Language Testing in Asia

RESEARCH Open Access

Development and validation of an English test measuring EFL learners' critical thinking skills

Akiyo Hirai[1*], Hideaki Oka[2], Takeshi Kato[3] and Hiroki Maeda[4]

*Correspondence:
hirai.akiyo.ft@u.tsukuba.ac.jp

[1] Humanities and Social Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8571, Japan
[2] The University of Nagano, Nagano, Japan
[3] Graduate School, University of Tsukuba, Tsukuba, Ibaraki, Japan
[4] Matsuyama University, Matsuyama, Japan

Abstract

Addressing the lack of English tests that measure critical thinking (CT) abilities of EFL learners, this study aims to develop an English Critical Thinking Test (ECTT) designed to measure the consistency, analysis, and inference CT skills. It is a follow-up to a pilot study to accumulate validity evidence for the use of the ECTT. A total of 262 Japanese EFL learners who were first-year college students took the ECTT, a Japanese Critical Thinking Test (JCTT), and an English proficiency test (EPT). The result of the internal structures of ECTT and JCTT revealed that both CT tests fit two-factor models better than three-factor models, with consistency as one factor and analysis and inference as the other factor. This implies that it is difficult to measure the CT skills' hierarchical nature. However, the model that included the three tests fit the data well and indicated moderately high correlations between the ECTT and EPT ($r = .64$) and between the ECTT and JCTT ($r = .55$). In addition, we confirmed that the skills that test-takers needed most in solving questions in each of the three CT sections (i.e., consistency, analysis, and inference CT skills) of the ECTT were as intended. Based on these findings, the validation of the ECTT has been improved; ECTT can be used for EFL learners at upper secondary and university levels in Japan.

Keywords: Critical thinking test, Validation, Analysis, Consistency, Inference, Reading literacy

Introduction

Critical thinking (CT) has become increasingly important for making reasoned judgments in today's advanced information society, where all types of information, regardless of credibility, are easily accessible. The Assessment and Teaching of the Twenty-First Century Skills Project organized by educational experts identified the skills necessary for young people in the twenty-first century. Among them, primal importance was placed on CT skills, communication, and teamwork skills (Griffin et al., 2012). Furthermore, the Organization for Economic Co-Operation and Development (2017) identified CT skills as the basis of higher-order literacy. Following this trend, many universities regarded CT as essential skills in higher education and have made efforts to develop students' CT skills (e.g., Association of American Colleges and Universities [AACU], 2018).

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

With the worldwide recognition of the importance of CT skills, its pedagogy has been emphasized in various subject classes by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) (2016) in Japan, and EFL education is no exception. In global society, EFL learners will have opportunities to read English documents and discuss or negotiate in English with people from across the world, activating CT. Under such circumstances, the number of studies reporting on how to cultivate EFL learners' CT skills is increasing (e.g., Mineshima & Imai, 2017; Takeda, 2016). However, CT assessment for EFL learners is rarely discussed, largely owing to the lack of a widely accepted definition of CT (e.g., Nicholas & Labig Jr., 2013) and scarcity of assessment tools designed for EFL learners (Reid & Chin, 2021).

Thus, we launched a project to develop an assessment tool for EFL students at the upper secondary and tertiary levels in Japan. In the project, we created an English Critical Thinking Test (ECTT) and implemented it for EFL learners (Author, year). The result of the pilot study showed that three factors of "Consistency & Inference," "English ability & Consistency," and "Analysis & Calculation" were extracted and the ECTT moderately correlated with the English proficiency test (EPT) ($r = .43$). The current study is a follow-up of this pilot study to find further validity evidence for ECTT use at a larger scale than the pilot study. By conducting the present study, English teachers in EFL circumstances would be able to measure the degree of students' CT skills and review the effectiveness of their instruction for CT skills.

Literature review

This section provides relevant studies for the current study. First, cognitive skill dimensions of CT are explained. Second, existing CT tests are introduced, but we suggest the necessity of CT tests for EFL learners. Third, the result of ECTT in the pilot study is briefly reported. Finally, the similarities between CT process and reading literacy are explained to show how ECTT is related to reading literacy.

Cognitive skill dimensions of critical thinking

CT is a comprehensive cognitive ability that involves various CT skills and interrelates them. Halpern (1999) described CT as abilities related to problem solving, and making inferences and decisions. Although the definition of CT differs slightly depending on the researchers (Ennis, 1987; Facione, 1990; Kuhn, 2015; Kusumi, 2010, 2018), most researchers agree that CT is reflective, and reasonable thinking focuses on deciding what to believe or do (Ennis, 2018).

More specifically, CT is a skill-application cognitive process of c*larification*, *bases for a decision*, and *inference* (Ennis, 1987; Facione, 1990; Kusumi, 2010, 2018). *Clarification* ability is applied when critical thinkers try to understand the intention, warrant, or logic of information. *Bases for a decision* ability are activated when thinkers clarify presuppositions behind information and evaluate its reliability and the conclusion drawn by scientific methods. The final *inference* ability is applied when they attempt to draw the correct conclusion based on inductive and deductive reasoning. Each of these CT abilities involves multiple CT skills or aspects, and critical thinkers use these CT skills selectively to obtain the most appropriate solution based on the task at hand. In other words, CT skills greatly help people solve problems or lead to a rational conclusion. Thus, due to the complex process of CT, to define its domain constructs is difficult.

Critical thinking tests

Difficulties in defining constructs of CT make the assessment even more challenging (Ennis, 2003). Another difficulty in assessing CT is domain specificity. CT is generally perceived as a generic ability that can be applied to a wider range of contexts (e.g., Ennis, 2018; Kusumi, 2018). However, some argue that the domain-specific nature of CT skills is unavoidable, rendering little value for the results of a general assessment (e.g., Rear, 2019). A case in point is that even though a Japanese student who has high CT skills but low English ability may not perform CT skills well in the domain of an English language environment owing to language barriers. In other words, performing CT skills is influenced by some degree of domain-specific knowledge and skills. Despite these arguments, several standardized English CT tests are available in the market, such as the Cornell Critical Thinking Test (Cornell test), Watson-Glaser test, and California Critical Thinking Skills Test (California test).

The Watson-Glaser test was first developed in the 1930s and later revised (Miller, 1992). The test is commonly used in “graduate, professional, and managerial recruitment” (Assessment Day, 2022) and has five subtests: *assumptions*, *analyzing arguments*, *deductions*, *inferences*, and *interpreting information.* Miller (1992) investigated the reliability and validity of the Watson-Glaser test and reported that the reliability calculated by split-half coefficients of the whole test ranged from .85 to .87, but those of the five subsets (e.g., assumptions and inferences) were lower, ranging from .41 to .74. Therefore, these subtests are recommended to be used all together, not separately. The Cornell test was first developed by Ennis and Millman (1985) to help teachers determine their students’ CT ability and predict their future performance and is now available at two levels: level X for Grades 5 to 12 and level Z for Grades 11 to 12. Level Z measures seven skills: *induction*, *deduction*, *credibility*, *identification of assumptions*, *semantics*, *definition*, *and prediction in planning* (Tests.com LLC, 2020). The split-half reliabilities of the test ranged from .55 to .76, and the correlation with the Watson-Glaser test was .48 for undergraduates (Rane-Szostak & Robertson, 1996). Like the Watson-Glaser test, the subscale scores of the Cornell test are also not recommended to be used separately because of the lack of evidence that they measure distinct skills from factor analysis studies (Reid, 2000).

More recently, the California test was developed based on the Delphi Expert Consensus Definition of Critical Thinking (Facione, 1990), in which core CT skills (i.e., interpretation, analysis, and evaluation) and sub-skills (e.g., categorization, examining ideas, and assessing claims) were defined to integrate CT into school curricula and assess students’ CT skills. The California test measures the following skills: *overall reasoning skills*, *analysis*, *interpretation*, *evaluation*, *explanation*, *inference*, *deduction*, *induction*, and *numeracy* (Insight Assessment, 2022). However, concerns have also been raised because of the relatively low pre-post correlation of .70 (Adams et al., 1996), and the test did not seem to be able to determine which item measures which subskill because of the multiple skills necessary to answer an item (Reid, 2000).

Among other validation studies on the California test, Facione (2000) confirmed a relatively weak relationship with the Watson-Glaser test ($r = .41$). This might be because CT is a complex ability and the two tests do not measure exactly the same CT skills (such as *explanation* and *induction* of the California test and *assumption* and *analyzing arguments* of the Watson-Glaser test). Further information is reported in the California test manual (Insight Assessment, 2022) that there was a relatively high correlation with the GRE total score ($r = .72$) but low correlation ($r = .10$ to $.20$) with grade point average (GPA). This may be because GPA is a holistic measure that usually involves various factors of student performance (Insight Assessment, 2022). Looking into existing tests, some skills are commonly tested, such as *analysis* (or analyzing arguments) and *inference* (including *induction* and *deduction*); other skills are uniquely measured in each test.

Besides these three CT tests, some psychometric tests include sections measuring CT skills. For example, Graduate Record Examinations (GRE), used for admission to many graduate schools, include verbal reasoning, quantitative reasoning, and analytical writing (ETS, 2020). Thinking Skills Assessment, an admission test required by major UK universities, measures CT and problem-solving abilities (Cambridge Assessment Admissions Testing, 2020).

While the studies above provided evidence regarding theses tests' validity and reliability, they seem insufficient based on the recent ideas of validation approaches (e.g., Chapelle & Voss, 2021). In other words, CT tests mentioned above have not clearly reported either the test scores' scope of interpretation and use or the distinctions and relationships among intended factors. Thus, it would be hard to conduct validation studies of ECTT using them. In addition, these CT tests are not suitable for classroom use for the following reasons. One is that English and some contents used in the tests would be too difficult for EFL learners because they are for native speakers or high-proficient speakers, and some of the contents do not culturally fit to Japanese EFL learners. Another point is that obtaining such overseas commercial tests is difficult (e.g., Stroupe, 2006) and buying them for all the students causes a great financial burden for teachers. To overcome these problems, it would be necessary to develop a CT test suitable for EFL learners.

CT skills of the ECTT in the pilot study

In the pilot study (Authors, year), we developed an English Critical Thinking Test (ECTT) to assess EFL learners' CT ability at the upper secondary or university level in Japan. Upper secondary students in Japan are expected to achieve the Common European Framework of Reference for Languages (CEFR) B1 level by Ministry of Education, Culture, Sports, Science and Technology (MEXT) (2016), and most first-year university students at the authors' university are in this range.

We administered the ECTT to 80 first-year university students in the pilot study, and they achieved 64% of the correct responses on the ECTT. Additionally, the students took TOEIC IP ($M = 488.38$, $SD = 7.93$) and the English proficiency test (EPT, see "Materials" section), whose mean correct response was 67%. Regarding CT skills to be measured in the ECTT, we needed to determine which CT skills were integrated into the test. The ECTT is designed for EFL learners to demonstrate the CT skills required to solve problems by forming reasoned judgments. For that purpose, we chose three aspects of

CT: *consistency*, *analysis*, and *inference* skills, referring to previous literature and existing tests, and CT skills that can be measured in an English reading test (Table 1).

First, *consistency* skills are mainly used in the *clarification* stage mentioned above (Ennis, 1987, 2018; Kusumi, 2010, 2018) and involve clarifying the meaning of texts and their logical structure, such as cohesion and coherence. We believe that it is an important and basic CT skill for EFL learners to read texts critically. According to Clemson University (2020), which has tried to enhance students through CT, one of the mainstream important concepts of CT is "Recognize flaws and inconsistencies in an argument" (quoted by Ennis, 2018; p. 182).

The second CT skill we chose is *analysis* or analytical skill, which involves analyzing the text and gathering the most relevant necessary information to develop a reasonable interpretation (Insight Assessment, 2022). The information extracted includes words or phrases and numbers in a chart, which is similar to the *numeracy* and *analysis* in the California test and *analysis* in the Watson-Glaser test. The analysis also covers skills to evaluate whether pieces of information in a text are correct, based on the subskill of "Judge the credibility of a source" in *bases for a decision* phase (Ennis, 2018; p. 167).

The third CT skill we chose is *inference*, which involves deductive and inductive reasoning to make a reasonable decision, corresponding to *inference*, *deduction*, and *induction* in the California test and in the *inference* stage of the CT process. These CT skills are assumed to be domain universal skills that may be utilized in other subject areas (e.g., Ennis, 1987, 2018). Based on the domain definition of the three CT skills followed by the item specification we made (Table 1), ECTT items were created and tested in the pilot study (see the "Materials" section).

Relationship between critical thinking skills and reading literacy

ECTT is closely related to reading literacy because CT skills are used when reading texts critically (e.g., Aloqaili, 2012; Lee, 2015). The relationship between critical reading and CT can be explained in the Programme for International Student Assessment (PISA) reading literacy assessment framework (Fig. 1), which illustrates the process of reading or critical reading. The PISA is an international test that measures 15-year-old students' math, science, and reading literacy skills to evaluate education systems worldwide. The items in the reading literacy assessment were categorized into three types. As shown in Fig. 1, the types of "Access and retrieve" and "Interpret and integrate" items are associated with the readers' literal and interpretative comprehension

Table 1 Item specification of the ECTT

CT skills	Item specification
(a) Consistency	1. Rearrange the sentences to make the text more logical and coherent. 2. Fill in the blank to make more cohesive relations between the sentences. 3. Exclude a sentence to make the paragraph more coherent.
(b) Analysis	1. Interpret graphs and charts. 2. Calculate using numbers in the text to find the answer. 3. Extract necessary information and judge its credibility.
(c) Inference	1. Deduce objective conclusion based on reliable information. 2. Infer the meaning of an unknown word from a context. 3. Choose the most reasonable argument from a context.

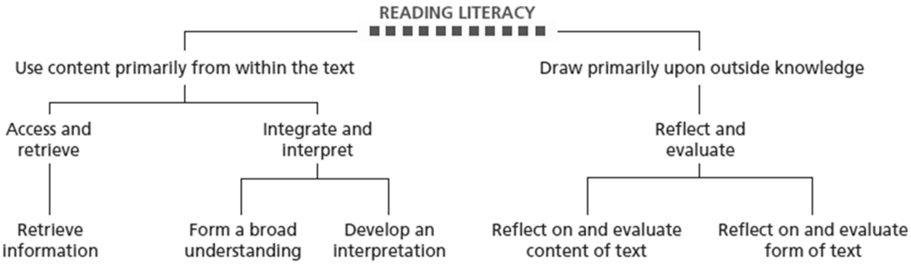

READING LITERACY
Use content primarily from within the text
Draw primarily upon outside knowledge
Access and retrieve
Integrate and interpret
Reflect and evaluate
Retrieve information
Form a broad understanding
Develop an interpretation
Reflect on and evaluate content of text
Reflect on and evaluate form of text

Fig. 1 Types of items in the PISA Reading Literacy Assessment Framework. *Note.* Adopted from Organization for Economic Co-Operation and Development (OECD) (2017)

within a text. Some "Interpret and integrate" items may require lower-level CT skills such as interpretation, synthesis, and logical thinking skills. The third type, "Reflect and evaluate," requires readers' outside knowledge and higher-level CT skills such as evaluation and inference (Khamkhong, 2018). From a cognitive processing perspective, readers must first comprehend the text to interpret it, and only those who can interpret the text can evaluate it (Organization for Economic Co-Operation and Development (OECD), 2017). While reading the text, if necessary, readers may go back and forth between the comprehension and evaluation. In this regard, these three types are not hierarchical, but semi-hierarchical. This semi-hierarchical process of critical reading may prompt readers to use CT skills as necessary, such as *clarification*, *basis for a decision*, and *inference* stages (Ennis, 1987; Kusumi, 2010).

Regarding empirical studies on the relationship between CT and reading literacy, Facione (1990) reported that CT skills were moderately correlated with the breadth of vocabulary knowledge and reading comprehension ($r = .46$ and $r = .43$, respectively). However, the relationship between CT skills and second language (L2) reading literacy is not the same. Critical reading in L2 inevitably becomes much more complex than that in L1 because readers require both L2 proficiency and CT skills (Khamkhong, 2018).

The influence of L2 proficiency on CT use becomes evident when the CT use is compared between L1 and L2 (e.g., Floyd, 2011; Grosser & Nel, 2013). Floyd (2011) compared Chinese students' Watson-Glaser Critical Thinking Appraisal (Watson-Glaser test) in Chinese (L1) and English (L2). The students' IELTS scores were between 6 and 6.5. The results showed that the students performed significantly better in the L1 version of the Watson-Glaser test than the L2 version, suggesting that they cannot exhibit their CT skills in L2 reading as fully as they can in L1. Floyd attributes this result to their slower L2 word recognition speed and overload of their L2 working memory capacity. Zhou et al. (2015) also investigated CT ability in English reading skills of 224 university students in non-English majors in China, using questionnaires and interviews. The results show that most students lack CT ability in English reading, from which they claim to urgently need relevant training and to raise students' awareness of CT in EFL reading classes. They extended their opinions on the influence of the current English examinations in China. Since students read English texts to cope with various English tests, much attention should be paid to the improvement of students' CT ability through English examination.

With regard to the L2 production skill, the difference in CT use between L1 and L2 becomes much more obvious. Manalo and Sheppard (2016) investigated the difference between Japanese students' CT use in their L1 and L2 compositions. The result was that the Japanese students wrote significantly more sentences in English than in Japanese, but they produced more of the target evaluative language in Japanese. In addition, the students' TOEIC scores correlated significantly with the number of evaluative sentences in their English compositions. Manalo and Sheppard (2016) assumed that cognitive resources available in working memory for critical thinking depletes to a greater extent by lower proficiency learners since they need more resources on writing English composition itself than proficient learners whose English production skills have become more automatized. Thus, these studies suggest that limited L2 ability may slow down higher-order operations (Just & Carpenter, 1992).

Current study

Most English reading tests for EFL learners measure literal comprehension including vocabulary and grammar knowledge, focusing only on domain-specific knowledge and skills. However, in the real world, we go beyond text comprehension when reading, by checking the credibility of the source to write an academic paper or a business report, or by extracting necessary facts or figures to lead to a reasonable judgment. To prepare for this, students must be trained to use CT skills when reading L2 texts (e.g., Marin et al., 2017). In this respect, developing a CT test for EFL learners is significant for assessing these skills, as this type of test is currently scarce.

To this end, first in the pilot study, we conducted a validation study for the ECTT and extracted three factors through exploratory factor analysis (EFA). However, the participants were relatively homogeneous in their English levels as indicated with a small standard deviation of the TOEIC IP, and their number was small ($N = 81$) for investigating the factor structure using structural equation modeling. Compensating for these points, the current study attempts to obtain further validity evidence for the use of the ECTT to assess the intended L2 CT skills of more heterogeneous EFL learners. In addition, since the ECTT is an English reading test, it needs to be clear whether it is a CT test or an English test. Thus, the following RQs are addressed:

RQ1. Is the difficulty of ECTT items appropriate for the target EFL learners?
RQ2. What is the internal structure of the ECTT?
RQ3. How are ECTT scores correlated with Japanese Critical Thinking Test (JCTT) and English proficiency test (EPT)?
RQ4. Do participants answer ECTT items based on the skills we intended?

Method

Participants

A total of 262 Japanese EFL learners from three universities participated. Most of the students were freshmen but some were sophomores, and their majors were diverse, including engineering, economics, and literature. They took the same EPT to compare with students in the pilot study. The EPT comprises 24 items (10 vocabulary, 6 grammar, and 8 reading comprehension items), which were taken from a Test of English for International Communication (TOEIC) practice test, and 2nd and pre-1st grade EIKEN prep-tests, the item levels of which were at CEFR B1 (EIKEN Foundation of Japan, 2022). EIKEN tests are one of the most popular and widely used commercial English proficiency tests in Japan. Based on the average and standard deviation of EPT score ($M = 12.13$ (51%), $SD = 4.90$), the range of the learners' English levels were found to be wider than that of the participants in the pilot study ($M = 16.20$ [67%], $SD = 3.50$). This indicates some improvement in sample heterogeneity, which was a concern in the pilot study.

Materials

English Critical Thinking Test

The ECTT was originally constructed based on the authors' specifications (year) (Table 1). Items with poor discrimination at less than $r_{\mathrm{pb}} = .30$ were eliminated based on the suggestion by Fulcher and Davidson (2007) that items are acceptable if they have an r_{pb} of around .25 or greater. Consequently, 30 items (14 for consistency, 8 for analysis, and 8 for inference skills) were examined here (see Additional files 1, 2, and 3).

Japanese Critical Thinking Test

To examine the participants' L1 CT ability and specifically compare it to their L2 CT ability, a Japanese CT test (JCTT) was developed based on the specification used for the ECTT. The JCTT comprises 20 items (6 for consistency, 7 for analysis, and 7 for inference), which were mainly taken from or referred to past Synthetic Personality Inventory tests, an aptitude examination that many university students in Japan must take during the recruitment process (Recruit, 2022).

Questionnaire

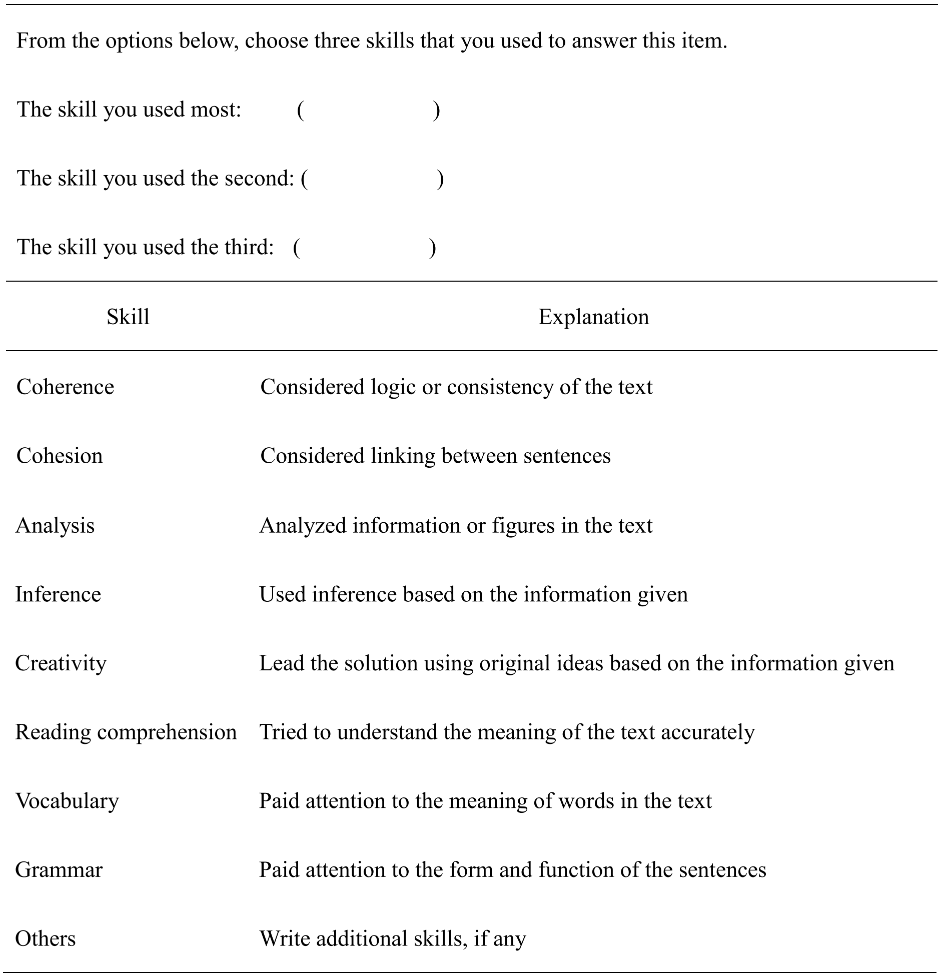
To make it clear what skills each CT item measures (RQ4), a questionnaire (Table 2) was inserted under each ECTT item. Every time participants answered an ECTT item, they chose skills they thought they used out of eight options (from the most necessary skills to less, up to three skills).

The skills included in the questionnaire (Table 2) were, based on the theory (i.e., the CT process and the reading literacy framework in OECD), ECTT target constructs (i.e., coherence, cohesion, analysis, and inference) and language knowledge and skills (i.e., reading comprehension, vocabulary, and grammar). The explanation shown in the Table 2 describes each construct's main characteristics. One skill, creativity, was added to the option list, although the ECTT was not intended to measure it. The reason for inclusion is that it is relevant to inference skills (e.g., Anderson et al., 2001), and we wanted to judge how students perceived the inference items. Lastly, there was an option to choose others, in which the participants could write skills other than the abovementioned choices.

Procedure

Participants took EPT, ECTT, and JCTT in English courses to enhance their CT skills. While taking ECTT and JCTT, the students could take notes to think or calculate. In

Table 2 A questionnaire inserted after every ECTT item



From the options below, choose three skills that you used to answer this item.

The skill you used most: ( )

The skill you used the second: ( )

The skill you used the third: ( )

Skill	Explanation
Coherence	Considered logic or consistency of the text
Cohesion	Considered linking between sentences
Analysis	Analyzed information or figures in the text
Inference	Used inference based on the information given
Creativity	Lead the solution using original ideas based on the information given
Reading comprehension	Tried to understand the meaning of the text accurately
Vocabulary	Paid attention to the meaning of words in the text
Grammar	Paid attention to the form and function of the sentences
Others	Write additional skills, if any

the pilot study, it took them more time to solve each ECTT item than a regular English reading item, such as an EPT reading passage. Thus, in the present study, we divided the ECTT items into four sets and JCTT items into two sets and administered them separately once a week for 6 weeks to avoid test-takers' mental fatigue. All the tests were provided on learning management systems, Manaba or Moodle, for automatic calculation.

Results

Item analysis

To examine whether the items functioned properly for RQ1, item analysis was conducted using a correct response (CR) rate and a point-biserial correlation coefficient (r_{pb}). The CR rate is the percentage of CRs of an item and indicates its difficulty. The one between 30 and 70% is acceptable as a test (Brown, 2013). Since the number of items in each section differs, the mean CR rate can be used to compare the difficulties of the three sections. The results of the CR rates for the consistency, analysis, and inference sections were 57, 52, and 47%, respectively (Table 3), which were all within the acceptable range.

Table 3 Percentages of correct responses and point-biserial correlation coefficients of the three skills' items

Consistency	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	*M*
% of CR	75	67	66	33	44	62	43	74	56	67	51	50	73	33	**57**
r_{pb}	.42	.60	.46	.43	.45	.45	.52	.45	.43	.48	.49	.43	.38	.35	
Analysis	A1	A2	A3	A4	A5	A6	A7	A8	*M*						
% of CR	57	64	72	68	40	25	45	46	**52**						
r_{pb}	.45	.48	.56	.62	.30	.48	.51	.44							
Inference	I1	I2	I3	I4	I5	I6	I7	I8	*M*						
% of CR	23	35	32	29	81	76	69	32	**47**						
r_{pb}	.31	.55	.49	.41	.35	.45	.43	.45							

CR correct response, r_{pb} point-biserial correlation coefficient

Table 4 Descriptive statistics of ECTT, JCTT, and EPT

ECTT (*k*)	Consistency (14)		Analysis (8)		Inference (8)		Total (30)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Score	7.95	3.02	4.17	1.82	3.75	1.54	15.87	5.13
% of CR	57	21.55	52	22.8	47	19.73	**53**	17.11
JCTT (*k*)	Consistency (6)		Analysis (7)		Inference (7)		Total (20)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Score	4.08	1.29	4.62	1.66	3.73	1.72	12.43	3.67
% of CR	68	21.52	66	23.74	53	24.53	**62**	0.35
EPT (*k*)	Vocabulary (10)		Grammar (6)		Reading (8)		Total (24)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Score	5.15	2.19	2.16	1.64	4.82	1.97	12.13	4.9
% of CR	52	21.89	36	27.32	60	24.65	**51**	20.41

k number of items, *CR* correct response, r_{pb} point-biserial correlation coefficient

The point-biserial correlation coefficient was used to examine the item discrimination, which is the correlation coefficient between the individual items and the overall score. As mentioned in the "Materials" section, all the items used for further analysis meet the criterion of $r_{\mathrm{pb}} = .30$ or more.

The Cronbach's alpha of the whole test (i.e., 30 items) was .79, which was sufficiently high for internal consistency reliability. When observing each CT section, Cronbach's alpha for 14 consistency items was the highest among the three sections ($\alpha = .70$), and the mean difficulty of the sections was 57%, with items ranging from 33 to 75%. The analysis section had the reliability of $\alpha = .52$, and the mean CR rate of 52%, ranging from 25 to 72%. However, the reliability for the eight inference items was the lowest ($\alpha = .38$). The CR rate of this section was also the lowest at 47%, but the items widely ranged from 23 to 81%.

Table 4 presents the descriptive statistics of the three tests. The easiest test for the students was JCTT (62%), followed by ECTT (53%), and EPT (51%). A repeated-measures one-way analysis of variance (ANOVA) was conducted to compare the difficulty of the three tests. The results showed that the ECTT (53%) was significantly more difficult than the JCTT (62%, $p < .001$), but was not statistically different from EPT (51%, $p = .09$).

We further examined whether the three sections of the ECTT and JCTT were significantly different in difficulty, with a two-way repeated ANOVA: one independent variable was language (English and Japanese) and the other was CT skill (consistency, analysis, and inference). As a result, there was no significant interaction, but significant main effects of language existed (F (1, 202) = 61.22, $p < .001$, $\eta_p^2 = .23$) and CT skill (F (2, 404) = 62.93, $p < .001$, $\eta_p^2 = .24$). The post hoc test (Table 5) showed that, except for the differences between consistency and analysis in both ECTT (57% and 52%) and JCTT (68% and 66%), all the differences across the sections were significant at $p < .05$. Thus, inference turned out to be significantly more difficult than the other two skills in both tests.

Factor analysis

Next, to examine RQ2, we conducted exploratory factor analysis (EFA) to find the internal factor structure of the ECTT using R (R Core Team, 2019). We chose the diagonally weighted least squares (DWLS; Jöreskog & Sörbom, 1996) as an estimation method because it can treat the observed dichotomous data and is robust for multivariate normality (Forero et al., 2009; Mîndrilă, 2010). As a factor rotation method, we used the oblique geomin rotation because it has recently been recommended for complicated factor solutions (e.g., Hattori et al., 2017).

To determine the number of factors in the ECTT, we conducted parallel analysis based on the squared multiple correlations (SMC method) and Velicer's minimum average partial (MAP) test. Based on the Hori's (2005) recommendation, we performed a series of EFAs, decreasing the number of factors extracted from six factors (suggested by the SMC method) to one factor (suggested by the MAP test). Items were removed from the model if they did not have primary factor loadings of .30 or higher, or if the items loaded on more than one factor; based on this criteria, 11 items (i.e., A1, A2, A5, I1, I4, I5, I7, C5, C6, C13, and C14) were excluded. As a result, the two factors solution was the most interpretable (Table 6). The first factor consists of all analysis and inference items, except for one consistency item with the lowest loading. Thus, the first factor was named Analysis and Inference (A and I). The second factor consists of only consistency items; thus, it was named the Consistency factor. The cumulative proportion of variance explained was 32%, and the factor correlation was .45.

Table 5 Post hoc comparisons between three CT skills in each ECTT and JCTT

		Mean		**95% CI**				
Comparison		difference	*SE*	Lower	Upper	*t*	*p*	*d*
E.Cns	E.Anl	4.70	1.67	−0.23	9.63	0.058	.077	.21
	E.Inf	9.87	1.67	4.94	14.80	5.90	< .001***	.48
E.Anl	E.Inf	5.17	1.67	0.24	10.10	3.09	.031*	.28
J.Cns	J.Anl	1.97	1.67	−2.96	6.90	1.18	1.000	.09
	J.Inf	14.64	1.67	9.71	19.57	8.74	< .001***	.63
J.Anl	J.Inf	12.67	1.67	7.74	17.60	7.57	< .001***	.53

* $p < .05$; $^{***}p < .001$. E = English; J = Japanese; *d* = Cohen's *d*; *p*-values and confidence intervals were adjusted using the Bonferroni method

Table 6 Exploratory factor analysis of ECTT

Item	Description	Factor 1	Factor 2	Commu
		(α=.69)	(α= .66)	nality
A4	Extract information for shopping of school festival	0.75	-0.01	0.56
A3	Extract information for shopping of school festival	0.71	-0.02	0.49
I2	Choose reasonable argument about the number of fruits	0.66	0.14	0.54
A6	Interpret the table of yearly sale with some calculation	0.59	-0.28	0.27
I3	Choose reasonable argument about the number of fruits	0.54	0.02	0.30
A7	Interpret the table of the test scores	0.48	0.16	0.33
I6	Infer the meaning of unknown words	0.47	0.11	0.28
A8	Extract necessary information and calculate bicycle speed	0.46	0.05	0.24
I8	Deduce conclusion based on several propositions	0.44	-0.18	0.16
C7	Rearrange sentences to make the visitor's story logical and coherent	0.32	0.23	0.22
C10	Fill in the conj. to make cohesive sentences about calling ambulance	-0.12	0.70	0.43
C11	Fill in the conj. to make cohesive sentences about calling ambulance	0.04	0.61	0.40
C3	Arrange sentences to make museum conversation logical and coherent	0.08	0.55	0.34
C8	Arrange sentences to make a logical and coherent essay about Eri	0.01	0.55	0.30
C2	Arrange sentences to make tele conversation logical and coherent	0.20	0.50	0.38
C12	Fill in the conj. to make cohesive sentences about calling ambulance	0.21	0.42	0.30
C1	Arrange sentences to make tele conversation logical and coherent	0.04	0.40	0.18
C4	Arrange sentences to make museum conversation logical and coherent	0.00	0.37	0.14
C9	Arrange sentences to make a logical and coherent essay about Eri	0.27	0.32	0.25

With factor correlations of	Factor 1	Factor 2
Factor 1	1	
Factor 2	0.448	1

SEM for ECTT

We also examined whether the factor structure of ECTT fit the three-factor or two-factor structure using structural equation modeling (SEM) with the lavaan package (Rosseel, 2012) in the R environment. The model fit was checked by a comparative fit index (CFI) of .90 or above and a root mean square error of approximation (RMSEA) of .08 or below (Browne & Cudeck, 1993). The sample size exceeded 200 (i.e., 203 observations in this study), which is considered “large” according to Kline’s (2005) guidelines.

We first tested a three-factor model in which the three CT skills correlate with each other, and each item reflects its measuring CT skill, using the DWLS estimation method for the same reason as the EFA mentioned above. The result of this initial model fit the data well with CFI = 1.00, RMSEA = 0.00 [0.00, 0.03], but revealed that the correlation between ECTT analysis (E_Anl) and inference (E_Inf) was too high (r = 1.07). This might be because the mathematical behavior of the two latent variables is quite similar, causing a problem with the model specification. Thus, we decided to merge them into one as E_A&I and tested a two-factor model whose factors correlated with each other.

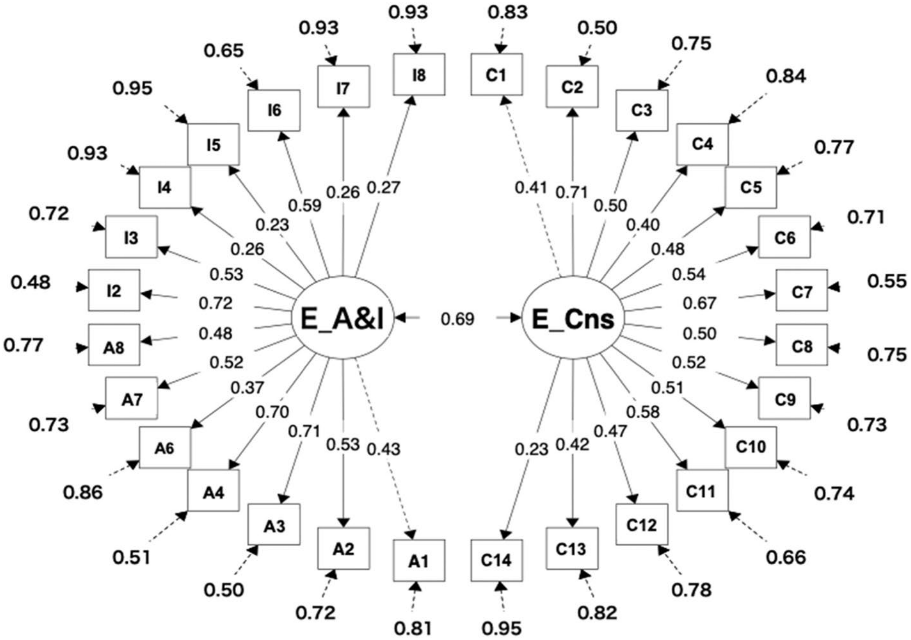


Fig. 2 Final two-factor correlated model of ECTT. *Notes.* All the estimates are standardized and significant ($p < .001$); E_A&I = ECTT Analysis and Inference, E_Cns = ECTT Consistency. χ^2 (349) = 346.88, p <.001, GFI = .86, AGFI = .84, SRMR = .10, CFI = 1.000, RMSEA= .00 [.00, .03]

Since the paths from "E_A&I" to item A5 (β = .08, p = .117) and I1 (β = .10, p = .085) were not significant, the two items were excluded from the model. As a result, the model shown in Fig. 2 fits the data reasonably well, where CFI = 1.00, RMSEA = 0.00 [0.00, 0.03].

SEM for JCTT

We also examined the factor structure of the JCTT using the same procedure as that of ECTT. As a result, similar to the case of the ECTT, the three-factor model fit the data well with CFI = 1.00, RMSEA = 0.00 [0.00, 0.03], but a high correlation (r = .93) between the JCTT analysis (J_Anl) and inference (J_Inf) was detected. Thus, we constructed a two-factor model, that is, Consistency and "Analysis and Inference," which showed a reasonable fit to the data, with CFI = 1.00 and RMSEA = 0.00 [0.00, 0.03] (Fig. 3). Thus, the internal factor structures of ECTT and JCTT were found to be quite similar even though the languages used in the tests were different.

SEM for all three tests

Next, to examine RQ3, we tested a model that included the total scores of each section of the ECTT, JCTT, and EPT and the corresponding latent variables named "L2 Critical Thinking skills (L2CT)," "L1 Critical Thinking skills (L1CT)," and "L2 Reading Literacy (L2RL)," assuming correlations among the three latent variables. In this analysis, to avoid an excessive increase in model complexity, the total scores of the constructs in each test were used as observed variables. As univariate and multivariate normality were

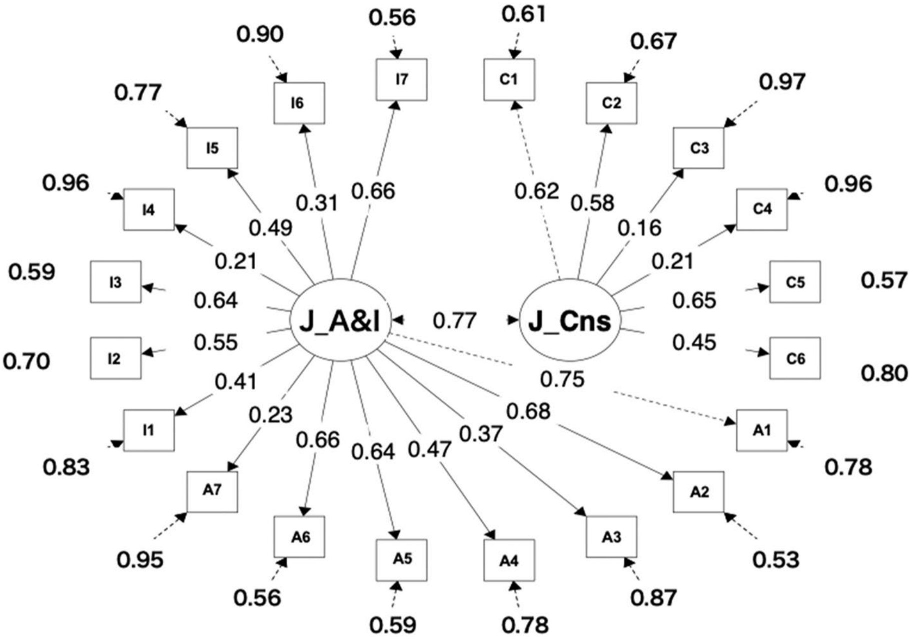


Fig. 3 Final two-factor correlated model for JCTT. *Notes.* All the estimates are standardized and significant (p <.001); J_A&I = JCTT Analysis and Inference; J_Cns = JCTT Consistency. χ^2 (169) = 167.64, p = .53, GFI = .90, AGFI = .88, SRMR = .11, CFI = 1.000, RMSEA = .00 [.00, 03]

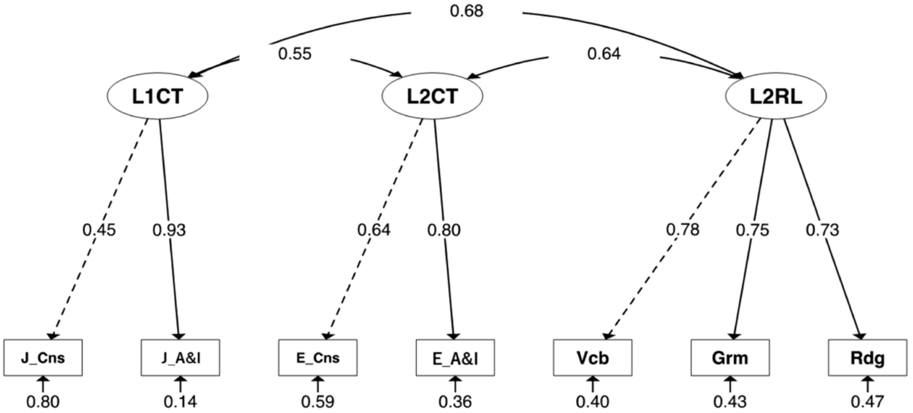


Fig. 4 Three-factor correlated model. *Notes.* All path coefficients are standardized and significant (p <.001); L2CT = L2 critical thinking; L1CT = L1 critical thinking; L2RL = L2 reading literacy. χ^2 (11) = 16.79, p = .114, GFI = .98, AGFI = .94, SRMR = .03, CFI = 0.86, RMSEA = .051 [.00, .10]

confirmed for the data, the maximum likelihood (ML) method was used for parameter estimation.

The results showed a good fit to the data, with CFI = .99 and RMSEA = .05 [.00, .10] (Fig. 4). The model indicates similar strengths of relationships with each latent variable, but L2CT was more closely related to L2RL (r = .64) than to L1CT (r = .55). In both ECTT and JCTT, the path coefficients from L1CT to J_A&I (β = .93) and from L2CT to E_A&I (β = .80) were larger than those from J_Cns (β = .45) or E_Cns (β = .64). This may indicate that analysis and inference influenced CT skills more than consistency.

Questionnaire survey

Lastly, we investigated whether the skills used by test-takers were aligned with the authors' expectations to answer RQ4. Since each item measures complicated CT skills and English abilities, we allowed them to choose three out of eight options or write freely. Participants ranked the three selected skills based on subjective importance they perceived when answering each item. For the sake of calculation, three points were given to the most needed skill, two points to the second-most needed, and one point to the third. The total scores for each option were then divided by the number of responses.

Figure 5 shows the proportion of skills that participants felt they used to answer the items. As expected, among the skills used in the consistency items, coherence and cohesion skills accounted for 53%, followed by reading comprehension (18%). This suggests that the participants needed to read connections and logicality between the sentences or paragraphs carefully when they dealt with consistency items. As for the analysis items, analysis skills were chosen as high as 37%, indicating that the participants analyzed texts, tables, or figures to retrieve necessary information. In addition, the percentages of inference (23%) and creativity (9%) were also chosen, which suggests that students made some inference or used imagination when they answered these items.

The results of the inference items, however, were somewhat different. In addition to inference skill (22%), other CT skills such as analysis (16%) and coherence and cohesion (23%) were also used equally. This indicates that inference items may require a series of multiple skills, such as sorting and evaluating information after understanding the logical structure of sentences. In addition, some students felt they used creativity (7%) though it was not an intended skill for the inference items. Lastly, regardless of any CT skills items, reading skills and some grammar and vocabulary knowledge were consistently used, which was a reasonable result as ECTT is an English reading test.

Discussion

To develop an English CT test for EFL learners, the current study examined the validity of the ECTT, with the following four RQs:

First, regarding RQ1 (Is the difficulty and reliability of ECTT items appropriate for target EFL learners?), the test was reasonably challenging for the target EFL learners, as

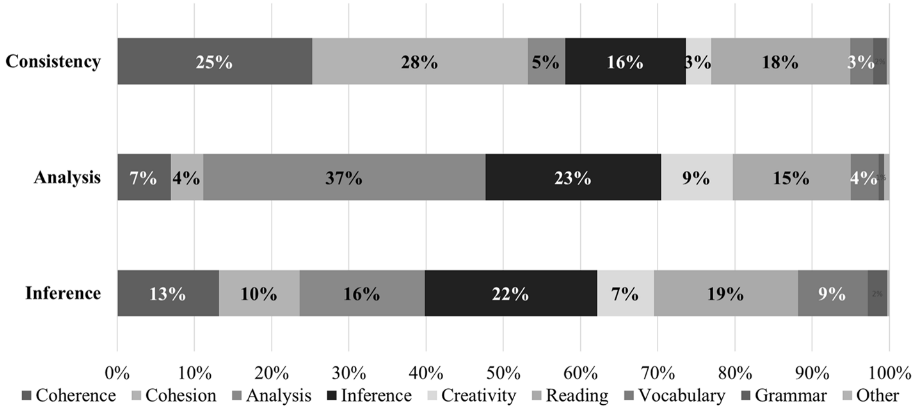


Fig. 5 Percentages of skills participants felt they needed to respond to the items in each section of ECTT

their total correct response was 53%, which is an ideal difficulty of a test for the target learners (Brown, 2013). In addition, item difficulties of ECTT ranged widely from 23 to 81%, which covered the variability of participants' CT and English proficiency levels. In addition, the criterions of r_{pb} values (Fulcher & Davidson, 2007) for all the items were acceptable. Thus, the statistical characteristics of the ECTT items seem to be appropriate for the target EFL learners.

However, as expected, ECTT (53% of CR) was found to be significantly more difficult than JCTT (62%) for Japanese participants. This result accords with Floyd (2011) in that students' low L2 ability depressed CT performance in L2 compared to their CT performance in L1 because cognitive resources are consumed so much before they evaluate the content critically because of lack of L2 proficiency. As noted, the item difficulties across the three CT sections of the ECTT and JCTT had a similar tendency. The correct responses decreased from consistency and analysis to inference in ECTT (57, 52, and 47%, respectively) and JCTT (68, 66, and 53%), which may also be an indication of different levels of CT skills. Thus, regardless of L1 or L2, higher levels of CT skills (e.g., inference) seem to be more difficult than lower levels of CT skills (e.g., consistency).

While the difficulty would be appropriate for the present participants, Cronbach's α in each construct varied. The result revealed that Cronbach's α in Consistency was .70, but the coefficients in Analysis and Inference were relatively low ($\alpha = .52$ and $\alpha = .38$). A similar tendency was observed in the pilot study (Authors, year), in which the internal consistency of each section was $\alpha = .61$ for Consistency, $\alpha = .52$ for Analysis, and $\alpha = .40$ for Inference. This decreasing tendency of the internal consistency reliabilities may also indicate that as the level of the CT process advances, more complex cognitive abilities are necessary (Khamkhong, 2018). In other words, owing to multiple variables involving the success or failure of each item, it seems difficult to achieve high internal consistency reliability for high levels of CT skills. Frisby (1991) noted that validity-related coefficients tend to become low based on the complexity and many dimensions of CT. For the same reason, other CT tests, such as the Cornell test and the Watson-Glaser test, may fail to detect distinct CT skills and reported low reliability of some subtests (Miller, 1992). Therefore, as pointed out earlier (Insight Assessment, 2022; Reid, 2000), we need to secure a sufficient level of reliability of the total ECTT items ($\alpha = .79$) and use it as a whole, not use each CT section separately.

Next, we examined RQ2 (What is the internal structure of the ECTT?), using EFA and SEM, and confirmed that both ECTT and JCTT fit to a two-factor model ("Analysis & Inference" factor and "Consistency" factor). There may be three reasons for not separating the analysis and inference items in the ECTT.

First, the difference between consistency and the other two skills would be much larger than the difference between the analysis and inference skills for the current participants. Consistency items may be at the lower level of CT skills and use "content primarily from within the text" in the reading literacy framework (Organization for Economic Co-Operation and Development (OECD), 2017; Fig. 1), while analysis and inference items both require students to "draw primarily upon outside knowledge," as in the latter phase of the framework. Looking into consistency items, students had to pay attention to discourse markers of the text and its content to arrange sentences in a logical order or fill in the blank cohesively. By contrast, analysis and inference items require students to not only pay attention to text information but also make judgments or produce reasonable solutions or figures by synthesizing text information (Khamkhong, 2018). This difference probably made Analysis and Inference the same factor and Consistency as the other.

Second, the specification on Analysis and Inference items (see Table 1) may not be clearly distinguished, so the constructs these items measure may overlap. For example, even though analysis items relate to charts and tables and involve some calculations, manipulating information is common for the inference items. In fact, the questionnaire revealed that participants felt that the Analysis items needed inference skills as much as inference items (see Fig. 5).

The third cause may stem from participants' low English abilities (EPT = 51% of CR). It is assumed that some participants may not interpret texts deeply to display the CT skills required by analysis or inference items. In other words, the differences in CT abilities might be undermined to some degree in the present study. This result is in line with Just and Carpenter (1992), Floyd (2011), and Manalo and Sheppard (2016), who found that EFL learners must first comprehend the text to produce something, but their L2 ability significantly influenced their display of higher-level CT ability.

Regarding RQ3 (How are ECTT scores correlated with JCTT and EPT?), moderate correlations between the ECTT and EPT ($r = .64$) and between the ECTT and JCTT ($r = .55$) indicate that the ECTT variance can be explained by English ability ($r^2 = .41$) and CT skills ($r^2 = .30$) and also that participants with high CT ability may also have high English reading ability, and vice versa. Since the ECTT is a test written in English, it is reasonable to assume that the ECTT is closely related to the EPT. The results are also similar to previous studies that have examined the relationship between academic English language proficiency and CT (e.g., Grosser & Nel, 2013; Manalo & Sheppard, 2016).

Furthermore, the ECTT was shown to be related to the JCTT, which was created in Japanese with the same constructs of the ECTT. In other words, it can be said that people with high CT performance in Japanese may also have high CT performance in English (and vice versa). However, the results of only moderate correlation between ECTT and JCTT ($r = .55$) and the CR rate of ECTT (53%) being significantly much lower than that of JCTT (62%) are the clear indications of the influence of L2 ability on the performance of CT (Luk & Lin, 2015; Nel & Nel, 2012).

From a pedagogical perspective on the closer relationship between ECTT and EPT and the gap between ECTT and JCTT scores, utilizing such L2 CT tests may provide an opportunity to train EFL learners to read the text deeply and critically and raise both English ability and CT ability. Zare et al. (2021) discussed the idea of dynamic assessment, that is, assessing EFL learners' CT skills and nurturing their English reading ability, and reported positive results on both assessments and the development of learners' reading performance. Another aspect of dynamic assessment is that the process itself is important, adjusting an individual's learning stage. Owing to such benefits, dynamic assessment has gained attention from other researchers (Poehner et al., 2015; Poehner & Lantolf, 2013).

In our case, for example, using the ECTT analysis section not only as a test but also as a teaching material, teachers can train students to enhance CT skills in evaluating textual evidence, and simultaneously, the students learn deep and critical reading skills (Brown et al., 1981). In addition, for those requiring additional help, teachers can flexibly provide supplementary reading materials or questions focusing on the specific CT skills the learners need to strengthen.

Finally, concerning RQ4 (Do participants answer ECTT items based on the skills we intended?) overall, the proportion of the skill we had intended was the largest in each section, but at the same time, it reflected the degree of complicated cognitive levels. The most straightforward items were found to be consistency items because the participants felt that they used consistency-related skills such as coherence and cohesion abilities (53%) in these items.

As for the analysis items, as expected, analytical skills (37%) were used most, followed by inference skills (23%). However, the proportion of the inference skill was as large as that of the Inference section (22%). This may indicate that analysis skill is a construct related to the foundation of inference in the process of CT and that inference may naturally occur after the text is analyzed and understood. This close relationship of analysis and inference skills can also be the evidence why the analysis and inference constructs formed one factor in the EFA and the SEM.

The most complicated skill was inference. Unlike the other two skills, it seemed to require all types of skills with similar proportions (see Fig. 5). Even though inference skill was used the most (22%), other CT skills were equally necessary. In addition, the proportion of English skills and knowledge (i.e., reading 19% and vocabulary 9%) was the greatest among the three sections, which indicates that deep reading was carried out to enhance comprehension. Inference skills correspond to the inference part of the critical thinking process. When inference activity is performed, participants must draw objective conclusions from multiple examples and results and derive reasonable solutions from several assumptions. Analytical skills are inevitably required to understand the results and assumptions written in numbers and letters. Hence, it is likely that inference items require a variety of skills. From these results, it was found that the students activated a variety of skills and knowledge to solve the inference items. In that sense, the inference is a higher-order comprehensive skill in the process.

Interestingly, some participants felt that they used "creativity" in the analysis (9%), followed by the inference items (7%). The result of such small percentages was not surprising since the ECTT is a multiple-choice format test, not a performance test, and the items had not been intended to measure creativity. Even so, the fact that test-takers felt they used creativity to answer the items can be a piece of positive washback or an indication of active reading since creativity is one of the most important skills to acquire when dealing with problems. According to Finke et al. (1992), creativity involves flexible thinking and can be used in problem-solving strategies to generate novel insights and solutions, and it is placed at the highest level of cognitive activity (Anderson et al., 2001).

Conclusion

Considering the recent emphasis on nurturing CT skills in the globalized information age, this study examined the validity of the ECTT, a newly developed CT test, for the use of EFL learners. As a result, we identified the following important points and implications. First, the reliability of the whole test was sufficiently high, and the level of the ECTT was appropriate for upper secondary and university students of English in Japan. Therefore, we were able to enhance the generalizability of the test result. In other words, the test can measure such students' CT skills reliably, so that it may help teachers diagnose their CT skills and choose appropriate instructions to enhance them. Second, the ECTT measured both CT skills and L2 proficiency (i.e., L2 literacy), but had a stronger relationship with L2 proficiency than CT skills for the current participants. Thus, to develop CT skills in English, students need to improve English proficiency accordingly. In this regard, teachers should work on both CT skills and English proficiency in English class, choosing appropriate materials that can train both, such as ECTT. Third, test-takers' most needed skills were aligned with the skills we intended to measure, and higher-order CT skills require them to use a greater variety of skills. This is strong evidence that ECTT is a valid test, and the test results can be interpreted as such.

However, the study also poses some challenges with regard to creating items that measure higher-order CT skills. We could not obtain a sufficient level of internal consistency of the inference section, nor could we separate the inference and analysis sections as separate factors. One reason is that CT skills may be hierarchical or semi-hierarchical and, because of that, such higher-order CT skills require multiple sub-CT skills, which makes it harder to extract pure constructs. Therefore, to obtain a reliable score, it is better to use the test as a whole (e.g., Reid, 2000) or combine various levels of CT skills. Another limitation of this study was that we made a Japanese CT test to compare with ECTT, instead of using existing CT tests. This is because we wanted to make a specific comparison with ECTT, but the result may lack extrapolation in the validity argument (e.g., Chapelle et al., 2015). Therefore, in the future, some comparative research between the ECTT and the other external English CT tests is necessary.

In addition, the washback and diagnostic functions of the ECTT should be examined, as the test is different from conventional English tests. In accordance with the dynamic assessment idea (e.g., Brown et al., 1981; Zare et al., 2021), ECTT can be recommended for use in English Medium Instruction (EMI) programs or in English for Academic Purposes (EAP) courses. Students who may read critically in their L1 may not necessarily do so in their L2; thus, it is worthwhile to utilize this type of test to raise awareness of CT skills and for assessment purposes. It can also be used as an EAP placement test to estimate the kinds of CT-related activities that need to be integrated in the course. In response to the current demand for cultivating students' CT skills in English education, classroom teachers may want to introduce practices and tests that develop students' CT skills. In this regard, this study hopes to provide a significant contribution to the evaluation of CT skills in the context of EFL education.

Abbreviations

AGFI	Adjusted goodness-of-fit index
CFI	Comparative fit index
CR	Correct response
CT	Critical thinking
DWLS	Diagonally weighted least squares
EFL	English as a foreign language
EPT	English proficiency test
ECTT	English Critical Thinking Test
EFA	Exploratory factor analysis
JCTT	Japanese Critical Thinking Test
ML	Maximum likelihood
SEM	Structural equation modeling
RMSEA	Root mean square error of approximation
SRMR	Standardized root mean square residual

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40468-022-00193-2.

Additional file 1.

Additional file 2.

Additional file 3.

Acknowledgements

We would like to express our gratitude to Kosuke Nagasue for his cooperation in data collection and Yoshishige Suda for his assistance in making and revising the CT items. Without their help, we could not proceed with our research. This research was supported by JSPS KAKENHI Grant Number JP22K00758.

Authors' contributions

AH implemented the experiment and drafted the manuscript. HO assisted AH in drafting the literature review section. TK assisted AH in performing the statistical analysis and drafted the result section. HM assisted AH in drafting the "Method" section. All authors prepared the materials for the experiments and read and approved the final manuscript.

Funding

This research was supported by JSPS KAKENHI (Grant Number 16K02953).

Availability of data and materials

Data and matrials cannot be shared currently due to confidentiality agreement with participants and material developers.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 7 March 2022 Accepted: 3 September 2022
Published online: 05 October 2022

References

Adams, M. H., Whitlow, J. F., Stover, L. M., & Johnson, K. W. (1996). Critical thinking as an educational outcome: An evaluation of current tools of measurement. *Nurse Educator*, *21*(3), 23–32. https://doi.org/10.1097/00006223-199605000-00009.

Aloqaili, A. S. (2012). The relationship between reading comprehension and critical thinking: A theoretical study. *Journal of King Saud University - Languages and Translation*, *24*(1), 35–41. https://doi.org/10.1016/j.jksult.2011.01.001.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Wittrock, M. C. (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

AssessmentDay. (2020). Watson–Glaser Critical Thinking Appraisal. Retrieved from https://www.assessmentday.co.uk/watson-glaser-critical-thinking.htm.

Association of American Colleges and Universities [AACU] (2018). Fulfilling the American Dream: Liberal Education and the Future of Work. https://www.aacu.org/research/2018-future-of-work.

Brown, A. L., Campione, J. C., & Day, J. D. (1981). Learning to learn: On training students to learn from texts. *Educational Researcher*, *10*, 14–21. https://doi.org/10.3102/0013189X010002014.

Brown, J. D. (2013). Classical test theory. In G. Fulcher, & F. Davidson (Eds.), *The Routledge handbook of language testing*, (pp. 337–349). Routledge.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation modeling*, (pp. 136–162). Sage.

Cambridge Assessment Admissions Testing. (2020). Thinking Skills Assessment. https://www.admissionstesting.org/for-test-takers/thinking-skills-assessment/.

Chapelle, C. A., & Voss, E. (2021). *Validity argument in language testing: Case studies of validation research*. Cambridge University Press.

Clemson University (2022). Clemson Thinks2.https://www.clemson.edu/academics/programs/thinks2/index.html.

EIKEN Foundation of Japan (2022). EIKEN Tests. https://www.eiken.or.jp/eiken/en/eiken-tests/.

Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and abilities. In J. B. Baron, & R. J. Sternberg (Eds.), *Teaching thinking skills: Theory and practice*, (pp. 9–26). W. H. Freeman and Company.

Ennis, R. H. (2003). Critical thinking assessment. In D. Fasko (Ed.), *Critical thinking and reasoning: Current theories, research, and practice*. Hampton Press.

Ennis, R. H. (2018). Critical thinking across the curriculum: A vision. *TOPOI*, *37*(1), 165–184. https://doi.org/10.1007/s11245-016-9401-4.

Ennis, R. J., & Millman, J. (1985). *Cornell tests of critical thinking*. Midwest Publications.

ETS (2020). About the GRE General Test. https://www.ets.org/gre/revised_general/about.

Facione, P. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. Executive summary the Delphi Report*. California Academic Press https://www.qcc.cuny.edu/socialsciences/ppecorino/CT-Expert-Report.pdf.

Facione, P. A. (2000). The disposition toward critical thinking: Its character, measurement, and relationship to critical thinking skill. *Informal Logic*, *20*(1), 61–84. https://doi.org/10.22329/il.v20i1.2254.

Finke, R. A., Ward, T. B., & Smith, S. M. (1992). *Creative cognition: Theory, research, and applications*. MIT Press.

Floyd, C. B. (2011). Critical thinking in a second language. *Higher Education Research and Development*, *30*(3), 289–302. https://doi.org/10.1080/07294360.2010.501076.

Forero, C., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 625–641. https://doi.org/10.1080/10705510903203573.

Frisby, C. L. (1991). A meta-analytic investigation of the relationship between grade level and mean scores on the Cornell Critical Thinking Test (Level X). *Measurement and Evaluation in Counseling and Development*, *23*, 162–170.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. Routledge.

Griffin, P., McGaw, B., & Care, E. (Eds.) (2012). *Assessment and teaching of 21st century skills*. Springer.

Grosser, M., & Nel, M. (2013). The relationship between the critical thinking skills and the academic language proficiency of prospective teachers. *South African Journal of Education, 33*(2), 1–17.

Halpern, D. F. (1999). Teaching for critical thinking: Helping college students develop the skills and dispositions of a critical thinker. *New Directions for Teaching and Learning*, *80*, 69–74 https://doi-org.ezproxy.tulips.tsukuba.ac.jp/10.1002/tl.8005.

Hattori, M., Zhang, G., & Preacher, K. J. (2017). Multiple local solutions and Geomin rotation. *Multivariate Behavioral Research*, *52*(6), 720–731. https://doi.org/10.1080/00273171.2017.1361312.

Hori, K. (2005). Inshi bunseki ni okeru inshisuu ketteihou: Heikou bunseki o chushin ni shite [Determining the number of factors in exploratory factor analysis: Focusing on parallel analysis]. *Kagawa University Economic Review*, *77*, 35–70.

Insight Assessment (2022). CCTST test manual. http://eccdl.dcccd.edu/DArumugam/Recovered/Desktop/CCTST%20Assessment%20Data/CCTST%20Test%20Manual%202013_2.pdf.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL, 8. User's reference guide*. Scientific Software https://www.jstor.org/stable/10.5325/jgeneeduc.62.4.0297#metadata_info_tab_contents.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*(1), 122–149. https://doi.org/10.1037/0033-295X.99.1.122.

Khamkhong, S. (2018). Developing English L2 critical reading and thinking skills through the Pisa reading literacy assessment framework: A case study of Thai EFL learners. *3L The Southeast Asian Journal of English Language Studies*, *24*(3), 83–94. https://doi.org/10.17576/3L-2018-2403-07.

Kline, R. B. (2005). *Principles and practice of structural equation modeling (2nd. ed.)*. Guilford Press.

Kuhn, D. (2015). Thinking together and alone. *Educational Researcher*, *44*(1), 46–53. https://doi.org/10.3102/0013189X15569530.

Kusumi, T. (2010). Hihantekishikou to koujI riterashii [Critical thinking and higher-order literacy]. In T. Kusumi (Ed.), *Shikou to gengo*, (pp. 134–160). Kitaoji Shobou.

Kusumi, T. (2018). Riterashii wo sasaeru hihantekishikou: Dokusho kagaku heno shisa [Critical thinking in support of literacy development: Implications for reading science]. *The Science of Reading*, *60*, 129–137. https://doi.org/10.19011/sor.60.3_129.

Lee, Y. H. (2015). Facilitating critical thinking using the C-QRAC collaboration script: Enhancing science reading literacy in a computer-supported collaborative learning environment. *Computers & Education*, *88*, 182–191. https://doi.org/10.1016/j.compedu.2015.05.004.

Luk, J., & Lin, A. (2015). Voices without words: Doing critical literate talk in English as a second language. *TESOL Quarterly*, *49*(1), 67–91.

Manalo, E., & Sheppard, C. (2016). How might language affect critical thinking performance? *Thinking Skills and Creativity*, *21*, 41–49. https://doi.org/10.1016/j.tsc.2016.05.005.

Marin, M. A., de la Pava, L., & DI. (2017). Conceptions of critical thinking from university EFL teachers. *English Language Teaching*, *10*(7), 78–88. https://doi.org/10.5539/elt.v10n7p78.

Miller, M. A. (1992). Outcomes evaluation: Measuring critical thinking. *Journal of Advanced Nursing*, *17*(12), 1401–1407. https://doi.org/10.1111/j.1365-2648.1992.tb02810.x.

Mîndrilă, D. (2010). Maximum likelihood (ML) and diagonally weighted least squares (DWLS) estimation procedures: A comparison of estimation bias with ordinal and multivariate non-normal data. *International Journal of Digital Society (IJDS)*, *1*(1) https://infonomics-society.org/wp-content/uploads/ijds/published-papers/volume-1-2010/Maximum-Likelihood-ML-and-Diagonally-Weighted-Least-Squares-DWLS-Estimation-Procedures-A-Comparison-of-Estimation-Bias-with-Ordinal-and-Multivariate-Non-Normal-Data.pdf.

Mineshima, M., & Imai, R. (2017). Hihanteki sihkouryoku wo ikuseisuru koudairenkei no kokoromi [Developing Critical Thinking Skills in High School and University: How to Make English Lessons More Intelligent]. *The Chubu English Language Education Society*, *46*, 133–140. https://doi.org/10.20713/celes.46.0_133.

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2016). Jikigakushu shidouyouryoutou ni muketa koremadeno shingino matome hosokushiryo [Summary of the discussion on the next Course of Study]. https://www.mext.go.jp/content/1377021_4_2.pdf.

Nel, N., & Nel, M. (2012). English language. In N. Nel, M. Nel, & A. Hugo (Eds.), *Learner Support in a Diverse Classroom: A Guide for Foundation, Intermediate and Senior Phase Teachers of Language and Mathematics*, (pp. 3–18). Van Schaik Publishers.

Nicholas, M. C., & Labig Jr., C. E. (2013). Faculty approaches to assessing critical thinking in the humanities and the natural and social sciences: Implications for general education. *The Journal of General Education*, *62*, 297–319.

Organization for Economic Co-Operation and Development (OECD) (2017). *PISA 2015 Assessment and analytical framework: Science, reading, mathematic, financial literacy and collaborative problem solving*. OECD Publishing. https://doi.org/10.1787/19963777.

Poehner, M. E., & Lantolf, J. P. (2013). Bringing the ZPD into the equation: Capturing L2 development during Computerized Dynamic Assessment. *Language Teaching Research*, *17*(3), 323–342. https://doi.org/10.1177/1362168813482935.

Poehner, M. E., Zhang, J., & Lu, X. (2015). Computerized dynamic assessment (C-DA): Diagnosing L2 development according to learner responsiveness to mediation. *Language Testing*, *32*(3), 337–357. https://doi.org/10.1177/0265532214560390.

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing https://www.R-project.org/.

Rane-Szostak, D., & Robertson, J. F. (1996). Issues in measuring critical thinking: Meeting the challenge. *Journal of Nursing Education*, *35*(1), 5–11. https://europepmc.org/article/med/8926519. https://doi.org/10.3928/0148-4834-19960101-04.

Rear, D. (2019). One size fits all? The limitations of standardised assessment in critical thinking. *Assessment & Evaluation in Higher Education*, *44*(5), 664–675. https://doi.org/10.1080/02602938.2018.1526255.

Recruit. (2022). Synthetic personality inventory. https://www.spi.recruit.co.jp.

Reid, H. (2000). *The correlation between a general critical thinking skills test and a discipline specific critical thinking test for associate degree nursing students [Doctoral dissertation, University of North Texas]*. UNT Digital Library https://digital.library.unt.edu/ark:/67531/metadc2505/m1/.

Reid, S., & Chin, P. (2021). Assessing critical thinking in L2: An exploratory study. *Shiken*, *25*, 8–21. https://doi.org/10.37546/JALTSIG.TEVAL25.1.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36 https://www.jstatsoft.org/article/view/v048i02.

Stroupe, R. R. (2006). Integrating critical thinking throughout ESL curricula. *TESL Reporter*, *39*(2), 42–61.

Takeda, I. (2016). Kotogakko ni okeru hihantekishikoryoku wo hagukumu eigojugyokaihatsu [Developing English classes to cultivate critical thinking skills in high school]. *Graduate School of Education, Shimane University, One-year Incumbent Teacher Course Report*, *7*, 51–60 https://ir.lib.shimane-u.ac.jp/ja/journal/E-KKS/7/--/article/36344.

Tests.com LLC. (2020). Cornell critical thinking test guide. https://www.tests.com/Cornell-Critical-Thinking-Testing.

Zare, M., Barjesteh, H., & Biria, R. (2021). Enhancing EFL learners' reading comprehension skill through critical thinking-oriented dynamic assessment. *Teaching English Language*, *15*, 189–214. https://doi.org/10.22132/TEL.2021.133238.

Zhou, J., Jiang, Y., & Yao, Y. (2015). The investigation on critical thinking ability in EFL reading class. *English Language Teaching*, *8*, 83–94.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission

► Rigorous peer review

► Open access: articles freely available online

► High visibility within the field

► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com