access and answer them from their ICT devices, and accumulating, analyzing the answer data to gain a detailed understanding of students' learning status, and improving instruction based on the obtained results. It is expected that in the future, CBT will be utilized for periodic tests conducted in classrooms.

Many empirical studies in language testing and assessment have examined whether performances on PBT and CBT differ by directly changing existing PBTs to CBTs since the early 2000s. Recently, however, research interests have begun to shift to assessments taking advantage of CBT features including those that are difficult to capture through PBT, that grasp the test-takers' progress from several perspectives, and that automatically score free descriptions (e.g., Douglas & Hegelheimer, 2007; Jamieson, 2005). Against this social and academic background, empirical studies are being conducted on specific questions and assessment methods through CBT in various fields (e.g., Masukawa et al., 2021; Ukon et al., 2019; Yamashita, 2017).

However, to the best of our knowledge, investigations have not yet been conducted on the assessment of English achievement for Japanese junior high school students through CBT; very few empirical findings have been compiled on the questions and assessment methods that take advantage of CBT features. Therefore, this paper reports the exploratory case and initial evaluation of the prototypical CBT writing task, an alternative to the conventional PBT in the classroom. The purpose of this paper was addressed in the research question—to what extent can CBT visualize students' thought processes and provide pedagogical information that informs instruction?

The advantages of CBT over PBT

Although computerizing tests would involve some changes, test developers and evaluators should clarify the expected differences between PBT and CBT to introduce CBT for assessing the next generation. The advantages of CBT over PBT are as follows: first, CBT allows us to pose questions and answers in a variety of ways such as via audio, computer graphics, video, and dynamic objects. Second, CBT also allows us to control and measure each item and obtain answer operation logs, which enable us to understand the thought processes of the test-takers (e.g., National Center for university entrance examinations, 2021; Nishigori et al., 2017). Third, by analyzing the log data of test-takers along the thought process to the final answer, CBT allows us to clarify the tendency of students to make certain types of mistakes, highlights their insufficient understanding of the question items, and provides useful information for improving instruction and diagnosing learning (e.g., The Japan Association for Research on Testing, 2007).

Most of the CBT studies on writing have been conducted by converting the PBT into CBT and analyzing whether there are any differences in the quality of writing between them (e.g., Brunfaut et al., 2018; Chan et al., 2018). Few studies have developed questions

that take advantage of the CBT features and analyzed the test-takers' writing. In recent years, many studies have attempted to elucidate learners' writing process by analyzing keystrokes obtained by having them type English on a computer (e.g., Bennett et al., 2022; Talebinamvar & Zarrabi, 2022). These studies, however, are rarely intended to provide teachers with a simple way to identify students' obstacles in the process leading up to the product or connect them to classroom instruction. CBT questions are not designed with the students' thought processes in mind, which the teachers/evaluators aim to elicit.

At present, only one case study in Japan—by Masukawa et al. (2021)—has attempted to utilize the CBT features to assess ability. According to Masukawa et al. (2021), the advantage of CBT is that it can control and record the answering process—which is not possible in PBT—and bring out the cognitive process intended by evaluators better than PBT can. CBTs of the revised version of the Japanese reading comprehension questions (that made it impossible to see the following question and revise the answer) and the conventional version without such a function were administered to 40 middle-ranking university students and 39 high-ranking high school students. The university students scored higher in the conventional version by using strategies such as transcribing, which the evaluators did not anticipate, while their scores were lower in the revised version as a result of following the answering process as per the expectations of the evaluators. Conversely, high school students followed the answering process intended by the evaluators in both versions; their scores were higher in the revised version and lower in the conventional one.

Overall, PBT can observe only a fragmented part of an activity, while CBT can reproduce an activity on a computer and observe it intermittently. Therefore, it is possible to evaluate a series of thought processes that lead to problem-solving.

The use of response process data in educational evaluation

Another major advantage of using CBT is that it provides response process data, unlike PBT. Response process data refers to not only the data related to thought processes, strategies, and approaches used when reading, interpreting, and forming solutions to assessment tasks, but also to the behavior of the examinee (Ercikan & Pellegrino, 2018). For instance, Kitazawa and Shirouzu (2020) developed a CBT system that can measure the behavior and solution strategies between the question and the text for a previous Japanese language test of the National Center University Entrance Examination. They found three types of processes by visualizing the solution processes: reading the text to the end after checking the lead sentence, answering questions while checking the questions and options, and reading and answering the text after checking the questions. However, the effect of the type of solution process on the percentage of correct answers could vary considerably.

One of the typical elements of response process data is response time (RT), which is obtained by measuring the amount of time taken by a test-taker to solve a single problem or the amount of time taken between starting and finishing an item (e.g., Sahin & Colvin, 2020). Gong et al. (2020) automatically measured preparation time (PT), execution time (ET), and mean execution time (MET) per answering event on a scientific inquiry task used in a pilot study of the NAEP. PT reflects the process of understanding a question and

planning an answer, ET reflects the process of using strategies in answer writing, and MET reflects information about the efficiency of answer writing. Test takers who spent more time on PT tended to have lower scores on the task. In another RT-based writing study, Xu and Ding (2014) analyzed the time spent on prewriting in computer-assisted writing conducted among 24 Chinese learners of English and found that less-skilled writers spent more time on the prewriting stage than did skilled writers. These findings reveal that the ability of test-takers to efficiently plan their answers affects the outcome of the task. Lee et al. (2019) conducted a hierarchical cluster analysis on a single simulation-based task consisting of 11 items—using the RTs of the test takers on each item as a variable—and found three clusters: slow, fast, and moderate RT patterns. The RT patterns among the clusters showed a similar trend for most items, but certain items deviated from the trend and more time was spent on them. In addition, each cluster had lower scores for items that took more time.

Thus, if the advantage of assessing academic achievement through CBT is that it can evaluate not only the products of test-takers but also the production process, it is then possible to evaluate students' thought processes by controlling the order of answers and having them follow the thought process intended by the evaluator or by effectively obtaining response process data. By using CBT, we attempted to visualize students' thought processes and evaluate them for instruction by asking sequential questions following the thought process that occurs when they carry out a certain language activity.

Case description

Design of CBT

In this study, we decided to assess students' integrated skills related to reading information and writing down their opinions. These skills include understanding and selecting information according to the purpose of writing, organizing and integrating the selected content, and linking information to one's own ideas (e.g., Chan et al., 2015; Knoch & Sitajalabhorn, 2013; Plakans, 2008).

We chose to assess integrated skills for two reasons. First, MEXT (2017, p. 53) highlights the importance of English lessons in "expressing facts, one's own thoughts, and feelings through speaking and writing by selecting and extracting information and expressions obtained by listening to and reading English on everyday and social topics;" thus, there is a growing focus on integrated skills. Second, language activity involves a long thought process, which likely makes use of the features of CBT. By being able to control the order of questions—such as by not being able to return to the previous question or screen—it is possible to evaluate the thought process. In addition, we tried to evaluate the integrated skills that require a long thought process before answering by considering the possibility of evaluating the thought process itself using response process data such as RT and analyzing it for each item. Based on the evaluation target of this study, Fig. 1 shows the core thought process assumed in this study, referring to Matsuura (2021). Based on the contents of the foreign language lessons' expert meeting

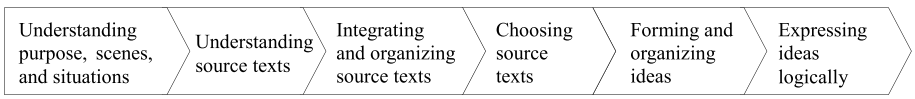

Fig. 1 The core thought process assumed in this study

on the preparation of the Japanese curriculum guidelines, Matsuura analyzed the way the thinking abilities are to be learned in foreign language lessons, as indicated in the National Curriculum Standards. Matsuura presented a series of core processes involving thinking abilities in foreign language lessons, which are as follows: (a) understanding purposes, situations, and circumstances, (b) extracting information according to the purpose, etc. (c) organizing and integrating knowledge and information, (d) forming and organizing opinions and ideas, and (e) expressing logically. The goal of this study is to create CBT questions for Japanese junior high school students that are compliant with the goals outlined in the National Curriculum Standards. Therefore, the thought process proposed by Matsuura is a useful reference for the creation of the CBT as it clarifies the thinking abilities, which are the first component of the academic skills to be learned in English language lessons in Japanese school education, and shows the core thought process involved in these abilities.

Based on the assessment abilities and the core thought processes described above, we implemented a continuous task on the TAO open-source CBT platform (Table 1). After answering the questions on each screen, the test-takers clicked on the "Next" button to proceed to the next screen and could not return to the previous screen.

In conventional PBT, students are often presented with an English text on a certain topic and asked to write their thoughts and opinions on it using English. This type of questioning, however, only evaluates English writing as a product; evaluators cannot obtain information on students' thought processes leading to English writing, such as whether they were able to understand the presented English text or whether they formulated their own ideas. To understand students' English achievements that cannot be captured by the conventional PBT, the CBT was designed to ask questions about the process leading to English writing, as shown in Table 1. Each of the screens shown in Table 1 is described in detail below, with figures.

With the theme of improving English, Screen 1 displays a scene/situation, as shown in Fig. 2. Screen 2 displays source text A written by Mr. A and Source text B written by Ms. B, which students are expected to read, as shown in Fig. 3. The source texts used in this study are summarized in Table 2. They both highlight the value of improving English skills, although Mr. A and Ms. B present different points. On Screen 3, students are asked to choose from four options regarding the two types of source texts (Fig. 4). Next, on Screen 4, they are asked to choose from three options related to the main points of Mr. A's opinion that it is better to study abroad and Ms. B's opinion that it is possible to

Table 1 Summary of the continuous task adopted in this study

S	Task description	Answer type
1	Understand the scene and situation	
2	Read two types of source texts	
3	Answer the question to read the outline of the two types of source texts	Multiple-choice (4 choices)
4	Answer the question to read the main points of the two types of source texts	Multiple-choice (3 choices)
5	State your position on each source text	Multiple-choice (2 choices)
6	Formulate and organize your own opinions, thoughts, and reasons	Essay style (in Japanese)
7	Connect the source texts with your ideas and state them logically in English	Essay style (in English)

"S" indicates screen

Fig. 2 Screen 1: Understanding the scene and situation

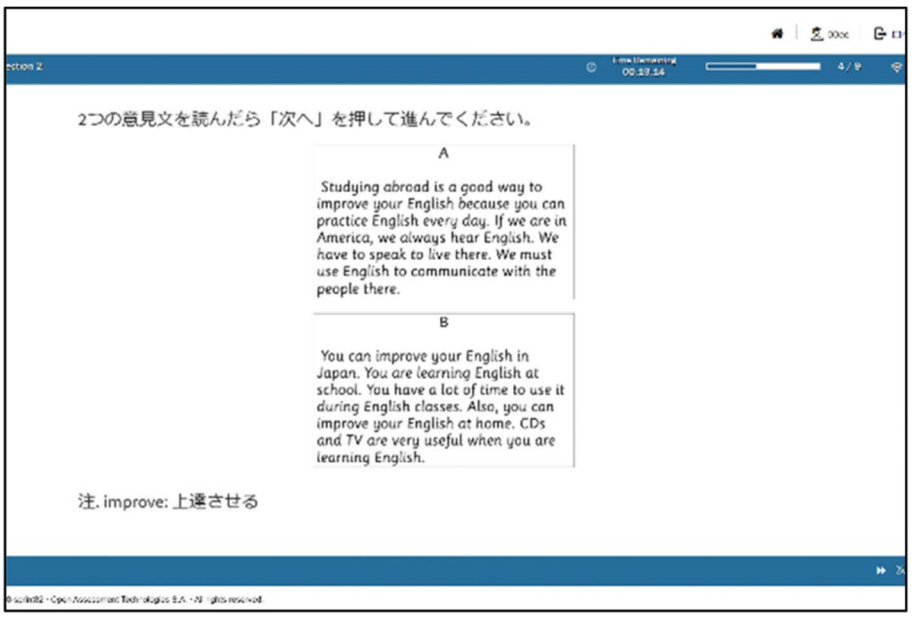

Fig. 3 Screen 2: Reading two types of source texts

Table 2 Summary of source texts used in this study

	Topic	Main point	Number of sentences	Token frequency
A	How can you improve your English?	You should study abroad	4	43
B		You can improve even in Japan	5	44

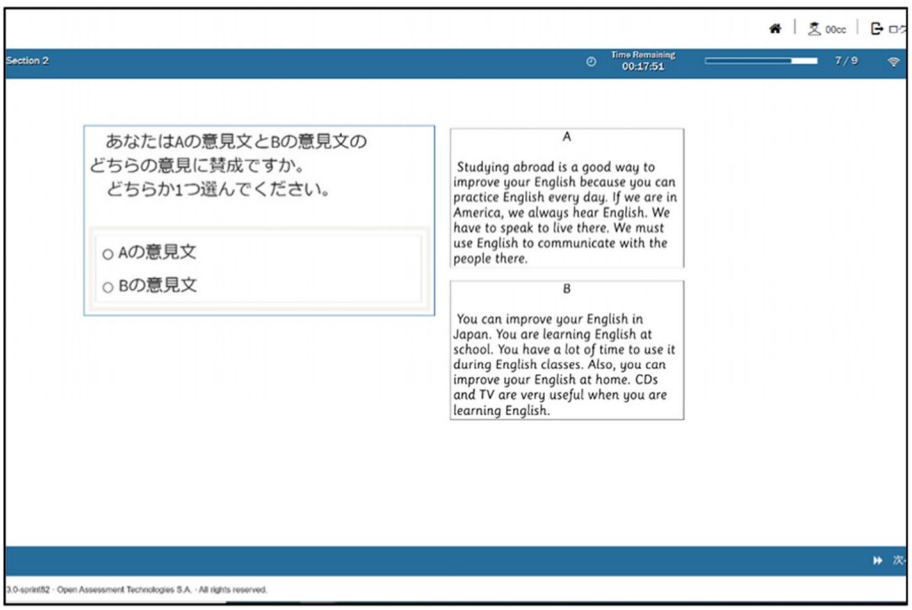

Fig. 4 Screen 3: Choose the outline of the source texts

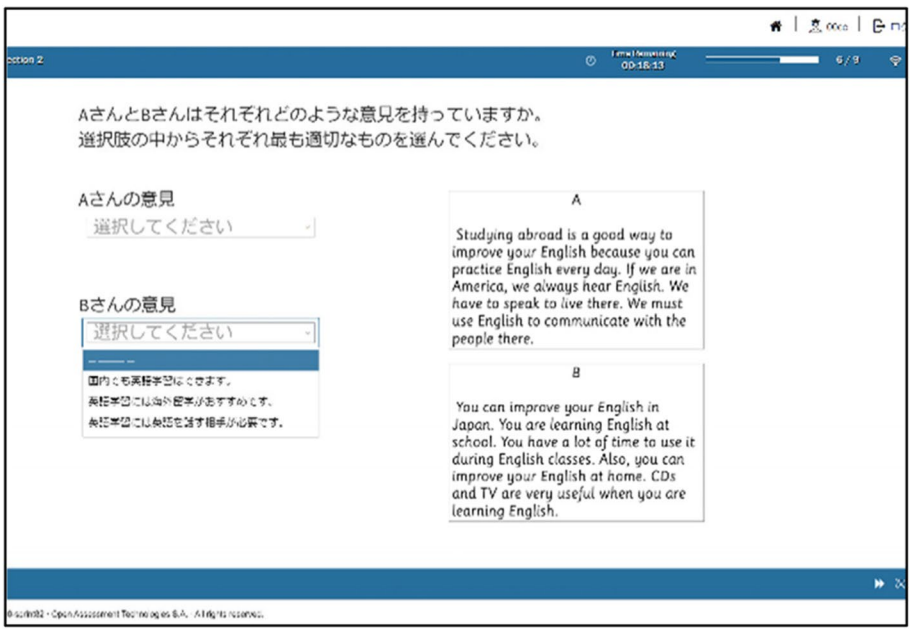

Fig. 5 Screen 4: Choose the main points of each

learn English in Japan (Fig. 5). Screens 1 to 4 focus on comprehension of the scene/situation and source texts, while Screen 5 focuses on producing English writing.

On Screen 5, students are asked to choose whether they agree with the opinion of Mr. A or Ms. B (Fig. 6). On Screen 6, they have to type the reasons for their selection in Japanese, comparing it with the source text of the other option (Fig. 7). Although students can formulate and organize their opinions in their minds, Screen 6 was created to

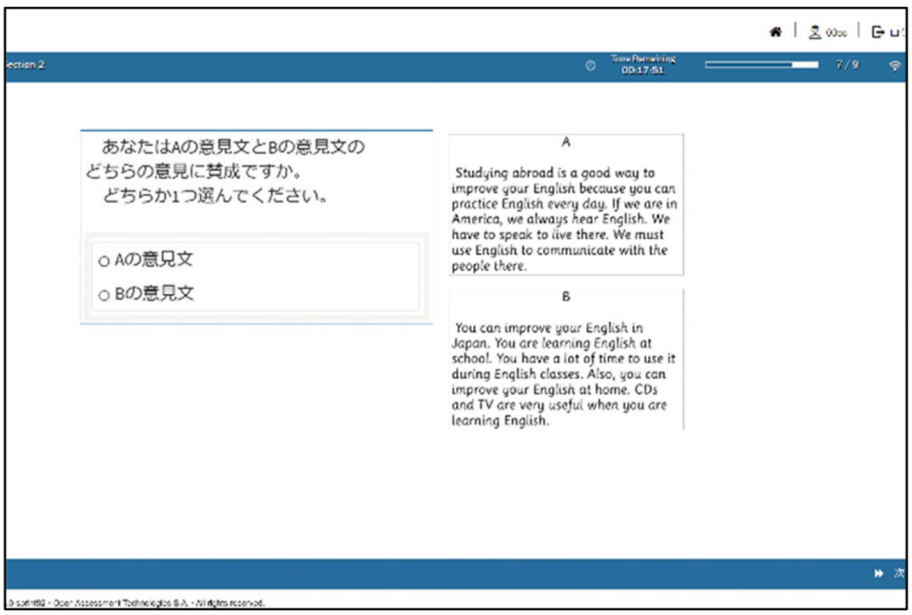

Fig. 6 Screen 5: Choose the source you agree with

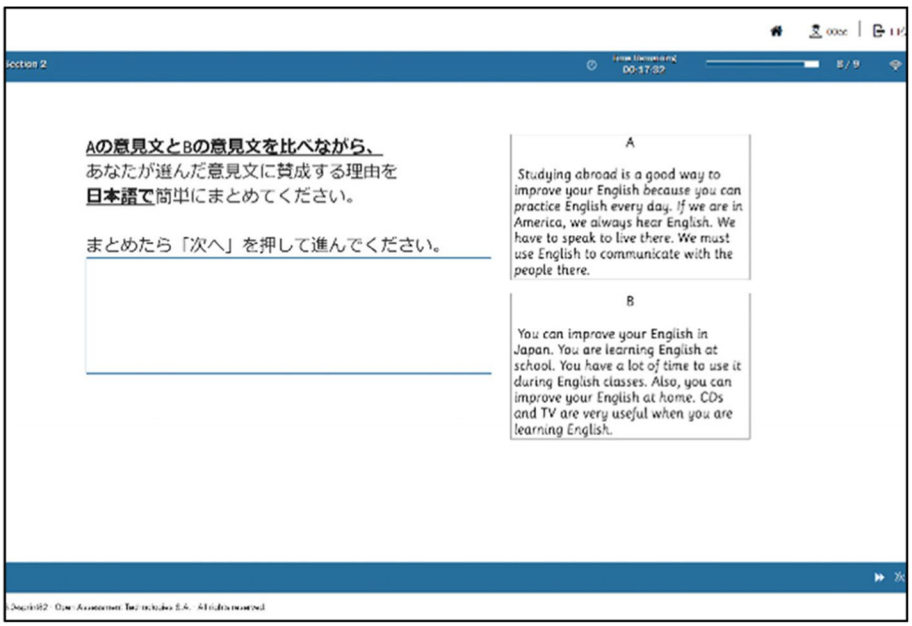

Fig. 7 Screen 6: Organize your ideas in Japanese

understand what the students were trying to express in English. Screen 7 requires students to type their opinions on what they need to do to improve their English, comparing their opinions with those of Mr. A and Ms. B (Fig. 8). The maximum number of words that can be typed was set at 65. Finally, the CBT is completed by clicking on the "End Test" button at the bottom right of the screen.

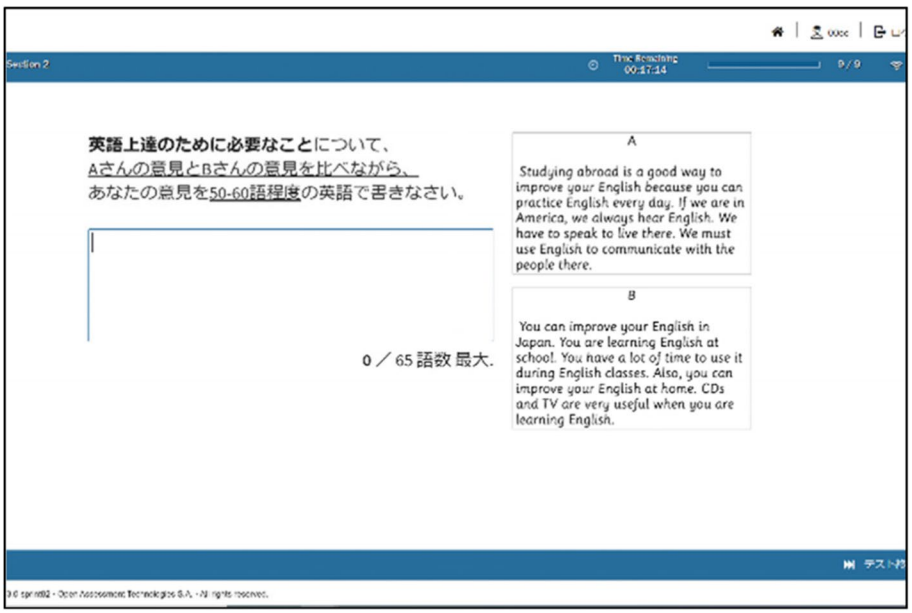

Fig. 8 Screen 7: State your ideas logically in English

Students' RTs, which can be obtained in CBTs, are automatically measured when the "Next" button is clicked to proceed to the next screen. In other words, RTs reveal how much time the test-takers spent on each screen to reach an answer.

We asked three graduate students majoring in English education to actually undergo the CBT and judge whether it was appropriate to assess integrated skills for third-year junior high school students. Based on their opinions, some corrections were made to the source texts and instructions to ensure the content validity.

Participants

A total of 32 students (15 boys and 17 girls) participated in the CBT. They were third-year students from the same class of a public junior high school in Hiroshima, Japan. According to their self-reports, one student had passed grade Pre-2 EIKEN Test, which is widely recognized as the domestic public test on practical English proficiency in Japan, four students had passed grade 3, four students had passed Grade 4, and one student had passed grade 5.

Before the CBT, the typing speeds of the students were measured using Typing Test Pro (https://pro.typingtest.com/), an online typing test for English. Students were asked to type the English text displayed on the screen as fast and accurately as possible in one minute, and their gross speed, accuracy, and net speed were measured. Gross speed is the number of words typed per minute (WPM), accuracy is the percentage of typing accuracy, and net speed is the number of words typed, adjusted for typing accuracy. Table 3 shows the descriptive statistics of the students' typing net and gross speeds, and Fig. 9 depicts the beeswarm plots. Among 32 students, 9 had gross and net speeds of zero WPM due to a network problem; these scores were treated as missing values and excluded from the analysis of typing speed.

Table 3 Descriptive statistics of students' typing net and gross speeds

	n	*M*	Median	SD	IQR
Net. speed	23	11.57	8.00	11.78	5.50
Gross. speed	23	12.70	9.00	11.66	8.00

The unit of measurement is words per minute. *IQR* refers to interquartile range

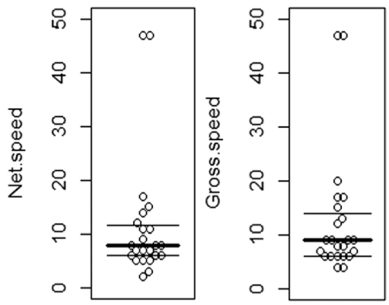

Fig. 9 Beeswarm plots representing students' typing net and gross speeds. ○ represents each student's word per minute. Solid lines represent the first and third quartiles. Thick lines represent the median

The participating students were able to type an average of only about 12 WPM. Judging from the fact that Japanese junior high school students were able to type approximately 17.4 Japanese characters per minute in the Information Use Proficiency Survey conducted by MEXT in 2013, the students' typing speed was quite slow. However, two students were able to type 47 WPM. Gross speed was slightly higher than net speed, confirming that in addition to slow typing speed, the students were also inadequate in handling keyboard operations such as typing errors, case conversion errors, and space errors. The above demographic information will not be used in the analysis but will be discussed.

Procedure

In mid-May 2021, the students underwent CBT using their laptops (Dynabook K50) in the classroom during class hours. It was conducted in the following order: pre-survey explanation, logging in to the test site, typing test for warm-up, and the CBT. In the pre-survey explanation, the students were instructed to press the "Next" button as quickly as possible after answering the questions to proceed to the next screen. They were informed that once they proceeded to the next screen, they could not return to the previous screen. We also informed them that CBT questions were to be read and answered in English and they would have about 25 min to answer the questions.

Before conducting the study, we requested the school administrators to cooperate and obtained their consent after providing them with oral and written explanations of the study and ensuring the protection of personal information. This study was approved by the Joint Committee for Ethical Review of the Graduate School of Humanities and Social Sciences, Hiroshima University (Approval No. 2021019).

Scoring

The multiple-choice questions in Screens 3 and 4 were scored automatically based on whether they were correct or incorrect. The essay questions on Screens 6 and 7 were scored by two graduate students, majoring in English language education, using an overall rating scale ranging from 0 to 5. For the rating scale, Yang's (2012) scoring rubric for the reading-based writing and graph-based writing tasks was used with some modifications to fit the essay questions created in this study (see Table 9 in Appendix). Prior to the start of scoring, a rater training session was conducted. The researcher explained the CBT questions and the rating scale and clarified the raters' doubts. In addition, while referring to the model English essays for each score, we discussed scoring based on the rating scale; eventually, scores were assigned based on a common understanding of the English essays for each score. Answers with a difference in points were discussed and re-rated by the researcher as the third rater. Cohen's weighted κ coefficient, which represents inter-rater reliability, was 0.81 for the question on Screen 6 and 0.88 for the question on Screen 7; thus, high reliability was ensured.

Data analysis

Since this study is a small case study, descriptive analysis and data visualization were actively conducted. First, regarding the quality of English writing produced by the students, we calculated descriptive statistics of the students' English writing scores on Screen 7 and visualized the distribution of scores. We showed several actual writings corresponding to those scores. Then, regarding the process leading to the English writing product, we showed the total number of correct and incorrect answers for the multiple-choice questions on Screen 3 through 5 and the distribution of scores for the question on Screen 6, in which students summarized their opinions in Japanese to understand the overall trend.

Next, to identify the stage of thought process reached by each student in this study, when a question was answered incorrectly for the first time in the transition from Screen 3 to Screen 7, it was excluded from the analysis regardless of whether it was answered correctly in the next screen. Based on the results, students were grouped to easily determine the stage of the reading and writing process within which a group of students with difficulties existed. Furthermore, we visualized the relationship between the results of the answers in each stage of the thought process and the English scores obtained.

Third, we calculated descriptive statistics for the amount of time students spent on each screen. Hierarchical cluster analysis (Ward's method, Euclidean distance) using seven variables of the percentage of time spent on each screen relative to the total time spent on screens 1 through 7 was used to analyze the relationship between the RT characteristics of the clusters found and the mean, score distribution of English writing scores for that cluster.

Discussion and evaluation

Overall results for items

The score distribution of the students' writings typed in Screen 7 is shown in Fig. 10. Excerpts from the actual English writings corresponding to each score are presented below.

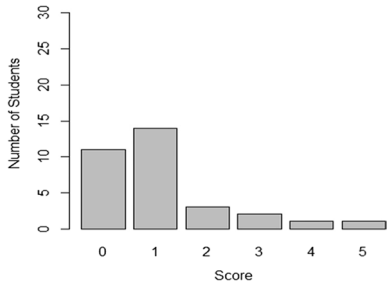

Fig. 10 Distribution of the scores in Screen 7

Table 4 Descriptive statistics of the scores in Screen 7 ($N = 32$)

	M	SD	Skewness	Kurtosis
Score	1.09	1.23	1.45	1.79

Excerpt 1 Student A's English writing (score 5).

I think we can improve our English in Japan. I have two reasons. First, we have English classes. We can hear English in it. Second, we must study other subjects. If I am in America, I cannot study these subjects hard

Excerpt 2 Student B's English writing (score 4).

I think that A's opinion is good because I can't to study hard. But if I am in America, I have to study and speak to live there. But I am in Japan, I am not to study hard

Excerpt 3 Student C's English writing (score 3).

I think B idea is very good. Many people can improve hear English. It can school and at home. We using TV or CDs to study English. We are lr

Excerpt 4 Student D's English writing (score 2).

I think you must learning English at school every day. And you have to more studying English. You should think that I can studying English

Excerpt 5 Student E's English writing (score 1).

I think is studying English talk is good

Table 4 shows the descriptive statistics of the scores. The number of students who scored 0 or 1 was 25 out of 32, indicating a floor effect. The conventional PBT, which assesses only production, can only provide information to this extent. The CBT in this study, however, provides a detailed picture of how far the students were able to proceed in their thought processes.

Regarding the results of the students' answers for each screen, Table 5 shows the number of correct and incorrect responses to the questions on Screen 3, which asks students to read the outline of source texts A and B, and Screen 4, which asks them to read the main points of source texts A and B. Figure 11 shows the distribution of scores for Screen 6, which asks students to summarize their opinions and thoughts

Table 5 Correct and incorrect responses in Screens 3 and 4 ($N = 32$)

	Task	Correct	Incorrect
Screen 3	Read the outline	25 (78.1%)	7 (21.9%)
Screen 4	Read the main points	17 (53.1%)	15 (46.9%)
	Read the main points of A	18 (56.3%)	14 (43.7%)
	Read the main points of B	30 (93.7%)	2 (6.3%)

Correct indicates the number who answered correctly both questions in the task "Read the main points"

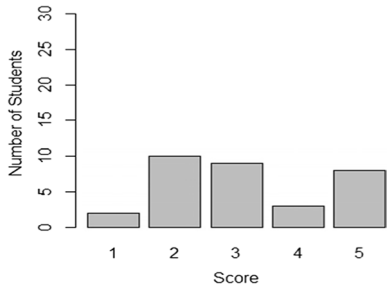

Fig. 11 Distribution of the scores in Screen 6

in Japanese. Overall, 25 out of 32 students (78%) were able to read the outline, but only 17 (53%) were able to read the main points, indicating greater difficulty in reading the main points than in reading the outline. As the number of correct and incorrect responses to the questions related to reading the main points of the source text A and B differed greatly, there may have been a difference in the difficulty level of English text or a problem in deciding on one option. However, considering that the options comprised three choices that are the same for both source texts A and B, it can be concluded that the students who were able to complete both questions were able to read the main points. In terms of summarizing one's opinion in Japanese, the scores varied from 1 to 5, and it was possible to judge whether the students were able to formulate and organize the contents they wanted to express.

Students' achievements based on the thought process

We ascertained the number of students who were able to reach which stage along the thought process. For this purpose, when a question was answered incorrectly for the first time in the transition from Screen 3 to Screen 7, it was excluded from the analysis, regardless of whether it was answered correctly on the next screen. The analysis revealed that seven out of 32 students (22%) answered Screen 3 incorrectly, which required them to read the outline, and were unable to read the outline of the two source texts. Of the remaining 25 students, 10 students (40%) answered Screen 3 correctly but Screen 4 incorrectly, indicating that although they were able to read the outline, they could not read the main points of each source text. Therefore, 15 out of 25 students (60%) were able to read the outline and main points of the source texts.

Furthermore, out of these 15 students, three (20%) who were able to read the outline and main points of the source texts were not able to formulate their opinions, thoughts, and reasons (scoring 2 points or less on Screen 6). Of the remaining 12 students, eight

Table 6 Each group and the characteristics discovered

Group	*n*	Characteristics
Group 1	7	This group is unable to even read the outline of the source texts
Group 2	10	This group is able to read the outline of the source texts, but is unable to read the main points of each of the two source texts
Group 3	5	This group is able to read the outline and main points of the two source texts and even express their position on each text, but is unable to form their own opinions, thoughts, and reasons
Group 4	7	This group is able to read the outline and main points of the two source texts, express their position on each source text, and even form their own opinions, thoughts, and reasons, but is unable to connect the source texts and their own thoughts and express them logically in English
Group 5	3	This group is fully able to connect source texts with their own ideas and state them logically in English

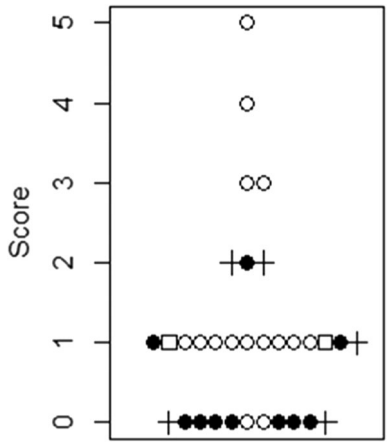

Fig. 12 Beeswarm plots representing the scores of the writing. ○ indicates students who could read the outline and main points, ● indicates those who could read the outline but not the main points, □ indicates those who could read the main points but not the outline, and + indicates those who could read neither the outline nor the main points

(67%) were able to read the outlines and main points of the source texts and also form their own opinions, thoughts, and reasons, but they were not able to connect them to the source texts and state them logically in English (scoring 2 points or less on Screen 7). Table 6 summarizes each group, the characteristics, and the number of students in the group we discovered.

Relation between the results of the thought process and the quality of English writing

The analysis of the relation between the results of each stage of the thought process and the quality of English writing is depicted in a beeswarm plot with different markers for each stage, as shown in Fig. 12.

We determined that students who answered all the questions in Screens 3 and 4 correctly were able to read the outline and main points; students who answered Screen 3 incorrectly but Screen 4 correctly were able to read the main points but not the outline, while students who answered both Screens 3 and 4 incorrectly were not able to read either. Only the students who were able to read the outline and main points (four students) scored 3 to 5 points in English writing, while students who could neither read the outline nor the main points (five students) scored 0 to 2 points. Conversely, 11 students

Table 7 Descriptive statistics of students' RTs (in seconds) on each screen ($N = 32$)

Screen	*M*	SD	Skewness	Kurtosis
1 (Understand scene and situation)	14.37	11.71	2.13	5.12
2 (Read two types of source texts)	102.05	50.18	0.94	0.65
3 (Choose the outline)	28.83	14.35	0.81	-0.22
4 (Choose each main point)	72.47	35.61	0.46	-1.14
5 (Choose the source text that you agree with)	19.73	22.54	2.80	8.67
6 (Organize your own ideas in Japanese)	371.49	176.01	0.86	0.94
7 (Write your own ideas in English)	560.8	284.99	0.36	0.73

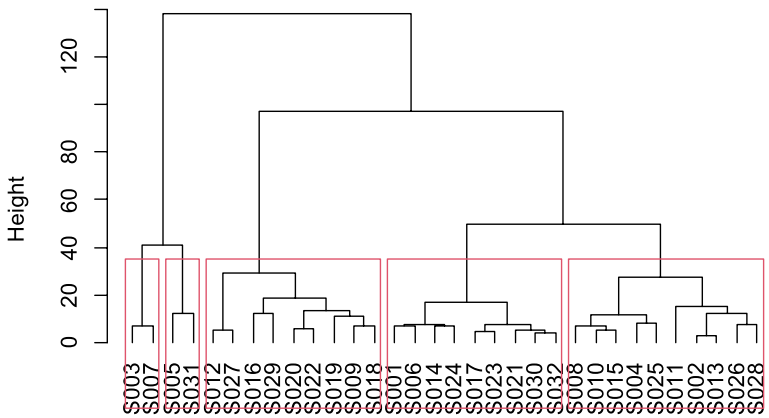

Fig. 13 Cluster dendrogram

were able to read the outline and main points but scored 0 or 1 in English writing. Out of them, three could not summarize their opinions and reasons in Japanese, while eight could. An example of one of the three students who could not summarize his/her opinions or reasons in Japanese and an excerpt of the English writing by the student (score 1) are shown below.

Excerpt 7 Student F's Japanese writing that could not summarize their opinions and reasons.

私は、Bの意見文の方がいいと思います。理由は家で英語を学ぶ方法について書いてあるからです [English translation: I think B's opinion is good because it mentions the way to learn English at home.]。

Excerpt 8 Student F's English writing (score 1).

I think B is good. became A is go to Amerika but B is learn English

Response time (RT)

The descriptive statistics of the RT of students on each screen are shown in Table 7.

To examine the results in more detail, a hierarchical cluster analysis (Ward method, Euclid distance) was conducted using the seven variables of percentages of each RT to total RT obtained from Screen 1–7. We visualized it with a dendrogram and finally decided that it was appropriate to divide it into five clusters (Fig. 13). Table 8 shows

the average percentage of each RT to total RT for Screens 1–7 for each cluster. We also plotted the average values for each cluster to visualize the characteristics, as shown in Fig. 14.

Evaluation

The CBT prototype in this study allowed us to identify students with difficulties in a multilayered manner by posing questions to determine their thought processes and by analyzing the relations among their answers based on their thought processes, RTs, and English writing scores. As a result of being able to analyze students' achievement using their thought processes, we were also able to identify the students who could not even read the outline, who could read the outline but not the main points of each source text, who could form their opinions, thoughts, and reasons but could not connect the source texts to their own ideas and state them logically in English, and those who could state their own ideas logically in English. Using PBT, it would be difficult to ascertain which of the above stages have been reached by the students because it does not necessarily require the test-takers to start from the beginning and the order of answers is left up to them. It is also possible to go back to previous questions and change an answer. Therefore, the test takers may not necessarily have solved the question as the evaluator intended. Controlling the answering process in CBT prevents the test-takers from engaging in an answering process that the evaluator does not intend (Masukawa et al., 2021), allowing for a more accurate diagnosis of their stumbling blocks. Moreover, due to analyzing the answers for each process, we found that only a few students answered incorrectly in one process but correctly in the next. There was only one student who could read the main points but not the outline, and four students who could not read the main points could summarize their opinions in Japanese sufficiently (one student scored 5 points and three students scored 3 points). This suggests that once they stumble in the middle of the thought process, it is difficult to answer correctly in the next stage.

The results of the cluster analysis using the ratio of RT for each screen to the total RT as a variable reveal five clusters. The clusters with low English writing scores were Clusters 3 and 4 (wherein students took too much time to form and organize opinions leaving little time to write in English). Students in Cluster 2 took too much time to understand the source text, which can show difficulties in English reading comprehension (e.g., Chan, 2017). Students in Clusters 3 and 4 spent more time forming and organizing their opinions on Screen 6 and had lower English writing scores, which can be interpreted as their inability to effectively conceptualize English writing in the prewriting stage (Xu & Ding, 2014). In this way, we can visualize the stages in which the students who could not write well in English stumbled.

There are several implications of CBT assessment for learning and teaching methods. In conventional PBT that only assesses language production, the main concern is whether a certain language performance task has been achieved; if not, the assessment is low. Tasks with a higher degree of difficulty are more likely to show a floor effect, as in this study; in fact, although eight students reached the stage of forming their opinions one step ahead, the overall task achievement was

Table 8 Means of each cluster

	S1 (%)	S2 (%)	S3 (%)	S4 (%)	S5 (%)	S6 (%)	S7 (%)	Total RT(s)	Score mean	Score distribution					
										0	1	2	3	4	5
Cluster 1 ($n=9$)	1.51	8.64	2.49	5.96	1.20	29.1	51.1	1159.8	0.9	2	6	1	0	0	0
Cluster 2 ($n=10$)	1.53	10.81	3.67	6.87	1.94	37.43	37.75	1056.8	1.1	4	4	0	1	1	0
Cluster 3 ($n=2$)	2.40	10.60	3.25	13.60	2.55	67.50	0.00	1026.4	0.0	2	0	0	0	0	0
Cluster 4 ($n=2$)	0.60	4.95	1.35	6.55	0.90	59.80	25.90	1202.7	0.5	1	1	0	0	0	0
Cluster 5 ($n=9$)	0.72	7.72	1.82	4.80	2.00	17.11	65.86	1329.7	1.7	2	3	2	1	0	1
Mean	1.29	8.95	2.65	6.43	1.72	32.65	46.31	1169.7	1.09						

S represents Screen. *Total RT(s)* represent the mean of total RT (in seconds) for each cluster. *Score mean* and *Score distribution* represent the mean score and score distribution of English writing for each cluster, respectively

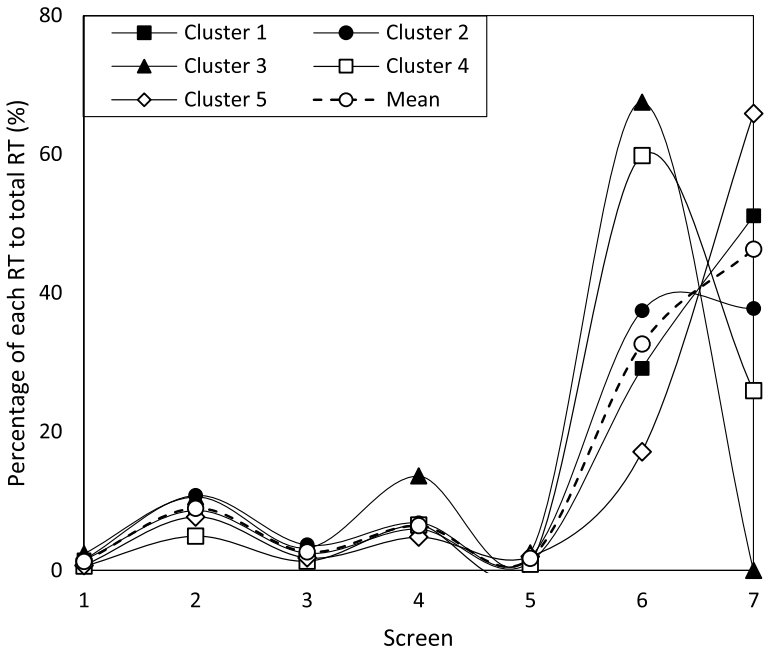

Fig. 14 Mean plots of each cluster

rated low. Therefore, it would be difficult to utilize the results in learning instruction because the points that needed to be taught were unclear. In a process-based evaluation method such as the prototype CBT, however, the main focus is to determine to what extent a student can solve a certain language performance task. Even if the task is not achieved in the end, the students can be actively evaluated according to different stages of achievement. Therefore, specific points that need to be taught will be identified and can be utilized in learning instruction (The Japan Association for Research on Testing, 2007). Specifically, CBT can be used to identify students who are not able to write English essays well but can complete the process of producing English writing. Those struggling with forming and organizing opinions can be provided instruction on generating and arranging ideas. Other students who are struggling only with logical expression in English can be provided explicit instruction on useful expressions. One last thing to take into consideration is that evaluation should be used for a specific purpose, and the actual evaluation method should be considered after the purpose has been defined. We can confidently say that the prototype CBT we created in this study has functions that can directly inform instruction.

In converting PBT, which has been utilized as the primary means of evaluation in Japanese school education, to CBT in the future, it is necessary to carefully consider the purpose of the CBT conversion. This case study purported to analyze the process leading to the product of each student and link it directly to the improvement of instruction instead of conventional PBT. In other words, the value of converting to CBT was not viewed simply in terms of efficiency, but in terms of its use in improving teachers' instruction and promoting students' learning. If CBTs are to be used

in this way, not only educators but the Japanese society as a whole will have to drastically change its perception of testing, and the shift to CBTs is an opportunity to modify this perception of testing in Japan. Until now, testing has been strongly associated with judging the ability of the examinees based on products and has been perceived simply as a means of grading them. However, depending on the test design, CBT can more powerfully link curriculum goals, teaching, learning, and assessment, allowing examinees to understand how far they have achieved their goals, where they need to actively improve, and then to envision the concrete actions needed next. To realize these goals, it will be necessary for Japanese society as a whole to take an interest in CBT and understand how CBT questions are designed, implemented, and validated. However, the aspect of the test that discriminates between high-ability and low-ability students will not be lost, so educators will need to build consensus on the purpose of CBT use.

Conclusions

This study proposed a CBT that overcomes the difficulties of PBT; it attempted to understand students' difficulties related to their thought processes in a multilayered manner. The results suggested that CBT can detect students who are able to complete the language performance task by controlling the order of answers, asking questions sequentially, and obtaining RT effectively. Therefore, CBTs could allow for the connection to instruction based on that student's achievement in the thought process.

As indicated by the demographic information of the participants, the students' typing speed was quite slow, which may have affected their English writing scores. The net speed of the students with English writing scores of 4 and 5 was 11 WPM and 15 WPM respectively, which was slightly higher than the average of 11.57 WPM. In contrast, the two students with a net speed of 47 WPM, which is much higher than the average, had English writing scores of 0 and 1. Except for these two students, the net speed (excluding missing values) of the 18 students who scored 0 or 1 on their English writing score was low ($M = 5.61$, $SD = 3.5$). Barkaoui (2014) suggests that poor keyboarding skills may have a small but negative effect on test-takers' language performance, which is why it is necessary to examine whether the same can be said in the context of English education in Japan. Although the time for answering the test was set at about 25 min, it is possible that the students with low English writing scores did not have sufficient time to answer the questions.

In this study, it was not possible to determine whether CBT with a controlled answer order captured students' thought processes and English achievements more accurately than PBT because we did not compare the results with PBT utilizing the same questions. Moreover, due to the small sample size, it was not possible to verify the validity of the CBT, which is another avenue for future research. Finally, due to the educational considerations of the participants, we were only able to propose one research question, which was very limited in scope. Since we only presented one assessment case, in the future, we would like to investigate whether a similar questioning and assessment approach would be effective for other skills.

Appendix

Table 9

Table 9 Reading-into-Writing Product Rubric

Score	Task Description	
	Screen 6: Japanese writing	Screen 7: English writing
5	A response at level 5 ● is effective in selecting major information from two source texts to support one another and connecting relevant ideas	A response level 5 ● is effective in selecting major information from two source texts to support one another and connecting relevant ideas ● demonstrates unity, coherence, syntactic variety, and appropriate word choice ● contains minor lexical or syntactical errors that do not interfere with meaning
4	A response at level 4 ● is effective in selecting major information from two source texts although some ideas may not be fully elaborated	A response level 4 ● is effective in selecting and connecting major information from two source texts although some ideas may not be fully elaborated ● demonstrates unity, coherence, syntactic variety, and appropriate word choice although it may contain few unclear connections or occasional redundancy ● contains few lexical or syntactical errors that do not interfere with meaning
3	A response at level 3 ● contains some but not all major points from two source texts and the points are imprecisely or incorrectly presented or connected	A response level 3 ● contains some but not all major points from two source texts and the points are imprecisely or incorrectly presented or connected ● demonstrates unity and coherence although it may contain few obscure connections and imprecise word choice ● displays limited syntactic structures and vocabulary ● contains some lexical or syntactical errors that occasionally obscure meaning
2	A response at level 2 ● contains limited relevant points from two source texts and they are significantly misrepresented ● displays little organization or inadequate connections of ideas	A response at level 2 ● contains limited relevant points from two source texts and they are significantly misrepresented ● displays little organization or inadequate connections of ideas ● contains inappropriate word choice ● displays many lexical or syntactical errors that largely obscure meaning
1	A response at level 1 ● contains little or no relevant information from two source texts is disorganized and underdeveloped ● is disorganized and underdeveloped	A response at level 1 ● contains little or no relevant information from two source texts ● is disorganized and underdeveloped ● displays serious and frequent lexical and syntactical errors that make understanding of the writing unlikely
0	A response at level 0 ● contains copied words from the source passages ● is left blank	A response at level 0 ● contains copied words from the source passages ● is written in a foreign language ● is left blank

Abbreviations

ET Execution time
MT Mean execution time
PT Preparation time
WPM Words typed per minute
CBT Computer-based testing
MEXT The Ministry of Education, Culture, Sports, Science, and Technology

NAEP	The National Assessment of Educational Progress
PBT	Paper-based testing
PISA	The Program for International Student Assessment
RT	Response time

Acknowledgements
The authors would like to appreciate the students and the teachers for their cooperation and participation. This work was supported by JST SPRING, Grant Number JPMJSP2132.

Authors' contributions
All authors read and approved the final manuscript.

Authors' information
Noriyasu Niimi is a Ph.D. student majoring in English Language Education at Hiroshima University, Japan. He is also a selected Next-Generation Fellow of the Program for Developing and Supporting the Next-Generation of Innovative Researchers at Hiroshima University from 2022 to 2025. His interests include English language teaching and assessment. Nobukazu Matsuura is a professor in English Language Education at Hiroshima University, Japan. He has specialized in Curriculum and Instruction, and language assessment of English Language Education.

Funding
Not applicable.

Availability of data and materials
The datasets used during this study can be availed from the corresponding author on reasonable request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 11 May 2022 Accepted: 26 August 2022
Published online: 13 September 2022

References

Barkaoui, K. (2014). Examining the impact of L2 proficiency and keyboarding skills on scores on TOEFL-iBT writing tasks. *Language Testing, 31*(2), 241–259. https://doi.org/10.1177/0265532213509810

Bennett, R., Zhang, M., Sinharay, S., Guo, H., & Deane, P. (2022). Are there distinctive profiles in examinee essay-writing processes? *Educational Measurement: Issues and Practice, 41*(2), 55–69. https://doi.org/10.1111/emip.12469

Brunfaut, T., Harding, L., & Batty, A. (2018). Going online: the effect of mode of delivery on performances and perceptions on an English L2 writing test suite. *Assessing Writing, 36*, 3–18. https://doi.org/10.1016/j.asw.2018.02.003

Chan, S. (2017). Using keystroke logging to understand writers' processes on a reading-into-writing test. *Language Testing in Asia, 7*, 10. https://doi.org/10.1186/s40468-017-0040-5

Chan, S., Inoue, C., & Taylor, L. (2015). Developing rubrics to assess the reading-into-writing skills: A case study. *Assessing Writing, 26*, 20–37. https://doi.org/10.1016/j.asw.2015.07.004

Chan, S., Bax, S., & Weir, C. (2018). Researching the comparability of paper-based and computer-based delivery in a high-stakes writing test. *Assessing Writing, 36*, 32–48. https://doi.org/10.1016/j.asw.2018.03.008

Douglas, D., & Hegelheimer, V. (2007). Assessing language using computer technology. *Annual Review of Applied Linguistics, 27*, 115–132. https://doi.org/10.1017/S0267190508070062

Ercikan, K., & Pellegrino, J. (Eds.) (2018). *Validation of score meaning for the next generations of assessments*: *The use of response processes*. Routledge

Gong, T., Shuai, L., Arslan, B., & Jiang, Y. (2020). Process based analysis on scientific inquiry tasks using large-scale national assessment dataset. *Proceedings of the 13th international conference on Educational Data Mining (EDM 2020)*, 417–423. https://educationaldatamining.org/files/conferences/EDM2020/papers/paper_7.pdf

Jamieson, J. (2005). Trends in computer-based second language assessment. *Annual Review of Applied Linguistics, 25*, 228–242. https://doi.org/10.1017/S0267190505000127

Kitazawa, T., & Shirouzu, H. (2020). CBT niyoru tashisentakushikimondai no kaiketsupurosesu no kaimei: Daigakunyuushi *Senta shikenmondai* no kokugokisyutsumondai wo katsuyou shite [Clarifying solving processes of multiple-choice problems by CBT: Using past Japanese language tests of the National Center University Entrance Examination]. *University Entrance Examination Research Journal, 30*, 52–58.

Knoch, U., & Sitajalabhorn, W. (2013). A closer look at integrated writing tasks: towards a more focussed definition for assessment purposes. *Assessing Writing, 18*(4), 300–308. https://doi.org/10.1016/j.asw.2013.09.003

Lee, Y. H., Hao, J., Man, K., & Ou, L. (2019). How do test takers interact with simulation-based tasks? A response-time perspective. *Frontiers in Psychology, 10*, 906. https://doi.org/10.3389/fpsyg.2019.00906

Masukawa, H., Shirouzu, H., Saito, M., Iikubo, S., & Amano, T. (2021). Development of CBT reading items to elicit "aggressive reading": Using language test of the university of Tokyo entrance examination. *Japanese Journal for Research on Testing, 17*(1), 25–44. https://doi.org/10.24690/jart.17.1_25

Matsuura, N. (2021). In Y, Ushiro., & M, Kashiba. (Eds.). *Shin kyosyoku katei ensyu: Vol. 18. Chuto eigoka kyouiku* [Middle English Education](pp. 10–13). Kyodo Shuppan.

MEXT. (2017). *Chugakkou Gakusyuu Shidou Youryou (Heisei 29 nendo kokuji) kaisetsu gaikoguko hen [The Explanation of The Course of Study for Junior High School (public notification in 29th year of Heisei period) Foreign Language Version]*.

MEXT. (2019). *Heisei 31 nendo (Reiwa gan nendo) zenkoku gakuryoku gakusyu jyokyo chosa hokokusyo Chugakko eigo [In 31th year of Heisei period (In 1st year of Reiwa period) The nationwide academic achievement test report Junior high school English]*.
National Center for university entrance examinations (2021). *Daikibo Nyugakusya Senbatsu Niokeru CBT Katsuyou no Kanousei Nitsuite (Hokoku)* [The Possibility of Using CBT in Large-Scale Admissions (Report)] https://www.dnc.ac.jp/albums/abm.php?f=abm00040361.pdf&n=%E3%80%90%E5%90%88%E4%BD%93%E7%89%88%E3%80%91CBT%E5%A0%B1%E5%91%8A%E6%9B%B8.pdf
Nishigori, D., Yamaguchi, A., Matsutaka, K., Osada, S., Sakaguchi, K., Fukui, T., Takamori, Y., Sonoda, Y., & Kodama, H. (2017). Dejitarugijyutsu wo katsuyou shita taburetto nyushi no kaihatsu: Tamenteki, Sougoutekihyouka ni muketa gijyutsuteki kentou [Development of examination using tablet device: technical considerations for multi-source and comprehensive assessment]. *University Entrance Examination Research Journal, 27*, 63–69.
Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing, 13*(2), 111–129. https://doi.org/10.1016/j.asw.2008.07.001
Sahin, F., & Colvin, K. F. (2020). Enhancing response time thresholds with response behaviors for detecting disengaged examinees. *Large-scale Assessments in Education, 8*, 5. https://doi.org/10.1186/s40536-020-00082-1
Talebinamvar, M., & Zarrabi, F. (2022). Clustering students' writing behaviors using keystroke logging: a learning analytic approach in EFL writing. *Language Testing in Asia, 12*, 6. https://doi.org/10.1186/s40468-021-00150-5
The Japan Association for research on testing (2007). *Test Standard: Nihon no test no shourai ni mukete* [Test Standards: Toward the future of testing in Japan], Kaneko Shobo.
Ukon, S., Kobayashi, M., Nakamura, Y., Okamoto, E., Yuji, S., Terasaki, S., Yasuda, J., & Yasuno, F. (2019). Taburetto tanmatsu wo mochiita eizou ya doutekiobujekuto wo fukumu CBT butsuri mondai no kaihatsu [Developing of physics problems with movies and dynamic objects for CBT using tablet computer]. Proceedings of the 43rd annual meeting of Japan Society for Science Education, 195–198. https://doi.org/10.14935/jssep.43.0_195
Xu, C., & Ding, Y. (2014). An exploratory study of pauses in computer-assisted EFL writing. *Language Learning & Technology, 18*(3), 80–96. 10125/44385.
Yamashita, T. (2017). Computer-based-testing (CBT) mondai no shisaku [A prototype of Computer-Based-Testing (CBT)]. *Chemistry in Education, 65*(7), 334–337. https://doi.org/10.20665/kakyoshi.65.7_334
Yang, H. (2012). A comparative study of composing processes in reading- and graph-based writing tasks. *Language Testing in Asia, 2*(3), 33–52. https://doi.org/10.1186/2229-0443-2-3-33

Publisher's Note